CERN–2010–001
22 March 2010

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

2009 CERN–Latin-American School of High-Energy Physics

Recinto Quirama, Antioquia Region, Colombia
15–28 March 2009

Proceedings

Editors: C. Grojean
M. Spiropulu

GENEVA
2010

CERN–100 copies printed–March 2010

In Memoriam

These Proceedings are dedicated to the memory of Juan Antonio Rubio Rodriguez whose personal contributions and support were fundamental to the establishment of the series of CERN–Latin-American Physics Schools.

Abstract

The CERN–Latin-American School of High-Energy Physics is intended to give young physicists an introduction to the theoretical aspects of recent advances in elementary particle physics. These proceedings contain lectures on quantum field theory, quantum chromodynamics, physics beyond the Standard Model, neutrino physics, flavour physics and CP violation, particle cosmology, high-energy astro-particle physics, and heavy-ion physics, as well as trigger and data acquisition, and commissioning and early physics analysis of the ATLAS and CMS experiments. Also included are write-ups of short review projects performed by the student discussions groups.

Preface

The fifth School in the series of Latin-American Schools of High-Energy Physics took place from 15 to 28 March 2009 in Recinto Quirama, Antioquia, Colombia. It was organized by CERN with the support of the Universidad Antonio Nariño, Bogotá and Universidad de Antioquia, Medellín.

The School was generously supported by CERN; CIEMAT, Spain; the Colombian Ministry of Education, together with the University of Antioquia, Antonio Nariño University and the National University of Colombia, Bogotá; the Brazilian Ministry of Science and Technology; and CLAF, the Centro Latino Americano de Física. Our sincere thanks go to all the sponsors for making it possible to organize the School and for contributing to its success.

Professor Marta Losada from the University Antonio Nariño, Bogotá, acted as local director for the School, strongly assisted by Enrico Nardi from the University of Antioquia, Medellín. The other local committee members were Carlos Quimbay from the National University of Colombia, Bogotá, and Juan Carlos Sanabria from the University Los Andes, Bogotá. We are extremely grateful to Marta Losada and Enrico Nardi for their excellent work in organizing the School and for creating such a wonderful atmosphere for the participants.

Fifty-nine students from 16 different countries, together with 11 Colombian 'listeners', attended the School. Following the tradition of the School the students shared twin rooms mixing nationalities, and in particular the Europeans mixed with Latin Americans.

The 12 lecturers came from Europe, Latin America and Israel. The lectures, which were in English, were complemented by daily discussion sessions led by three physicists from Latin America and one from the USA. The lectures were given in the main hall where the students also displayed their work in the form of posters on a special evening session during the first week. The posters were left on display until the end of the School.

Our thanks are due to the lecturers and discussion leaders for their active participation in the School and for making the scientific programme so stimulating. The students who in turn manifested their good spirits during two intense weeks undoubtedly appreciated their personal contributions in answering questions and explaining points of theory.

The School was hosted in the beautiful Hotel Recinto Quirama, a colonial-style hotel close to Medellín Airport. We are indebted to the hotel for its friendly staff who certainly contributed to the good spirit of the School and, in particular, to the hotel chef, Mr. Juan David Jaramillo, who provided a varied and much-appreciated menu throughout the school.

We are very grateful to Danielle Métral for her efforts in the lengthy preparations for the School and for her day-to-day care of the School. Her efficient work, friendly attitude, and continuous care of the participants and their needs were highly appreciated.

The students will certainly remember several interesting excursions, which included visits to Santa Fé de Antioquia, and El Peñol a 200-metre-high granite monolith with 650 steps that the participants climbed. They greatly appreciated the excellent social programme, including horse riding and performances by local groups that were organized by Enrico Nardi.

However, the success of the School was to a large extent due to the students themselves. Their poster session was very well prepared and highly appreciated, and throughout the School they participated actively during the lectures, in the discussion sessions, and in the different activities and excursions.

Egil Lillestøl and Nick Ellis
on behalf of the Organizing Committee

People in the photograph

1 Nick ELLIS
2 Javier Alberto DUARTE CHAVEZ
3 Diego RESTREPO
4 Egil LILLESTØL
5 Marina VON STEINKIRCH
6 Hayk HAKOBYAN
7 William PONCE
8 Pedro QUINTERO
9 José Alejandro ROSABAL RODRIGUEZ
10 Guillermo FIORENTINI
11 Jhovanny Andres MEJIA GUISAO
12 Jose Andres MONROY MONTAÑEZ
13 Miguel PINO ROZAS
14 Javier BROCHERO CIFUENTES
15 Patricia REBELLO TELES
16 Diego Alonso ROMERO MALTRANA
17 Carlos FLOREZ BUSTOS
18 Yosef NIR
19 Diego Julian RODRIGUEZ PATARROYO
20 Jonathan IMONG
21 Fernando QUINONEZ GRANADOS
22 Alexander AUSTREGESILO
23 Cristian MARTINEZ
24 Mary DIAZ
25 Fatima PADILLA CABAL
26 Mike SEYMOUR
27 Pía ZURITA
28 Pavel JEZ
29 Fabián Darío VILLALBA-PARDO
30 Wayne DE PAULA
31 Estela Alejandra GARCES-GARCIA
32 Bruce YEE RENDON
33 Hector MARTINEZ
34 Ulises Jesus SALDAÑA-SALAZAR
35 Alberto GAGO MEDINA
36 Enrico NARDI
37 Carlos MEDINA HERNANDEZ
38 Elias RON
39 Kate SHAW
40 Angela BUECHLER
41 Ivan ARRAUT
42 Antonio ORTIZ VELASQUEZ
43 Miguel MONCADA
44 Danielle METRAL
45 Jose David RUIZ ALVAREZ
46 Paulo Henrique FLOSE REIMBERG
47 Antonio RIOTTO
48 Christophe SALZMANN
49 Gaston Leonardo ROMEO
50 Flavia Alejandra GOMEZ ALBARRACIN
51 German David CARRILLO MONTOYA
52 Katharine LENEY
53 Geraldo Magela SEVERINO VASCONCELOS
54 John SWAIN
55 Javier TIFFENBERG
56 Mauricio BUSTAMANTE
57 Maria Laura GONZALEZ SILVA
58 Lea CAMINADA
59 Jordan MARTINS
60 Federico BENITEZ
61 Victor Ivan GIRALDO RIVERA
62 Leandro CIERI
63 Andreas HOECKER
64 Kim ALWYN
65 Bruno GONÇALVES
66 Marta LOSADA
67 John ELLIS
68 Folkert KOETSVELD
69 Marek NOWAKOWSKI
70 Luis Alberto WILLS TORO
71 Hugo Raymundo MARQUEZ FALCON
72 Enrique ARRIETA DIAZ
73 Richard BENAVIDES PALACIOS
74 David Alejandro MARTINEZ CAICEDO
75 Hector Javier HORTUA ORJUELA
76 Cesar ARIAS
77 Carolina ARBELAEZ
78 Boris OSORNO TORRES
79 Jorge Luis NISPERUZA TOLEDO
80 Marta Liliana SANCHEZ PELAEZ
81 Daniel Fernando HIGUITA BORJA
82 Alejandro JARAMILLO MORENO
83 Yithsbey GIRALDO
84 Mauricio VELASQUEZ

PHOTOGRAPHS (MONTAGE)

Contents

Student project write-ups (edited by discussion leaders)

Introductory lectures on quantum field theory*

L. Álvarez-Gaumé[a,†] *and M. A. Vázquez-Mozo*[b‡]
[a] CERN, Geneva, Switzerland
[b] University of Salamanca, Salamanca, Spain

Abstract
In these lectures we present a few topics in quantum field theory in detail. Some of them are conceptual and some more practical. They have been selected because they appear frequently in current applications to particle physics and string theory.

1 Introduction

These notes summarize the lectures presented at the 2005 CERN–CLAF school in Malargüe, Argentina and the 2009 CERN–CLAF school in Medellín, Colombia. The audience on both occasions was composed to a large extent of students in experimental high-energy physics with an important minority of theorists. In nearly ten hours it is quite difficult to give a reasonable introduction to a subject as vast as quantum field theory. For this reason the lectures were intended to provide a review of those parts of the subject to be used later by other lecturers. Although a cursory acquaintance with the subject of quantum field theory is helpful, the only requirement to follow the lectures is a working knowledge of quantum mechanics and special relativity.

The guiding principle in choosing the topics presented (apart to serve as introductions to later courses) was to present some basic aspects of the theory that present conceptual subtleties. Those topics one often is uncomfortable with after a first introduction to the subject. Among them we have selected:

- The need to introduce quantum fields, with the great complexity this implies.
- Quantization of gauge theories and the rôle of topology in quantum phenomena. We have included a brief study of the Aharonov–Bohm effect and Dirac's explanation of the quantization of the electric charge in terms of magnetic monopoles.
- Quantum aspects of global and gauge symmetries and their breaking.
- Anomalies.
- The physical idea behind the process of renormalization of quantum field theories.
- Some more specialized topics, like the creation of particles by classical fields and the very basics of supersymmetry.

These notes have been written following closely the original presentation, with numerous clarifications. Sometimes the treatment given to some subjects has been extended, in particular the discussion of the Casimir effect and particle creation by classical backgrounds. Since no group theory was assumed, we have included an Appendix with a review of the basic concepts.

For lack of space and on purpose, few proofs have been included. Instead, very often we illustrate a concept or property by describing a physical situation where it arises. Full details and proofs can be found in the many textbooks in the subject, and in particular in the ones provided in the bibliography [1–10].

*Based on lectures delivered by L.A.-G. at the 3rd CERN–CLAF School of High-Energy Physics, Malargüe (Argentina), February 27th–March 12th, 2005 and at the 5th CERN–CLAF School of High-Energy Physics, Medellín (Colombia), 15th–28th March, 2009

†Luis.Alvarez-Gaume@cern.ch
‡Miguel.Vazquez-Mozo@cern.ch, vazquez@usal.es

Specially modern presentations, very much in the spirit of these lectures, can be found in Refs. [4,5,9,10]. We should nevertheless warn the reader that we have been a bit cavalier about references. Our aim has been to provide mostly a (non-exhaustive) list of references for further reading. We apologize to those authors who feel misrepresented.

1.1 A note about notation

Before starting it is convenient to review the notation used. Throughout these notes we will be using the metric $\eta_{\mu\nu} = \mathrm{diag}\,(1,-1,-1,-1)$. Derivatives with respect to the four-vector $x^\mu = (ct, \vec{x})$ will be denoted by the shorthand

$$\partial_\mu \equiv \frac{\partial}{\partial x^\mu} = \left(\frac{1}{c}\frac{\partial}{\partial t}, \vec{\nabla}\right). \tag{1.1}$$

As usual space-time indices will be labelled by Greek letters ($\mu, \nu, \ldots = 0, 1, 2, 3$) while Latin indices will be used for spatial directions ($i, j, \ldots = 1, 2, 3$). In many expressions we will use the notation $\sigma^\mu = (\mathbf{1}, \sigma^i)$ where σ^i are the Pauli matrices

$$\sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \sigma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{1.2}$$

Sometimes we make use of the Feynman slash notation $\not{a} = \gamma^\mu a_\mu$. Finally, unless stated otherwise, we work in natural units $\hbar = c = 1$.

2 Why do we need quantum field theory after all?

Despite the impressive success of quantum mechanics in describing atomic physics, it was immediately clear after its formulation that its relativistic extension was not free of difficulties. These problems were clear already to Schrödinger, whose first guess for a wave equation of a free relativistic particle was the Klein–Gordon equation

$$\left(\frac{\partial^2}{\partial t^2} - \nabla^2 + m^2\right)\psi(t, \vec{x}) = 0. \tag{2.1}$$

This equation follows directly from the relativistic 'mass-shell' identity $E^2 = \vec{p}^{\,2} + m^2$ using the correspondence principle

$$\begin{aligned} E &\rightarrow i\frac{\partial}{\partial t}, \\ \vec{p} &\rightarrow -i\vec{\nabla}. \end{aligned} \tag{2.2}$$

Plane wave solutions to the wave equation (2.1) are readily obtained

$$\psi(t, \vec{x}) = e^{-ip_\mu x^\mu} = e^{-iEt + i\vec{p}\cdot\vec{x}} \qquad \text{with} \qquad E = \pm\omega_p \equiv \pm\sqrt{\vec{p}^{\,2} + m^2}. \tag{2.3}$$

In order to have a complete basis of functions, one must include plane waves with both $E > 0$ and $E < 0$. This implies that given the conserved current

$$j_\mu = \frac{i}{2}\Big(\psi^*\partial_\mu\psi - \partial_\mu\psi^*\,\psi\Big), \tag{2.4}$$

its time-component is $j^0 = E$ and therefore does not define a positive-definite probability density.

A complete, properly normalized, continuous basis of solutions of the Klein-Gordon equation (2.1) labelled by the momentum $\vec{p}$ can be defined as

$$f_p(t, \vec{x}) \;=\; \frac{1}{(2\pi)^2\sqrt{2\omega_p}}\, e^{-i\omega_p t + i\vec{p}\cdot\vec{x}},$$

$$f_{-p}(t,\vec{x}) \;=\; \frac{1}{(2\pi)^2\sqrt{2\omega_p}}\, e^{i\omega_p t - i\vec{p}\cdot\vec{x}}. \tag{2.5}$$

Given the inner product

$$\langle\psi_1|\psi_2\rangle = i\int d^3x\Big(\psi_1^*\partial_0\psi_2 - \partial_0\psi_1^*\,\psi_2\Big)$$

the states (2.5) form an orthonormal basis

$$\begin{aligned}
\langle f_p|f_{p'}\rangle &= \delta(\vec{p}-\vec{p}\,'), \\
\langle f_{-p}|f_{-p'}\rangle &= -\delta(\vec{p}-\vec{p}\,'), &(2.6)\\
\langle f_p|f_{-p'}\rangle &= 0. &(2.7)
\end{aligned}$$

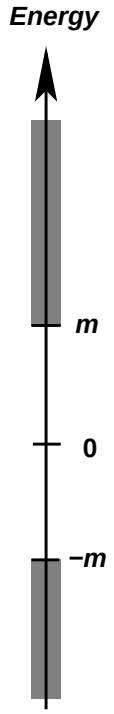


Fig. 1: Spectrum of the Klein–Gordon wave equation

The wave functions $f_p(t,x)$ describe states with momentum $\vec{p}$ and energy given by $\omega_p = \sqrt{\vec{p}^{\,2}+m^2}$. On the other hand, the states $|f_{-p}\rangle$ not only have a negative scalar product but they actually correspond to negative energy states

$$i\partial_0 f_{-p}(t,\vec{x}) = -\sqrt{\vec{p}^{\,2}+m^2}\, f_{-p}(t,\vec{x}). \tag{2.8}$$

Therefore the energy spectrum of the theory satisfies $|E| > m$ and is unbounded from below (see Fig. 1). Although in a case of a free theory the absence of a ground state is not necessarily a fatal problem, once the theory is coupled to the electromagnetic field this is the source of all kinds of disasters, since nothing can prevent the decay of any state by emission of electromagnetic radiation.

The problem of the instability of the 'first-quantized' relativistic wave equation can be heuristically tackled in the case of spin-$\frac{1}{2}$ particles, described by the Dirac equation

$$\left(-i\beta\frac{\partial}{\partial t} + \vec{\alpha}\cdot\vec{\nabla} - m\right)\psi(t,\vec{x}) = 0, \tag{2.9}$$

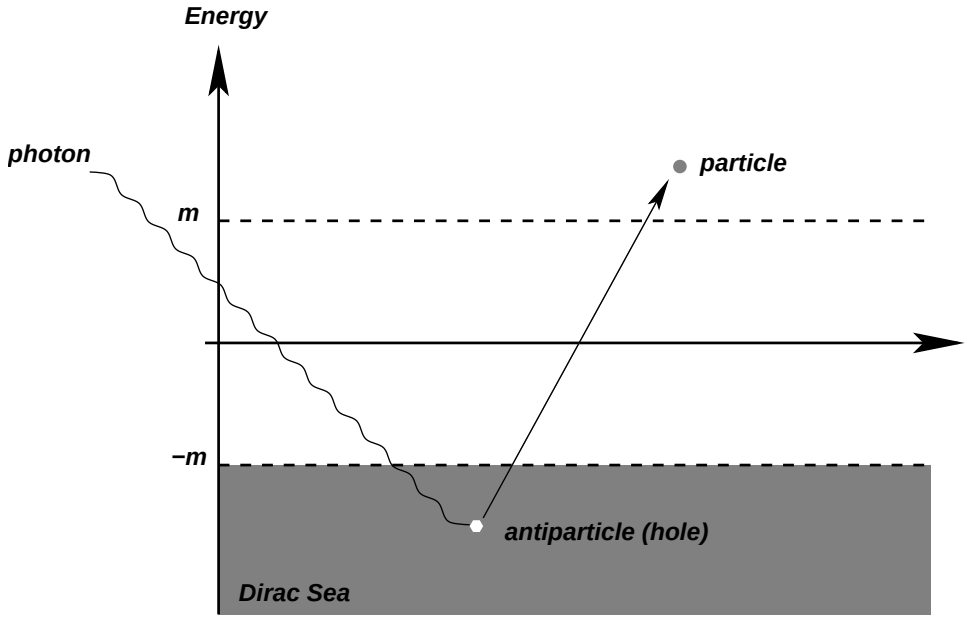


Fig. 2: Creation of a particle–antiparticle pair in the Dirac sea picture

where $\vec{\alpha}$ and β are 4×4 matrices

$$\alpha^i = \begin{pmatrix} 0 & i\sigma^i \\ -i\sigma^i & 0 \end{pmatrix}, \qquad \beta = \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix}, \tag{2.10}$$

with σ^i the Pauli matrices, and the wave function $\psi(t, \vec{x})$ has four components. The wave equation (2.9) can be thought of as a kind of 'square root' of the Klein–Gordon equation (2.1), since the latter can be obtained as

$$\left(-i\beta\frac{\partial}{\partial t} + \vec{\alpha}\cdot\vec{\nabla} - m\right)^\dagger \left(-i\beta\frac{\partial}{\partial t} + \vec{\alpha}\cdot\vec{\nabla} - m\right)\psi(t,\vec{x}) = \left(\frac{\partial^2}{\partial t^2} - \nabla^2 + m^2\right)\psi(t,\vec{x}). \tag{2.11}$$

An analysis of Eq. (2.9) along the lines of the one presented above for the Klein–Gordon equation leads again to the existence of negative energy states and a spectrum unbounded from below as in Fig. 1. Dirac, however, solved the instability problem by pointing out that now the particles are fermions and therefore they are subject to Pauli's exclusion principle. Hence, each state in the spectrum can be occupied by at most one particle, so the states with $E = m$ can be made stable if we assume that *all* the negative energy states are filled.

If Dirac's idea restores the stability of the spectrum by introducing a stable vacuum where all negative energy states are occupied, the so-called Dirac sea, it also leads directly to the conclusion that a single-particle interpretation of the Dirac equation is not possible. Indeed, a photon with enough energy ($E > 2m$) can excite one of the electrons filling the negative energy states, leaving behind a 'hole' in the Dirac sea (see Fig. 2). This hole behaves as a particle with equal mass and opposite charge that is interpreted as a positron, so there is no escaping the conclusion that interactions will produce particle–antiparticle pairs out of the vacuum.

In spite of the success of the heuristic interpretation of negative energy states in the Dirac equation, this is not the end of the story. In 1929 Oskar Klein stumbled into an apparent paradox when trying to describe the scattering of a relativistic electron by a square potential using Dirac's wave equation [11] (for pedagogical reviews see Refs. [12, 13]). In order to capture the essence of the problem without

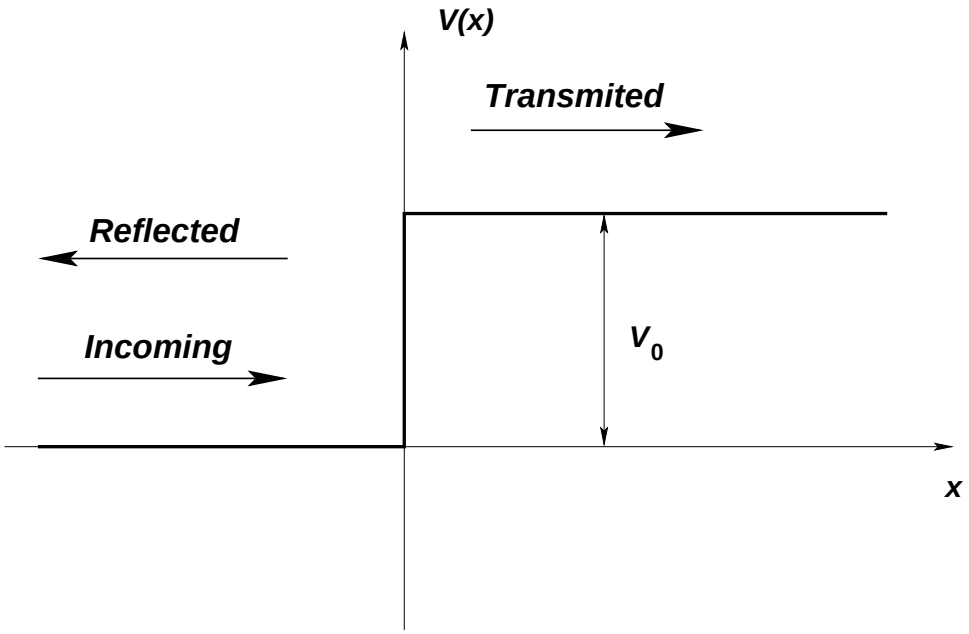


Fig. 3: Illustration of the Klein paradox

entering into unnecessary complication we will study Klein's paradox in the context of the Klein–Gordon equation.

Let us consider a square potential with height $V_0 > 0$ of the type shown in Fig. 3. A solution to the wave equation in regions I and II is given by

$$\begin{aligned} \psi_I(t,x) &= e^{-iEt+ip_1x} + Re^{-iEt-ip_1x}, \\ \psi_{II}(t,x) &= Te^{-iEt+p_2x}, \end{aligned} \tag{2.12}$$

where the mass-shell condition implies that

$$p_1 = \sqrt{E^2 - m^2}, \qquad p_2 = \sqrt{(E-V_0)^2 - m^2}. \tag{2.13}$$

The constants R and T are computed by matching the two solutions across the boundary $x = 0$. The conditions $\psi_I(t,0) = \psi_{II}(t,0)$ and $\partial_x\psi_I(t,0) = \partial_x\psi_{II}(t,0)$ imply that

$$T = \frac{2p_1}{p_1+p_2}, \qquad R = \frac{p_1-p_2}{p_1+p_2}. \tag{2.14}$$

At first sight one would expect a behavior similar to the one encountered in the non-relativistic case. If the kinetic energy is bigger than V_0 both a transmitted and reflected wave are expected, whereas when the kinetic energy is smaller than V_0 one only expects to find a reflected wave, the transmitted wave being exponentially damped within a distance of a Compton wavelength inside the barrier.

Indeed this is what happens if $E - m > V_0$. In this case both p_1 and p_2 are real and we have a partly reflected, and a partly transmitted wave. In the same way, if $E - m < V_0$ and $E - m < V_0 - 2m$ then p_2 is imaginary and there is total reflection.

However, in the case when $V_0 > 2m$ and the energy is in the range $V_0 - 2m < E - m < V_0$ a completely different situation arises. In this case one finds that both p_1 and p_2 are real and therefore the incoming wave function is partially reflected and partially transmitted across the barrier. This is a shocking result, since it implies that there is a nonvanishing probability of finding the particle at any point across the barrier with negative kinetic energy ($E - m - V_0 < 0$)! This weird result is known as Klein's paradox.

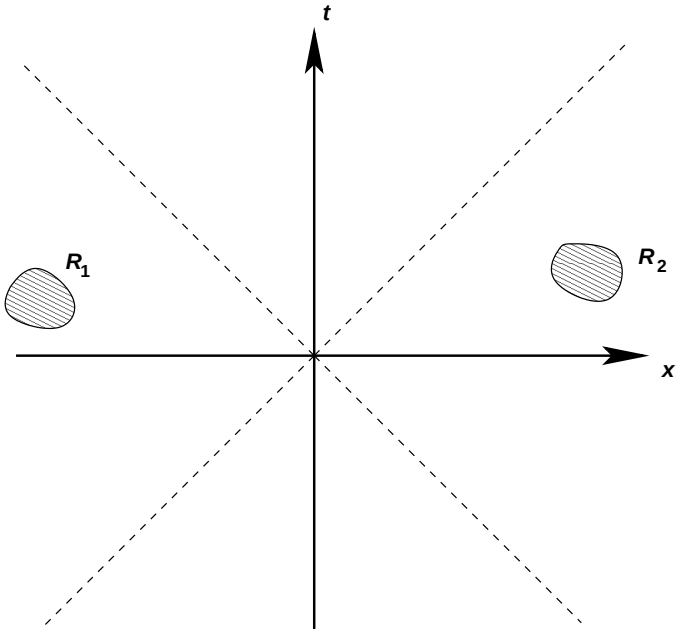


Fig. 4: Two regions R_1, R_2 that are causally disconnected

As with the negative energy states, the Klein paradox results from our insistence in giving a single-particle interpretation to the relativistic wave function. Actually, a multiparticle analysis of the paradox [12] shows that what happens when $E - m > V_0 - 2m$ is that the reflection of the incoming particle by the barrier is accompanied by the creation of particle–antiparticle pairs out of the energy of the barrier (notice that for this to happen it is required that $V_0 > 2m$, the threshold for the creation of a particle–antiparticle pair).

Actually, this particle creation can be understood by noticing that the sudden potential step in Fig. 3 localizes the incoming particle with mass m in distances smaller than its Compton wavelength $\lambda = \frac{1}{m}$. This can be seen by replacing the square potential by another one where the potential varies smoothly from 0 to $V_0 > 2m$ in distances scales larger than $1/m$. This case was worked out by Sauter shortly after Klein pointed out the paradox [14]. He considered a situation where the regions with $V = 0$ and $V = V_0$ are connected by a region of length d with a linear potential $V(x) = \frac{V_0 x}{d}$. When $d > \frac{1}{m}$ he found that the transmission coefficient is exponentially small[1].

The creation of particles is impossible to avoid whenever one tries to locate a particle of mass m within its Compton wavelength. Indeed, from Heisenberg's uncertainty relation we find that if $\Delta x \sim \frac{1}{m}$, the fluctuations in the momentum will be of order $\Delta p \sim m$ and fluctuations in the energy of order

$$\Delta E \sim m \tag{2.15}$$

can be expected. Therefore, in a relativistic theory, the fluctuations of the energy are enough to allow the creation of particles out of the vacuum. In the case of a spin-$\frac{1}{2}$ particle, the Dirac sea picture shows clearly how, when the energy fluctuations are of order m, electrons from the Dirac sea can be excited to positive energy states, thus creating electron–positron pairs.

It is possible to see how the multiparticle interpretation is forced upon us by relativistic invariance. In non-relativistic quantum mechanics observables are represented by self-adjoint operators that in the Heisenberg picture depend on time. Therefore measurements are localized in time but are global in space. The situation is radically different in the relativistic case. Because no signal can propagate faster

[1] In Section (9.1) we will see how, in the case of the Dirac field, this exponential behavior can be associated with the creation of electron–positron pairs due to a constant electric field (Schwinger effect).

than the speed of light, measurements have to be localized both in time and space. Causality demands then that two measurements carried out in causally-disconnected regions of space-time not interfere with each other. In mathematical terms this means that if $\mathcal{O}_{R_1}$ and $\mathcal{O}_{R_2}$ are the observables associated with two measurements localized in two causally-disconnected regions R_1, R_2 (see Fig. 4), they satisfy

$$[\mathcal{O}_{R_1}, \mathcal{O}_{R_2}] = 0, \qquad \text{if } (x_1 - x_2)^2 < 0, \text{ for all } x_1 \in R_1,\, x_2 \in R_2. \tag{2.16}$$

Hence, in a relativistic theory, the basic operators in the Heisenberg picture must depend on the space-time position x^μ. Unlike the case in non-relativistic quantum mechanics, here the position $\vec{x}$ *is not* an observable, but just a label, similar to the case of time in ordinary quantum mechanics. Causality is then imposed microscopically by requiring

$$[\mathcal{O}(x), \mathcal{O}(y)] = 0, \qquad \text{if } (x - y)^2 < 0. \tag{2.17}$$

A smeared operator $\mathcal{O}_R$ over a space-time region R can then be defined as

$$\mathcal{O}_R = \int d^4x\, \mathcal{O}(x)\, f_R(x) \tag{2.18}$$

where $f_R(x)$ is the characteristic function associated with R,

$$f_R(x) = \begin{cases} 1 & x \in R \\ 0 & x \notin R \end{cases} . \tag{2.19}$$

Equation (2.16) follows now from the microcausality condition (2.17).

Therefore, relativistic invariance forces the introduction of quantum fields. It is only when we insist on keeping a single-particle interpretation that we crash against causality violations. To illustrate the point, let us consider a single-particle wave function $\psi(t, \vec{x})$ that initially is localized in the position $\vec{x} = 0$

$$\psi(0, \vec{x}) = \delta(\vec{x}). \tag{2.20}$$

Evolving this wave function using the Hamiltonian $H = \sqrt{-\nabla^2 + m^2}$ we find that the wave function can be written as

$$\psi(t, \vec{x}) = e^{-it\sqrt{-\nabla^2+m^2}}\delta(\vec{x}) = \int \frac{d^3k}{(2\pi)^3} e^{i\vec{k}\cdot\vec{x} - it\sqrt{k^2+m^2}}. \tag{2.21}$$

Integrating over the angular variables, the wave function can be recast in the form

$$\psi(t, \vec{x}) = \frac{1}{2\pi^2|\vec{x}|} \int_{-\infty}^{\infty} k\, dk\, e^{ik|\vec{x}|}\, e^{-it\sqrt{k^2+m^2}}. \tag{2.22}$$

The resulting integral can be evaluated using the complex integration contour C shown in Fig. 5. The result is that, for any $t > 0$, one finds that $\psi(t, \vec{x}) \neq 0$ for any $\vec{x}$. If we insist on interpreting the wave function $\psi(t, \vec{x})$ as the probability density of finding the particle at the location $\vec{x}$ in the time t we find that the probability leaks out of the light cone, thus violating causality.

3 From classical to quantum fields

We have learned how the consistency of quantum mechanics with special relativity forces us to abandon the single-particle interpretation of the wave function. Instead we have to consider quantum fields whose elementary excitations are associated with particle states, as we will see below.

In any scattering experiment, the only information available to us is the set of quantum numbers associated with the set of free particles in the initial and final states. Ignoring for the moment other

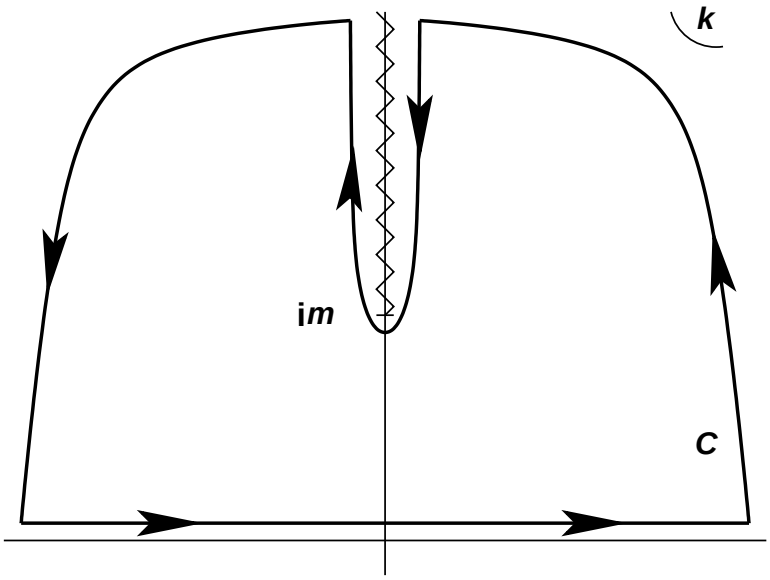


Fig. 5: Complex contour C for the computation of the integral in Eq. (2.22)

quantum numbers like spin and flavor, one-particle states are labelled by the three-momentum $\vec{p}$ and span the single-particle Hilbert space $\mathcal{H}_1$

$$|\vec{p}\rangle \in \mathcal{H}_1, \qquad \langle\vec{p}|\vec{p}\,'\rangle = \delta(\vec{p}-\vec{p}\,') \,. \tag{3.1}$$

The states $\{|\vec{p}\rangle\}$ form a basis of $\mathcal{H}_1$ and therefore satisfy the closure relation

$$\int d^3p \, |\vec{p}\rangle\langle\vec{p}| = \mathbf{1} \,. \tag{3.2}$$

The group of spatial rotations acts unitarily on the states $|\vec{p}\rangle$. This means that for every rotation $R \in \mathrm{SO}(3)$ there is a unitary operator $\mathcal{U}(R)$ such that

$$\mathcal{U}(R)|\vec{p}\rangle = |R\vec{p}\rangle \tag{3.3}$$

where $R\vec{p}$ represents the action of the rotation on the vector $\vec{k}$, $(R\vec{p})^i = R^i{}_j k^j$. Using a spectral decomposition, the momentum operator $\widehat{P}^i$ can be written as

$$\widehat{P}^i = \int d^3p \, |\vec{p}\rangle \, p^i \, \langle\vec{p}| \,. \tag{3.4}$$

With the help of Eq. (3.3) it is straightforward to check that the momentum operator transforms as a vector under rotations:

$$\mathcal{U}(R)^{-1} \, \widehat{P}^i \, \mathcal{U}(R) = \int d^3p \, |R^{-1}\vec{p}\rangle \; p^i \; \langle R^{-1}\vec{p}| = R^i{}_j \widehat{P}^j, \tag{3.5}$$

where we have used that the integration measure is invariant under $\mathrm{SO}(3)$.

Since, as we argued above, we are forced to deal with multiparticle states, it is convenient to introduce creation–annihilation operators associated with a single-particle state of momentum $\vec{p}$

$$[a(\vec{p}), a^\dagger(\vec{p}\,')] = \delta(\vec{p}-\vec{p}\,'), \qquad [a(\vec{p}), a(\vec{p}\,')] = [a^\dagger(\vec{p}), a^\dagger(\vec{p}\,')] = 0, \tag{3.6}$$

such that the state $|\vec{p}\rangle$ is created out of the Fock space vacuum $|0\rangle$ (normalized such that $\langle 0|0\rangle = 1$) by the action of a creation operator $a^\dagger(\vec{p})$

$$|\vec{p}\rangle = a^\dagger(\vec{p})|0\rangle, \qquad a(\vec{p})|0\rangle = 0 \;\; \forall\vec{p}. \tag{3.7}$$

Covariance under spatial rotations is all we need if we are interested in a non-relativistic theory. However, in a relativistic quantum field theory we must preserve more than $\mathrm{SO}(3)$, actually we need the expressions to be covariant under the full Poincaré group $\mathrm{ISO}(1,3)$ consisting of spatial rotations, boosts and space-time translations. Therefore, in order to build the Fock space of the theory we need two key ingredients: first an invariant normalization for the states, since we want a normalized state in one reference frame to be normalized in any other inertial frame. And secondly a relativistic invariant integration measure in momentum space, so the spectral decomposition of operators is covariant under the full Poincaré group.

Let us begin with the invariant measure. Given an invariant function $f(p)$ of the four-momentum p^μ of a particle of mass m with positive energy $p^0 > 0$, there is an integration measure which is invariant under proper Lorentz transformations[2]

$$\int \frac{d^4p}{(2\pi)^4}\,(2\pi)\delta(p^2-m^2)\,\theta(p^0)\,f(p), \tag{3.8}$$

where $\theta(x)$ represent the Heaviside step function. The integration over p^0 can be easily done using the δ-function identity

$$\delta[f(x)] = \sum_{x_i=\text{zeros of } f} \frac{1}{|f'(x_i)|}\delta(x-x_i), \tag{3.9}$$

which in our case implies that

$$\delta(p^2-m^2) = \frac{1}{2p^0}\,\delta\left(p^0-\sqrt{\vec{p}^{\,2}+m^2}\right) + \frac{1}{2p^0}\,\delta\left(p^0+\sqrt{\vec{p}^{\,2}+m^2}\right). \tag{3.10}$$

The second term in the previous expression corresponds to states with negative energy and therefore does not contribute to the integral. We can then write

$$\int \frac{d^4p}{(2\pi)^4}\,(2\pi)\delta(p^2-m^2)\,\theta(p^0)\,f(p) = \int \frac{d^3p}{(2\pi)^3}\frac{1}{2\sqrt{\vec{p}^{\,2}+m^2}}\,f\left(\sqrt{\vec{p}^{\,2}+m^2},\vec{p}\right). \tag{3.11}$$

Hence, the relativistic invariant measure is given by

$$\int \frac{d^3p}{(2\pi)^3}\frac{1}{2\omega_p} \qquad \text{with} \qquad \omega_p \equiv \sqrt{\vec{p}^{\,2}+m^2}. \tag{3.12}$$

Once we have an invariant measure the next step is to find an invariant normalization for the states. We work with a basis $\{|p\rangle\}$ of eigenstates of the four-momentum operator $\widehat{P}^\mu$

$$\widehat{P}^0|p\rangle = \omega_p|p\rangle, \qquad\qquad \widehat{P}^i|p\rangle = \vec{p}^{\,i}|p\rangle. \tag{3.13}$$

Since the states $|p\rangle$ are eigenstates of the three-momentum operator we can express them in terms of the non-relativistic states $|\vec{p}\rangle$ that we introduced in Eq. (3.1)

$$|p\rangle = N(\vec{p})|\vec{p}\rangle \tag{3.14}$$

with $N(\vec{p})$ a normalization to be determined now. The states $\{|p\rangle\}$ form a complete basis, so they should satisfy the Lorentz-invariant closure relation

$$\int \frac{d^4p}{(2\pi)^4}\,(2\pi)\delta(p^2-m^2)\,\theta(p^0)\,|p\rangle\,\langle p| = \mathbf{1}\,. \tag{3.15}$$

[2]The factors of 2π are introduced for later convenience.

At the same time, this closure relation can be expressed, using Eq. (3.14), in terms of the non-relativistic basis of states $\{|\vec{p}\rangle\}$ as

$$\int \frac{d^4p}{(2\pi)^4}\,(2\pi)\delta(p^2-m^2)\,\theta(p^0)\,|p\rangle\,\langle p| = \int \frac{d^3p}{(2\pi)^3}\frac{1}{2\omega_p}|N(p)|^2\,|\vec{p}\rangle\,\langle\vec{p}|. \tag{3.16}$$

Using now Eq. (3.4) for the non-relativistic states, expression (3.15) follows provided

$$|N(\vec{p})|^2 = (2\pi)^3\,(2\omega_p). \tag{3.17}$$

Taking the overall phase in Eq. (3.14) so that $N(p)$ is real, we define the Lorentz-invariant states $|p\rangle$ as

$$|p\rangle = (2\pi)^{\frac{3}{2}}\,\sqrt{2\omega_p}\,|\vec{p}\rangle, \tag{3.18}$$

and given the normalization of $|\vec{p}\rangle$ we find the normalization of the relativistic states to be

$$\langle p|p'\rangle = (2\pi)^3(2\omega_p)\delta(\vec{p}-\vec{p}\,'). \tag{3.19}$$

Although not obvious at first sight, the previous normalization is Lorentz invariant. Although it is not difficult to show this in general, here we consider the simpler case of 1+1 dimensions where the two components (p^0,p^1) of the on-shell momentum can be parametrized in terms of a single hyperbolic angle λ as

$$p^0 = m\cosh\lambda, \qquad\qquad p^1 = m\sinh\lambda. \tag{3.20}$$

Now, the combination $2\omega_p\delta(p^1-p^{1\prime})$ can be written as

$$2\omega_p\delta(p^1-p^{1\prime}) = 2m\cosh\lambda\,\delta(m\sinh\lambda - m\sinh\lambda') = 2\delta(\lambda-\lambda'), \tag{3.21}$$

where we have made use of the property (3.9) of the δ-function. Lorentz transformations in $1+1$ dimensions are labelled by a parameter $\xi\in\mathbb{R}$ and act on the momentum by shifting the hyperbolic angle $\lambda\to\lambda+\xi$. However, Eq. (3.21) is invariant under a common shift of λ and λ', so the whole expression is obviously invariant under Lorentz transformations.

To summarize what we did so far, we have succeed in constructing a Lorentz-covariant basis of states for the one-particle Hilbert space $\mathcal{H}_1$. The generators of the Poincaré group act on the states $|p\rangle$ of the basis as

$$\widehat{P}^\mu|p\rangle = p^\mu|p\rangle, \qquad\qquad \mathcal{U}(\Lambda)|p\rangle = |\Lambda^\mu{}_\nu\,p^\nu\rangle \equiv |\Lambda p\rangle \quad \text{with} \quad \Lambda\in \mathrm{SO}(1,3). \tag{3.22}$$

This is compatible with the Lorentz invariance of the normalization that we have checked above

$$\langle p|p'\rangle = \langle p|\mathcal{U}(\Lambda)^{-1}\mathcal{U}(\Lambda)|p'\rangle = \langle\Lambda p|\Lambda p'\rangle. \tag{3.23}$$

On $\mathcal{H}_1$ the operator $\widehat{P}^\mu$ admits the following spectral representation

$$\widehat{P}^\mu = \int \frac{d^3p}{(2\pi)^3}\frac{1}{2\omega_p}|p\rangle\,p^\mu\,\langle p|\,. \tag{3.24}$$

Using (3.23) and the fact that the measure is invariant under Lorentz transformation, one can easily show that $\widehat{P}^\mu$ transform covariantly under $\mathrm{SO}(1,3)$

$$\mathcal{U}(\Lambda)^{-1}\widehat{P}^\mu\mathcal{U}(\Lambda) = \int \frac{d^3p}{(2\pi)^3}\frac{1}{2\omega_p}|\Lambda^{-1}p\rangle\,p^\mu\,\langle\Lambda^{-1}p| = \Lambda^\mu{}_\nu\widehat{P}^\nu. \tag{3.25}$$

A set of covariant creation–annihilation operators can be constructed now in terms of the operators $a(\vec{p})$, $a^\dagger(\vec{p})$ introduced above

$$\alpha(\vec{p}) \equiv (2\pi)^{\frac{3}{2}}\sqrt{2\omega_p}a(\vec{p}), \qquad \alpha^\dagger(\vec{p}) \equiv (2\pi)^{\frac{3}{2}}\sqrt{2\omega_p}a^\dagger(\vec{p}) \tag{3.26}$$

with the Lorentz-invariant commutation relations

$$\begin{aligned}[\alpha(\vec{p}),\alpha^\dagger(\vec{p}\,')] &= (2\pi)^3(2\omega_p)\delta(\vec{p}-\vec{p}\,'),\\ [\alpha(\vec{p}),\alpha(\vec{p}\,')] &= [\alpha^\dagger(\vec{p}),\alpha^\dagger(\vec{p}\,')] = 0.\end{aligned} \tag{3.27}$$

Particle states are created by acting with any number of creation operators $\alpha(\vec{p})$ on the Poincaré invariant vacuum state $|0\rangle$ satisfying

$$\langle 0|0\rangle = 1, \qquad \widehat{P}^\mu|0\rangle = 0, \qquad \mathcal{U}(\Lambda)|0\rangle = |0\rangle, \quad \forall\Lambda\in \mathrm{SO}(1,3). \tag{3.28}$$

A general one-particle state $|f\rangle \in \mathcal{H}_1$ can be then written as

$$|f\rangle = \int \frac{d^3p}{(2\pi)^3}\frac{1}{2\omega_p}f(\vec{p})\alpha^\dagger(\vec{p})|0\rangle, \tag{3.29}$$

while a n-particle state $|f\rangle \in \mathcal{H}_1^{\otimes n}$ can be expressed as

$$|f\rangle = \int \prod_{i=1}^{n}\frac{d^3p_i}{(2\pi)^3}\frac{1}{2\omega_{p_i}}f(\vec{p}_1,\ldots,\vec{p}_n)\alpha^\dagger(\vec{p}_1)\ldots\alpha^\dagger(\vec{p}_n)|0\rangle. \tag{3.30}$$

That these states are Lorentz invariant can be checked by noticing that from the definition of the creation–annihilation operators follows the transformation

$$\mathcal{U}(\Lambda)\alpha(\vec{p})\mathcal{U}(\Lambda)^\dagger = \alpha(\Lambda\vec{p}) \tag{3.31}$$

and the corresponding one for creation operators.

As we have argued above, the very fact that measurements have to be localized implies the necessity of introducing quantum fields. Here we will consider the simplest case of a scalar quantum field $\phi(x)$ satisfying the following properties:

– **Hermiticity.**

$$\phi^\dagger(x) = \phi(x). \tag{3.32}$$

– **Microcausality.** Since measurements cannot interfere with each other when performed in causally disconnected points of space-time, the commutator of two fields has to vanish outside the relative light-cone

$$[\phi(x),\phi(y)] = 0, \qquad (x-y)^2 < 0. \tag{3.33}$$

– **Translation invariance.**

$$e^{i\widehat{P}\cdot a}\phi(x)e^{-i\widehat{P}\cdot a} = \phi(x-a). \tag{3.34}$$

– **Lorentz invariance.**

$$\mathcal{U}(\Lambda)^\dagger\phi(x)\mathcal{U}(\Lambda) = \phi(\Lambda^{-1}x). \tag{3.35}$$

– **Linearity.** To simplify matters we will also assume that $\phi(x)$ is linear in the creation–annihilation operators $\alpha(\vec{p})$, $\alpha^\dagger(\vec{p})$

$$\phi(x) = \int \frac{d^3p}{(2\pi)^3}\frac{1}{2\omega_p}\left[f(\vec{p},x)\alpha(\vec{p}) + g(\vec{p},x)\alpha^\dagger(\vec{p})\right]. \tag{3.36}$$

Since $\phi(x)$ should be hermitian we are forced to take $f(\vec{p},x)^* = g(\vec{p},x)$. Moreover, $\phi(x)$ satisfies the equations of motion of a free scalar field, $(\partial_\mu\partial^\mu + m^2)\phi(x) = 0$, only if $f(\vec{p},x)$ is a complete basis of solutions of the Klein–Gordon equation. These considerations lead to the expansion

$$\phi(x) = \int \frac{d^3p}{(2\pi)^3}\frac{1}{2\omega_p}\left[e^{-i\omega_p t + i\vec{p}\cdot\vec{x}}\alpha(\vec{p}) + e^{i\omega_p t - i\vec{p}\cdot\vec{x}}\alpha^\dagger(\vec{p})\right]. \tag{3.37}$$

Given the expansion of the scalar field in terms of the creation–annihilation operators it can be checked that $\phi(x)$ and $\partial_t\phi(x)$ satisfy the equal-time canonical commutation relations

$$[\phi(t,\vec{x}), \partial_t\phi(t,\vec{y})] = i\delta(\vec{x}-\vec{y})\ . \tag{3.38}$$

The general commutator $[\phi(x),\phi(y)]$ can also be computed to be

$$[\phi(x),\phi(x')] = i\Delta(x-x'). \tag{3.39}$$

The function $\Delta(x-y)$ is given by

$$\begin{aligned} i\Delta(x-y) &= -\mathrm{Im}\int \frac{d^3p}{(2\pi)^3}\frac{1}{2\omega_p}e^{-i\omega_p(t-t')+i\vec{p}\cdot(\vec{x}-\vec{x}\,')} \\ &= \int \frac{d^4p}{(2\pi)^4}(2\pi)\delta(p^2-m^2)\varepsilon(p^0)e^{-ip\cdot(x-x')}, \end{aligned} \tag{3.40}$$

where $\varepsilon(x)$ is defined as

$$\varepsilon(x) \equiv \theta(x) - \theta(-x) = \left\{ \begin{array}{cc} 1 & x>0 \\ -1 & x<0 \end{array}\right. . \tag{3.41}$$

Using the last expression in Eq. (3.40) it is easy to show that $i\Delta(x-x')$ vanishes when x and x' are space-like separated. Indeed, if $(x-x')^2 < 0$ there is always a reference frame in which both events are simultaneous, and since $i\Delta(x-x')$ is Lorentz invariant we can compute it in this reference frame. In this case $t = t'$ and the exponential in the second line of (3.40) does not depend on p^0. Therefore, the integration over k^0 gives

$$\begin{aligned} \int_{-\infty}^{\infty} dp^0\varepsilon(p^0)\delta(p^2-m^2) &= \int_{-\infty}^{\infty} dp^0\left[\frac{1}{2\omega_p}\varepsilon(p^0)\delta(p^0-\omega_p) + \frac{1}{2\omega_p}\varepsilon(p^0)\delta(p^0+\omega_p)\right] \\ &= \frac{1}{2\omega_p} - \frac{1}{2\omega_p} = 0. \end{aligned} \tag{3.42}$$

So we have concluded that $i\Delta(x-x') = 0$ if $(x-x')^2 < 0$, as required by microcausality. Notice that the situation is completely different when $(x-x')^2 \geq 0$, since in this case the exponential depends on p^0 and the integration over this component of the momentum does not vanish.

3.1 Canonical quantization

So far we have contented ourselves with requiring a number of properties in the quantum scalar field: existence of asymptotic states, locality, microcausality and relativistic invariance. With only these ingredients we have managed to go quite far. The former can also be obtained using canonical quantization. One starts with a classical free scalar field theory in Hamiltonian formalism and obtains the quantum theory by replacing Poisson brackets by commutators. Since this quantization procedure is based on the use of the canonical formalism, which gives time a privileged rôle, it is important to check at the end of the calculation that the resulting quantum theory is Lorentz invariant. In the following we will briefly overview the canonical quantization of the Klein–Gordon scalar field.

The starting point is the action functional $S[\phi(x)]$ which, in the case of a free real scalar field of mass m, is given by

$$S[\phi(x)] \equiv \int d^4x\, \mathcal{L}(\phi, \partial_\mu\phi) = \frac{1}{2}\int d^4x\, \left(\partial_\mu\phi\partial^\mu\phi - m^2\phi^2\right). \tag{3.43}$$

The equations of motion are obtained, as usual, from the Euler–Lagrange equations

$$\partial_\mu\left[\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\right] - \frac{\partial\mathcal{L}}{\partial\phi} = 0 \qquad \Longrightarrow \qquad (\partial_\mu\partial^\mu + m^2)\phi = 0. \tag{3.44}$$

The momentum canonically conjugated to the field $\phi(x)$ is given by

$$\pi(x) \equiv \frac{\partial\mathcal{L}}{\partial(\partial_0\phi)} = \frac{\partial\phi}{\partial t}. \tag{3.45}$$

In the Hamiltonian formalism the physical system is described not in terms of the generalized coordinates and their time derivatives but in terms of the generalized coordinates and their canonically conjugated momenta. This is achieved by a Legendre transformation after which the dynamics of the system is determined by the Hamiltonian function

$$H \equiv \int d^3x \left(\pi\frac{\partial\phi}{\partial t} - \mathcal{L}\right) = \frac{1}{2}\int d^3x \left[\pi^2 + (\vec{\nabla}\phi)^2 + m^2\right]. \tag{3.46}$$

The equations of motion can be written in terms of the Poisson brackets. Given two functional $A[\phi,\pi]$, $B[\phi,\pi]$ of the canonical variables

$$A[\phi,\pi] = \int d^3x\mathcal{A}(\phi,\pi), \qquad B[\phi,\pi] = \int d^3x\mathcal{B}(\phi,\pi). \tag{3.47}$$

Their Poisson bracket is defined by

$$\{A,B\} \equiv \int d^3x \left[\frac{\delta A}{\delta\phi}\frac{\delta B}{\delta\pi} - \frac{\delta A}{\delta\pi}\frac{\delta B}{\delta\phi}\right], \tag{3.48}$$

where $\frac{\delta}{\delta\phi}$ denotes the functional derivative defined as

$$\frac{\delta A}{\delta\phi} \equiv \frac{\partial\mathcal{A}}{\partial\phi} - \partial_\mu\left[\frac{\partial\mathcal{A}}{\partial(\partial_\mu\phi)}\right]. \tag{3.49}$$

Then, the canonically conjugated fields satisfy the following equal time Poisson brackets

$$\begin{aligned} \{\phi(t,\vec{x}),\phi(t,\vec{x}\,')\} &= \{\pi(t,\vec{x}),\pi(t,\vec{x}\,')\} = 0, \\ \{\phi(t,\vec{x}),\pi(t,\vec{x}\,')\} &= \delta(\vec{x}-\vec{x}\,'). \end{aligned} \tag{3.50}$$

Canonical quantization proceeds now by replacing classical fields with operators and Poisson brackets with commutators according to the rule

$$i\{\cdot,\cdot\} \longrightarrow [\cdot,\cdot]. \tag{3.51}$$

In the case of the scalar field, a general solution of the field equations (3.44) can be obtained by working with the Fourier transform

$$(\partial_\mu\partial^\mu + m^2)\phi(x) = 0 \qquad \Longrightarrow \qquad (-p^2+m^2)\widetilde{\phi}(p) = 0, \tag{3.52}$$

whose general solution can be written as[3]

$$\begin{aligned}\phi(x) &= \int \frac{d^4p}{(2\pi)^4}(2\pi)\delta(p^2-m^2)\theta(p^0)\left[\alpha(p)e^{-ip\cdot x}+\alpha(p)^*e^{ip\cdot x}\right] \\ &= \int \frac{d^3p}{(2\pi)^3}\frac{1}{2\omega_p}\left[\alpha(\vec{p})e^{-i\omega_p t+\vec{p}\cdot\vec{x}}+\alpha(\vec{p})^*e^{i\omega_p t-\vec{p}\cdot\vec{x}}\right]\end{aligned} \tag{3.53}$$

and we have required $\phi(x)$ to be real. The conjugate momentum is

$$\pi(x) = -\frac{i}{2}\int \frac{d^3p}{(2\pi)^3}\left[\alpha(\vec{p})e^{-i\omega_p t+\vec{p}\cdot\vec{x}}+\alpha(\vec{p})^*e^{i\omega_p t-\vec{p}\cdot\vec{x}}\right]. \tag{3.54}$$

Now $\phi(x)$ and $\pi(x)$ are promoted to operators by replacing the functions $\alpha(\vec{p})$, $\alpha(\vec{p})^*$ by the corresponding operators

$$\alpha(\vec{p}) \longrightarrow \widehat{\alpha}(\vec{p}), \qquad\qquad \alpha(\vec{p})^* \longrightarrow \widehat{\alpha}^\dagger(\vec{p}). \tag{3.55}$$

Moreover, demanding $[\phi(t,\vec{x}),\pi(t,\vec{x}\,')] = i\delta(\vec{x}-\vec{x}\,')$ forces the operators $\widehat{\alpha}(\vec{p})$, $\widehat{\alpha}(\vec{p})^\dagger$ to have the commutation relations found in Eq. (3.27). Therefore they are identified as a set of creation–annihilation operators creating states with well-defined momentum $\vec{p}$ out of the vacuum $|0\rangle$. In the canonical quantization formalism the concept of particle appears as a result of the quantization of a classical field.

Knowing the expressions of $\widehat{\phi}$ and $\widehat{\pi}$ in terms of the creation–annihilation operators we can proceed to evaluate the Hamiltonian operator. After a simple calculation one arrives at the expression

$$\widehat{H} = \int d^3p\,\left[\omega_p\widehat{\alpha}^\dagger(\vec{p})\widehat{\alpha}(\vec{p}) + \frac{1}{2}\omega_p\,\delta(\vec{0})\right]. \tag{3.56}$$

The first term has a simple physical interpretation since $\widehat{\alpha}^\dagger(\vec{p})\widehat{\alpha}(\vec{p})$ is the number operator of particles with momentum $\vec{p}$. The second divergent term can be eliminated if we defined the normal-ordered Hamiltonian $:\widehat{H}:$ with the vacuum energy subtracted

$$:\widehat{H}:\equiv \widehat{H} - \langle 0|\widehat{H}|0\rangle = \int d^3p\,\omega_p\,\widehat{\alpha}^\dagger(\vec{p})\,\widehat{\alpha}(\vec{p})\,. \tag{3.57}$$

It is interesting to try to make sense of the divergent term in Eq. (3.56). This term has two sources of divergence. One is associated with the delta function evaluated at zero coming from the fact that we are working in a infinite volume. It can be regularized for large but finite volume by replacing $\delta(\vec{0}) \sim V$. Hence, it is of infrared origin. The second one comes from the integration of ω_p at large values of the momentum and it is then an ultraviolet divergence. The infrared divergence can be regularized by considering the scalar field to be living in a box of finite volume V. In this case the vacuum energy is

$$E_{\rm vac} \equiv \langle 0|\widehat{H}|0\rangle = \sum_{\vec{p}}\frac{1}{2}\omega_p. \tag{3.58}$$

[3]In momentum space, the general solution to this equation is $\widetilde{\phi}(p) = f(p)\delta(p^2-m^2)$, with $f(p)$ a completely general function of p^μ. The solution in position space is obtained by inverse Fourier transform.

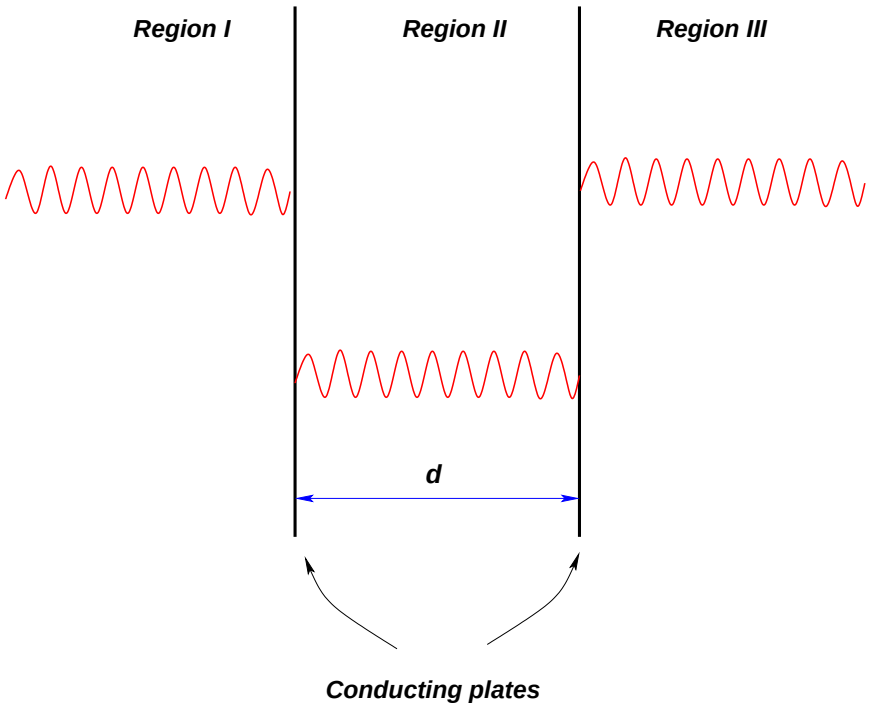


Fig. 6: Illustration of the Casimir effect. In regions I and II the spectrum of modes of the momentum $p_\perp$ is continuous, while in the space between the plates (region II) it is quantized in units of π/d.

Written in this way the interpretation of the vacuum energy is straightforward. A free scalar quantum field can be seen as an infinite collection of harmonic oscillators per unit volume, each one labelled by $\vec{p}$. Even if those oscillators are not excited, they contribute to the vacuum energy with their zero-point energy, given by $\frac{1}{2}\omega_p$. This vacuum contribution to the energy adds up to infinity even if we work at finite volume, since even then there are modes with arbitrarily high momentum contributing to the sum, $p_i = \frac{n_i \pi}{L_i}$, with L_i the sides of the box of volume V and n_i an integer. Hence, this divergence is of ultraviolet origin.

3.2 The Casimir effect

The presence of a vacuum energy is not characteristic of the scalar field. It is also present in other cases, in particular in quantum electrodynamics. Although one might be tempted to discard this infinite contribution to the energy of the vacuum as unphysical, it has observable consequences. In 1948 Hendrik Casimir pointed out [15] that although a formally divergent vacuum energy would not be observable, any variation in this energy would be (see [16] for comprehensive reviews).

To show this he devised the following experiment. Consider a couple of infinite, perfectly conducting plates placed parallel to each other at a distance d (see Fig. 6). Because the conducting plates fix the boundary condition of the vacuum modes of the electromagnetic field these are discrete in between the plates (region II), while outside there is a continuous spectrum of modes (regions I and III). In order to calculate the force between the plates we can take the vacuum energy of the electromagnetic field as given by the contribution of two scalar fields corresponding to the two polarizations of the photon. Therefore we can use the formulas derived above.

A naive calculation of the vacuum energy in this system gives a divergent result. This infinity can be removed, however, by substracting the vacuum energy corresponding to the situation where the plates are removed

$$E(d)_{\rm reg} = E(d)_{\rm vac} - E(\infty)_{\rm vac} \ . \tag{3.59}$$

This substraction cancels the contribution of the modes outside the plates. Because of the boundary conditions imposed by the plates the momentum of the modes perpendicular to the plates are quantized according to $p_\perp = \frac{n\pi}{d}$, with n a non-negative integer. If we consider that the size of the plates is much larger than their separation d we can take the momenta parallel to the plates $\vec{p}_\parallel$ as continuous. For $n > 0$ we have two polarizations for each vacuum mode of the electromagnetic field, each contributing like $\frac{1}{2}\sqrt{\vec{p}_\parallel^{\,2} + p_\perp^2}$ to the vacuum energy. On the other hand, when $p_\perp = 0$ the corresponding modes of the field are effectively (2+1)-dimensional and therefore there is only one polarization. Keeping this in mind, we can write

$$\begin{aligned} E(d)_{\rm reg} &= S\int \frac{d^2p_\parallel}{(2\pi)^2}\frac{1}{2}|\vec{p}_\parallel| + 2S\int \frac{d^2p_\parallel}{(2\pi)^2}\sum_{n=1}^{\infty}\frac{1}{2}\sqrt{\vec{p}_\parallel^{\,2} + \left(\frac{n\pi}{d}\right)^2} \\ &- 2Sd\int \frac{d^3p}{(2\pi)^3}\frac{1}{2}|\vec{p}| \end{aligned} \tag{3.60}$$

where S is the area of the plates. The factors of 2 take into account the two propagating degrees of freedom of the electromagnetic field, as discussed above. In order to ensure the convergence of integrals and infinite sums we can introduce an exponential damping factor[4]

$$\begin{aligned} E(d)_{\rm reg} &= \frac{1}{2}S\int \frac{d^2p_\perp}{(2\pi)^2}e^{-\frac{1}{\Lambda}|\vec{p}_\parallel|}|\vec{p}_\parallel| + S\sum_{n=1}^{\infty}\int \frac{d^2p_\parallel}{(2\pi)^2}e^{-\frac{1}{\Lambda}\sqrt{\vec{p}_\parallel^{\,2}+\left(\frac{n\pi}{d}\right)^2}}\sqrt{\vec{p}_\parallel^{\,2} + \left(\frac{n\pi}{d}\right)^2} \\ &- Sd\int_{-\infty}^{\infty}\frac{dp_\perp}{2\pi}\int \frac{d^2p_\parallel}{(2\pi)^2}e^{-\frac{1}{\Lambda}\sqrt{\vec{p}_\parallel^{\,2}+p_\perp^2}}\sqrt{\vec{p}_\parallel^{\,2} + p_\perp^2} \end{aligned} \tag{3.61}$$

where Λ is an ultraviolet cutoff. It is now straightforward to see that if we define the function

$$F(x) = \frac{1}{2\pi}\int_0^\infty y\,dy\,e^{-\frac{1}{\Lambda}\sqrt{y^2+\left(\frac{x\pi}{d}\right)^2}}\sqrt{y^2+\left(\frac{x\pi}{d}\right)^2} = \frac{1}{4\pi}\int_{\left(\frac{x\pi}{d}\right)^2}^\infty dz\,e^{-\frac{\sqrt{z}}{\Lambda}}\sqrt{z} \tag{3.62}$$

the regularized vacuum energy can be written as

$$E(d)_{\rm reg} = S\left[\frac{1}{2}F(0) + \sum_{n=1}^{\infty}F(n) - \int_0^\infty dx\,F(x)\right]\,. \tag{3.63}$$

This expression can be evaluated using the Euler–MacLaurin formula [17]

$$\begin{aligned} \sum_{n=1}^{\infty}F(n) - \int_0^\infty dx\,F(x) &= -\frac{1}{2}\left[F(0)+F(\infty)\right] + \frac{1}{12}\left[F'(\infty) - F'(0)\right] \\ &- \frac{1}{720}\left[F'''(\infty) - F'''(0)\right] + \dots \end{aligned} \tag{3.64}$$

Since for our function $F(\infty) = F'(\infty) = F'''(\infty) = 0$ and $F'(0) = 0$, the value of $E(d)_{\rm reg}$ is determined by $F'''(0)$. Computing this term and removing the ultraviolet cutoff, $\Lambda \to \infty$ we find the result

$$E(d)_{\rm reg} = \frac{S}{720}F'''(0) = -\frac{\pi^2 S}{720 d^3}. \tag{3.65}$$

Then, the force per unit area between the plates is given by

$$P_{\rm Casimir} = -\frac{\pi^2}{240}\frac{1}{d^4}. \tag{3.66}$$

The minus sign shows that the force between the plates is attractive. This is the so-called Casimir effect. It was experimentally measured in 1958 by Sparnaay [18] and since then the Casimir effect has been checked with better and better precision in a variety of situations [16].

[4]Actually, one could introduce any cutoff function $f(p_\perp^2 + p_\parallel^2)$ going to zero fast enough as $p_\perp$, $p_\parallel \to \infty$. The result is independent of the particular function used in the calculation.

4 Theories and Lagrangians

Up to this point we have used a scalar field to illustrate our discussion of the quantization procedure. However, nature is richer than that and it is necessary to consider other fields with more complicated behavior under Lorentz transformations. Before considering other fields we pause and study the properties of the Lorentz group.

4.1 Representations of the Lorentz group

In four dimensions the Lorentz group has six generators. Three of them correspond to the generators of the group of rotations in three dimensions SO(3). In terms of the generators J_i of the group a finite rotation of angle φ with respect to an axis determined by a unitary vector $\vec{e}$ can be written as

$$R(\vec{e},\varphi) = e^{-i\varphi\,\vec{e}\cdot\vec{J}}, \qquad \vec{J} = \begin{pmatrix} J_1 \\ J_2 \\ J_3 \end{pmatrix}. \tag{4.1}$$

The other three generators of the Lorentz group are associated with boosts M_i along the three spatial directions. A boost with rapidity λ along a direction $\vec{u}$ is given by

$$B(\vec{u},\lambda) = e^{-i\lambda\,\vec{u}\cdot\vec{M}}, \qquad \vec{M} = \begin{pmatrix} M_1 \\ M_2 \\ M_3 \end{pmatrix}. \tag{4.2}$$

These six generators satisfy the algebra

$$\begin{aligned} [J_i, J_j] &= i\epsilon_{ijk} J_k, \\ [J_i, M_k] &= i\epsilon_{ijk} M_k, \\ [M_i, M_j] &= -i\epsilon_{ijk} J_k\,. \end{aligned} \tag{4.3}$$

The first line corresponds to the commutation relations of SO(3), while the second one implies that the generators of the boosts transform like a vector under rotations.

At first sight, to find representations of the algebra (4.3) might seem difficult. The problem is greatly simplified if we consider the following combination of the generators

$$J_k^{\pm} = \frac{1}{2}(J_k \pm iM_k). \tag{4.4}$$

Using (4.3) it is easy to prove that the new generators $J_k^{\pm}$ satisfy the algebra

$$\begin{aligned} [J_i^{\pm}, J_j^{\pm}] &= i\epsilon_{ijk} J_k^{\pm}, \\ [J_i^{+}, J_j^{-}] &= 0. \end{aligned} \tag{4.5}$$

Then the Lorentz algebra (4.3) is actually equivalent to two copies of the algebra of $\mathrm{SU}(2) \approx \mathrm{SO}(3)$. Therefore the irreducible representations of the Lorentz group can be obtained from the well-known representations of SU(2). Since the latter ones are labelled by the spin $\mathbf{s} = k + \frac{1}{2}, k$ (with $k \in \mathbb{N}$), any representation of the Lorentz algebra can be identified by specifying $(\mathbf{s}_+, \mathbf{s}_-)$, the spins of the representations of the two copies of SU(2) that made up the algebra (4.3).

To get familiar with this way of labelling the representations of the Lorentz group we study some particular examples. Let us start with the simplest one $(\mathbf{s}_+, \mathbf{s}_-) = (\mathbf{0}, \mathbf{0})$. This state is a singlet under $J_i^{\pm}$ and therefore also under rotations and boosts. Therefore we have a scalar.

The next interesting cases are $(\frac{\mathbf{1}}{\mathbf{2}}, \mathbf{0})$ and $(\mathbf{0}, \frac{\mathbf{1}}{\mathbf{2}})$. They correspond respectively to a right-handed and a left-handed Weyl spinor. Their properties will be studied in more detail below. In the case of

Table 1: Representations of the Lorentz group

Representation	Type of field
$(\mathbf{0},\mathbf{0})$	Scalar
$(\frac{\mathbf{1}}{\mathbf{2}},\mathbf{0})$	Right-handed spinor
$(\mathbf{0},\frac{\mathbf{1}}{\mathbf{2}})$	Left-handed spinor
$(\frac{\mathbf{1}}{\mathbf{2}},\frac{\mathbf{1}}{\mathbf{2}})$	Vector
$(\mathbf{1},\mathbf{0})$	Selfdual antisymmetric 2-tensor
$(\mathbf{0},\mathbf{1})$	Anti-selfdual antisymmetric 2-tensor

$(\frac{\mathbf{1}}{\mathbf{2}},\frac{\mathbf{1}}{\mathbf{2}})$, since from Eq. (4.4) we see that $J_i = J_i^+ + J_i^-$ the rules of addition of angular momentum tell us that there are two states, one of them transforming as a vector and another one as a scalar under three-dimensional rotations. Actually, a more detailed analysis shows that the singlet state corresponds to the time component of a vector and the states combine to form a vector under the Lorentz group.

There are also more 'exotic' representations. For example we can consider the $(\mathbf{1},\mathbf{0})$ and $(\mathbf{0},\mathbf{1})$ representations corresponding respectively to a selfdual and an anti-selfdual rank-two antisymmetric tensor. In Table 1 we summarize the previous discussion.

To conclude our discussion of the representations of the Lorentz group we notice that under a parity transformation the generators of SO(1,3) transform as

$$P : J_i \longrightarrow J_i, \qquad P : M_i \longrightarrow -M_i \tag{4.6}$$

this means that $P : J_i^{\pm} \longrightarrow J_i^{\mp}$ and therefore a representation $(\mathbf{s_1},\mathbf{s_2})$ is transformed into $(\mathbf{s_2},\mathbf{s_1})$. This means that, for example, a vector $(\frac{\mathbf{1}}{\mathbf{2}},\frac{\mathbf{1}}{\mathbf{2}})$ is invariant under parity, whereas a left-handed Weyl spinor $(\frac{\mathbf{1}}{\mathbf{2}},\mathbf{0})$ transforms into a right-handed one $(\mathbf{0},\frac{\mathbf{1}}{\mathbf{2}})$ and vice versa.

4.2 Spinors

Weyl spinors. Let us go back to the two spinor representations of the Lorentz group, namely $(\frac{\mathbf{1}}{\mathbf{2}},\mathbf{0})$ and $(\mathbf{0},\frac{\mathbf{1}}{\mathbf{2}})$. These representations can be explicitly constructed using the Pauli matrices as

$$\begin{aligned} J_i^+ &= \frac{1}{2}\sigma^i, \qquad J_i^- = 0 \qquad \text{for} \quad (\tfrac{\mathbf{1}}{\mathbf{2}},\mathbf{0}), \\ J_i^+ &= 0, \qquad J_i^- = \frac{1}{2}\sigma^i \qquad \text{for} \quad (\mathbf{0},\tfrac{\mathbf{1}}{\mathbf{2}}). \end{aligned} \tag{4.7}$$

We denote by $u_\pm$ a complex two-component object that transforms in the representation $\mathbf{s}_\pm = \frac{1}{2}$ of $J_\pm^i$. If we define $\sigma_\pm^\mu = (\mathbf{1}, \pm\sigma^i)$ we can construct the following vector quantities

$$u_+^\dagger \sigma_+^\mu u_+, \qquad u_-^\dagger \sigma_-^\mu u_-. \tag{4.8}$$

Notice that since $(J_i^\pm)^\dagger = J_i^\mp$ the hermitian conjugated fields $u_\pm^\dagger$ are in the $(\mathbf{0},\frac{\mathbf{1}}{\mathbf{2}})$ and $(\frac{\mathbf{1}}{\mathbf{2}},\mathbf{0})$ respectively.

To construct a free Lagrangian for the fields $u_\pm$ we have to look for quadratic combinations of the fields that are Lorentz scalars. If we also demand invariance under global phase rotations

$$u_\pm \longrightarrow e^{i\theta} u_\pm \tag{4.9}$$

we are left with just one possibility up to a sign

$$\mathcal{L}^\pm_{\text{Weyl}} = i u_\pm^\dagger \left(\partial_t \pm \vec{\sigma}\cdot\vec{\nabla}\right) u_\pm = i u_\pm^\dagger \sigma_\pm^\mu \partial_\mu u_\pm. \tag{4.10}$$

This is the Weyl Lagrangian. In order to grasp the physical meaning of the spinors $u_\pm$ we write the equations of motion

$$\left(\partial_0 \pm \vec{\sigma}\cdot\vec{\nabla}\right) u_\pm = 0. \tag{4.11}$$

Multiplying this equation on the left by $\left(\partial_0 \mp \vec{\sigma}\cdot\vec{\nabla}\right)$ and applying the algebraic properties of the Pauli matrices we conclude that $u_\pm$ satisfies the massless Klein–Gordon equation

$$\partial_\mu \partial^\mu \, u_\pm = 0, \tag{4.12}$$

whose solutions are

$$u_\pm(x) = u_\pm(k) e^{-ik\cdot x}, \qquad \text{with} \quad k^0 = |\vec{k}|. \tag{4.13}$$

Plugging these solutions back into the equations of motion (4.11) we find

$$\left(|\vec{k}| \mp \vec{k}\cdot\vec{\sigma}\right) u_\pm = 0, \tag{4.14}$$

which implies

$$\begin{aligned} u_+ : \qquad & \frac{\vec{\sigma}\cdot\vec{k}}{|\vec{k}|} = 1, \\ u_- : \qquad & \frac{\vec{\sigma}\cdot\vec{k}}{|\vec{k}|} = -1. \end{aligned} \tag{4.15}$$

Since the spin operator is defined as $\vec{s} = \frac{1}{2}\vec{\sigma}$, the previous expressions give the chirality of the states with wave function $u_\pm$, i.e., the projection of spin along the momentum of the particle. Therefore we conclude that u_+ is a Weyl spinor of positive helicity $\lambda = \frac{1}{2}$, while u_- has negative helicity $\lambda = -\frac{1}{2}$. This agrees with our assertion that the representation $(\mathbf{\frac{1}{2}}, \mathbf{0})$ corresponds to a right-handed Weyl fermion (positive chirality) whereas $(\mathbf{0}, \mathbf{\frac{1}{2}})$ is a left-handed Weyl fermion (negative chirality). For example, in the Standard Model neutrinos are left-handed Weyl spinors and therefore transform in the representation $(\mathbf{0}, \mathbf{\frac{1}{2}})$ of the Lorentz group.

Nevertheless, it is possible that we were too restrictive in constructing the Weyl Lagrangian (4.10). There we constructed the invariants from the vector bilinears (4.8) corresponding to the product representations

$$(\mathbf{\tfrac{1}{2}}, \mathbf{\tfrac{1}{2}}) = (\mathbf{\tfrac{1}{2}}, \mathbf{0}) \otimes (\mathbf{0}, \mathbf{\tfrac{1}{2}}) \quad \text{and} \quad (\tfrac{1}{2}, \tfrac{1}{2}) = (\mathbf{0}, \mathbf{\tfrac{1}{2}}) \otimes (\mathbf{\tfrac{1}{2}}, \mathbf{0}). \tag{4.16}$$

In particular our insistence in demanding the Lagrangian to be invariant under the global symmetry $u_\pm \to e^{i\theta} u_\pm$ rules out the scalar term that appears in the product representations

$$(\mathbf{\tfrac{1}{2}}, \mathbf{0}) \otimes (\mathbf{\tfrac{1}{2}}, \mathbf{0}) = (\mathbf{1}, \mathbf{0}) \oplus (\mathbf{0}, \mathbf{0}), \qquad (\mathbf{0}, \mathbf{\tfrac{1}{2}}) \otimes (\mathbf{0}, \mathbf{\tfrac{1}{2}}) = (\mathbf{0}, \mathbf{1}) \oplus (\mathbf{0}, \mathbf{0}). \tag{4.17}$$

The singlet representations corresponds to the antisymmetric combinations

$$\epsilon_{ab} u^a_\pm u^b_\pm, \tag{4.18}$$

where ϵ_{ab} is the antisymmetric symbol $\epsilon_{12} = -\epsilon_{21} = 1$.

At first sight it might seem that the term (4.18) vanishes identically because of the antisymmetry of the ϵ-symbol. However, we should keep in mind that the spin-statistic theorem (more on this later) demands that fields with half-integer spin have to satisfy the Fermi–Dirac statistics and therefore satisfy anticommutation relations, whereas fields of integer spin follow the statistics of Bose–Einstein and, as a

consequence, quantization replaces Poisson brackets by commutators. This implies that the components of the Weyl fermions $u_\pm$ are anticommuting Grassmann fields

$$u_\pm^a u_\pm^b + u_\pm^b u_\pm^a = 0. \tag{4.19}$$

It is important to realize that, strictly speaking, fermions (i.e., objects that satisfy the Fermi–Dirac statistics) do not exist classically. The reason is that they satisfy the Pauli exclusion principle and therefore each quantum state can be occupied, at most, by one fermion. Therefore the naïve definition of the classical limit as a limit of large occupation numbers cannot be applied. Fermion fields do not really make sense classically.

Since the combination (4.18) does not vanish and we can construct a new Lagrangian

$$\mathcal{L}^\pm_{\text{Weyl}} = i u_\pm^\dagger \sigma_\pm^\mu \partial_\mu u_\pm + \frac{1}{2} m \epsilon_{ab} u_\pm^a u_\pm^b + \text{h.c.} \tag{4.20}$$

This mass term, called of Majorana type, is allowed if we do not worry about breaking the global U(1) symmetry $u_\pm \to e^{i\theta} u_\pm$. This is not the case, for example, of charged chiral fermions, since the Majorana mass violates the conservation of electric charge or any other gauge U(1) charge. In the Standard Model, however, there is no such problem if we introduce Majorana masses for right-handed neutrinos, since they are singlet under all Standard Model gauge groups. Such a term will, however, break the global U(1) lepton number charge because the operator $\epsilon_{ab}\nu_R^a \nu_R^b$ changes the lepton number by two units

Dirac spinors. We have seen that parity interchanges the representations $(\mathbf{\frac{1}{2}}, \mathbf{0})$ and $(\mathbf{0}, \mathbf{\frac{1}{2}})$, i.e., it changes right-handed with left-handed fermions

$$P : u_\pm \longrightarrow u_\mp. \tag{4.21}$$

An obvious way to build a parity-invariant theory is to introduce a pair or Weyl fermions u_+ and u_+. Actually, these two fields can be combined in a single four-component spinor

$$\psi = \begin{pmatrix} u_+ \\ u_- \end{pmatrix} \tag{4.22}$$

transforming in the reducible representation $(\mathbf{\frac{1}{2}}, \mathbf{0}) \oplus (\mathbf{0}, \mathbf{\frac{1}{2}})$.

Since now we have both u_+ and u_- simultaneously at our disposal the equations of motion for $u_\pm$, $i\sigma_\pm^\mu \partial_\mu u_\pm = 0$ can be modified, while keeping them linear, to

$$\left.\begin{aligned} i\sigma_+^\mu \partial_\mu u_+ &= m u_- \\ i\sigma_-^\mu \partial_\mu u_- &= m u_+ \end{aligned}\right\} \quad \Longrightarrow \quad i \begin{pmatrix} \sigma_+^\mu & 0 \\ 0 & \sigma_-^\mu \end{pmatrix} \partial_\mu \psi = m \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix} \psi. \tag{4.23}$$

These equations of motion can be derived from the Lagrangian density

$$\mathcal{L}_{\text{Dirac}} = i\psi^\dagger \begin{pmatrix} \sigma_+^\mu & 0 \\ 0 & \sigma_-^\mu \end{pmatrix} \partial_\mu \psi - m\psi^\dagger \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix} \psi. \tag{4.24}$$

To simplify the notation it is useful to define the Dirac γ-matrices as

$$\gamma^\mu = \begin{pmatrix} 0 & \sigma_-^\mu \\ \sigma_+^\mu & 0 \end{pmatrix} \tag{4.25}$$

and the Dirac conjugate spinor $\overline{\psi}$

$$\overline{\psi} \equiv \psi^\dagger \gamma^0 = \psi^\dagger \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix}. \tag{4.26}$$

Now the Lagrangian (4.24) can be written in the more compact form

$$\mathcal{L}_{\rm Dirac} = \overline{\psi}\,(i\gamma^\mu\partial_\mu - m)\,\psi. \tag{4.27}$$

The associated equations of motion give the Dirac equation (2.9) with the identifications

$$\gamma^0 = \beta, \qquad \gamma^i = i\alpha^i. \tag{4.28}$$

In addition, the γ-matrices defined in (4.25) satisfy the Clifford algebra

$$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}. \tag{4.29}$$

In D dimensions this algebra admits representations of dimension $2^{[\frac{D}{2}]}$. When D is even the Dirac fermions ψ transform in a reducible representation of the Lorentz group. In the case of interest, $D = 4$, this is easy to prove by defining the matrix

$$\gamma^5 = -i\gamma^0\gamma^1\gamma^2\gamma^3 = \begin{pmatrix} \mathbf{1} & 0 \\ 0 & -\mathbf{1} \end{pmatrix}. \tag{4.30}$$

We see that γ^5 anticommutes with all other γ-matrices. This implies that

$$[\gamma^5, \sigma^{\mu\nu}] = 0, \qquad \text{with} \qquad \sigma^{\mu\nu} = -\frac{i}{4}[\gamma^\mu, \gamma^\nu]. \tag{4.31}$$

Because of Schur's lemma (see Appendix) this implies that the representation of the Lorentz group provided by $\sigma^{\mu\nu}$ is reducible into subspaces spanned by the eigenvectors of γ^5 with the same eigenvalue. If we define the projectors $P_\pm = \frac{1}{2}(1 \pm \gamma^5)$ these subspaces correspond to

$$P_+\psi = \begin{pmatrix} u_+ \\ 0 \end{pmatrix}, \qquad P_-\psi = \begin{pmatrix} 0 \\ u_- \end{pmatrix}, \tag{4.32}$$

which are precisely the Weyl spinors introduced before.

Our next task is to quantize the Dirac Lagrangian. This will be done along the lines used for the Klein–Gordon field, starting with a general solution to the Dirac equation and introducing the corresponding set of creation–annihilation operators. Therefore we start by looking for a complete basis of solutions to the Dirac equation. In the case of the scalar field the elements of the basis were labelled by their four-momentum k^μ. Now, however, we have more degrees of freedom since we are dealing with a spinor which means that we have to add extra labels. Looking back at Eq. (4.15) we can define the helicity operator for a Dirac spinor as

$$\lambda = \frac{1}{2}\vec{\sigma}\cdot\frac{\vec{k}}{|\vec{k}|}\begin{pmatrix} \mathbf{1} & 0 \\ 0 & \mathbf{1} \end{pmatrix}. \tag{4.33}$$

Hence, each element of the basis of functions is labelled by its four-momentum k^μ and the corresponding eigenvalue s of the helicity operator. For positive energy solutions we then propose the ansatz

$$u(k,s)e^{-ik\cdot x}, \qquad s = \pm\frac{1}{2}, \tag{4.34}$$

where $u_\alpha(k,s)$ ($\alpha = 1,\ldots,4$) is a four-component spinor. Substituting in the Dirac equation we obtain

$$(\not{k} - m)u(k,s) = 0. \tag{4.35}$$

In the same way, for negative energy solutions we have

$$v(k,s)e^{ik\cdot x}, \qquad s = \pm\frac{1}{2}, \tag{4.36}$$

where $v(k,s)$ has to satisfy

$$(\not{k}+m)v(k,s)=0. \tag{4.37}$$

Multiplying Eqs. (4.35) and (4.37) on the left respectively by $(\not{k}\mp m)$ we find that the momentum is on the mass shell, $k^2=m^2$. Because of this, the wave function for both positive- and negative-energy solutions can be labeled as well using the three-momentum $\vec{k}$ of the particle, $u(\vec{k},s)$, $v(\vec{k},s)$.

A detailed analysis shows that the functions $u(\vec{k},s)$, $v(\vec{k},s)$ satisfy the properties

$$\begin{aligned}
\overline{u}(\vec{k},s)u(\vec{k},s)&=2m, & \overline{v}(\vec{k},s)v(\vec{k},s)&=-2m,\\
\overline{u}(\vec{k},s)\gamma^\mu u(\vec{k},s)&=2k^\mu, & \overline{v}(\vec{k},s)\gamma^\mu v(\vec{k},s)&=2k^\mu,\\
\sum_{s=\pm\frac{1}{2}} u_\alpha(\vec{k},s)\overline{u}_\beta(\vec{k},s)&=(\not{k}+m)_{\alpha\beta}, & \sum_{s=\pm\frac{1}{2}} v_\alpha(\vec{k},s)\overline{v}_\beta(\vec{k},s)&=(\not{k}-m)_{\alpha\beta},
\end{aligned} \tag{4.38}$$

with $k^0=\omega_k=\sqrt{\vec{k}^2+m^2}$. Then, a general solution to the Dirac equation including creation and annihilation operators can be written as:

$$\widehat{\psi}(t,\vec{x})=\int\frac{d^3k}{(2\pi)^3}\frac{1}{2\omega_k}\sum_{s=\pm\frac{1}{2}}\left[u(\vec{k},s)\,\widehat{b}(\vec{k},s)e^{-i\omega_k t+i\vec{k}\cdot\vec{x}}+v(\vec{k},s)\,\widehat{d}^\dagger(\vec{k},s)e^{i\omega_k t-i\vec{k}\cdot\vec{x}}\right]. \tag{4.39}$$

The operators $\widehat{b}^\dagger_\alpha(\vec{k},s)$, $\widehat{b}_\alpha(\vec{k})$ respectively create and annihilate a spin-$\frac{1}{2}$ particle (for example, an electron) out of the vacuum with momentum $\vec{k}$ and helicity s. Because we are dealing with half-integer spin fields, the spin-statistics theorem forces canonical anticommutation relations for $\widehat{\psi}$ which means that the creation–annihilation operators satisfy the algebra[5]

$$\begin{aligned}
\{b_\alpha(\vec{k},s),b^\dagger_\beta(\vec{k}\,',s')\}&=\delta(\vec{k}-\vec{k}\,')\delta_{\alpha\beta}\delta_{ss'},\\
\{b_\alpha(\vec{k},s),b_\beta(\vec{k}\,',s')\}&=\{b^\dagger_\alpha(\vec{k},s),b^\dagger_\beta(\vec{k}\,',s')\}=0.
\end{aligned} \tag{4.40}$$

In the case of $d_a(\vec{k},s)$, $d^\dagger_a(\vec{k},s)$ we have a set of creation–annihilation operators for the corresponding antiparticles (for example positrons). This is clear if we notice that $d^\dagger_a(\vec{k},s)$ can be seen as the annihilation operator of a negative energy state of the Dirac equation with wave function $v_a(\vec{k},s)$. As we saw, in the Dirac sea picture this corresponds to the creation of an antiparticle out of the vacuum (see Fig. 2). The creation–annihilation operators for antiparticles also satisfy the fermionic algebra

$$\begin{aligned}
\{d_\alpha(\vec{k},s),d^\dagger_\beta(\vec{k}\,',s')\}&=\delta(\vec{k}-\vec{k}\,')\delta_{\alpha\beta}\delta_{ss'},\\
\{d_\alpha(\vec{k},s),d_\beta(\vec{k}\,',s')\}&=\{d^\dagger_\alpha(\vec{k},s),d^\dagger_\beta(\vec{k}\,',s')\}=0.
\end{aligned} \tag{4.41}$$

All other anticommutators between $b_\alpha(\vec{k},s)$, $b^\dagger_\alpha(\vec{k},s)$ and $d_\alpha(\vec{k},s)$, $d^\dagger_\alpha(\vec{k},s)$ vanish.

The Hamiltonian operator for the Dirac field is

$$\widehat{H}=\sum_{s=\pm\frac{1}{2}}\int d^3k\,\left[\omega_k b^\dagger_\alpha(\vec{k},s)b_\alpha(\vec{k},s)-\omega_k d_\alpha(\vec{k},s)d^\dagger_\alpha(\vec{k},s)\right]. \tag{4.42}$$

At this point we realize again the necessity of quantizing the theory using anticommutators instead of commutators. Had we used canonical commutation relations, the second term inside the integral in (4.42) would give the number operator $d^\dagger_\alpha(\vec{k},s)d_\alpha(\vec{k},s)$ with a minus sign in front. As a consequence the Hamiltonian would be unbounded from below and we would be facing again the instability of the

[5]To simplify notation, and since there is no risk of confusion, we now drop the hat to indicate operators.

theory already noticed in the context of relativistic quantum mechanics. However, because of the *anti-commutation* relations (4.41), the Hamiltonian (4.42) takes the form

$$\widehat{H} = \sum_{s=\pm\frac{1}{2}} \int d^3k \, \left[\omega_k b^\dagger_\alpha(\vec{k},s) b_\alpha(\vec{k},s) + \omega_k d^\dagger_\alpha(\vec{k},s) d_\alpha(\vec{k},s) - \omega_k \delta(\vec{0}) \right]. \tag{4.43}$$

As with the scalar field, we find a divergent vacuum energy contribution due to the zero-point energy of the infinite number of harmonic oscillators. Unlike the Klein–Gordon field, the vacuum energy is negative. In Section 9.2 we will see that in certain types of theory called supersymmetric, where the number of bosonic and fermionic degrees of freedom is the same, there is a cancellation of the vacuum energy. The divergent contribution can be removed by the normal order prescription

$$:\widehat{H}:= \sum_{s=\pm\frac{1}{2}} \int d^3k \, \left[\omega_k b^\dagger_\alpha(\vec{k},s) b_\alpha(\vec{k},s) + \omega_k d^\dagger_\alpha(\vec{k},s) d_\alpha(\vec{k},s) \right]. \tag{4.44}$$

Finally, let us mention that using the Dirac equation it is easy to prove that there is a conserved four-current given by

$$j^\mu = \overline{\psi}\gamma^\mu\psi, \qquad \partial_\mu j^\mu = 0. \tag{4.45}$$

As we will explain further in Section 6 this current is associated to the invariance of the Dirac Lagrangian under the global phase shift $\psi \to e^{i\theta}\psi$. In electrodynamics the associated conserved charge

$$Q = e \int d^3x \, j^0 \tag{4.46}$$

is identified with the electric charge.

4.3 Gauge fields

In classical electrodynamics the basic quantities are the electric and magnetic fields $\vec{E}$, $\vec{B}$. These can be expressed in terms of the scalar and vector potential $(\varphi, \vec{A})$

$$\begin{aligned} \vec{E} &= -\vec{\nabla}\varphi - \frac{\partial \vec{A}}{\partial t}, \\ \vec{B} &= \vec{\nabla} \times \vec{A}. \end{aligned} \tag{4.47}$$

From these equations it follows that there is an ambiguity in the definition of the potentials given by the gauge transformations

$$\varphi(t,\vec{x}) \to \varphi(t,\vec{x}) + \frac{\partial}{\partial t}\epsilon(t,\vec{x}), \qquad \vec{A}(t,\vec{x}) \to \vec{A}(t,\vec{x}) + \vec{\nabla}\epsilon(t,\vec{x}). \tag{4.48}$$

Classically $(\varphi, \vec{A})$ are seen as only a convenient way to solve the Maxwell equations, but without physical relevance.

The equations of electrodynamics can be recast in a manifestly Lorentz-invariant form using the four-vector gauge potential $A^\mu = (\varphi, \vec{A})$ and the antisymmetric rank-two tensor: $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. Maxwell's equations become

$$\begin{aligned} \partial_\mu F^{\mu\nu} &= j^\mu, \\ \epsilon^{\mu\nu\sigma\eta}\partial_\nu F_{\sigma\eta} &= 0, \end{aligned} \tag{4.49}$$

where the four-current $j^\mu = (\rho, \vec{j})$ contains the charge density and the electric current. The field strength tensor $F_{\mu\nu}$ and the Maxwell equations are invariant under gauge transformations (4.48), which in covariant form read

$$A_\mu \longrightarrow A_\mu + \partial_\mu \epsilon. \tag{4.50}$$

Finally, the equations of motion of charged particles are given, in covariant form, by

$$m\frac{du^\mu}{d\tau} = eF^{\mu\nu}u_\nu, \tag{4.51}$$

where e is the charge of the particle and $u^\mu(\tau)$ its four-velocity as a function of the proper time.

The physical rôle of the vector potential becomes manifest only in quantum mechanics. Using the prescription of minimal substitution $\vec{p} \to \vec{p} - e\vec{A}$, the Schrödinger equation describing a particle with charge e moving in an electromagnetic field is

$$i\partial_t \Psi = \left[-\frac{1}{2m}\left(\vec{\nabla} - ie\vec{A}\right)^2 + e\varphi\right]\Psi. \tag{4.52}$$

Because of the explicit dependence on the electromagnetic potentials φ and $\vec{A}$, this equation seems to change under the gauge transformations (4.48). This is physically acceptable only if the ambiguity does not affect the probability density given by $|\Psi(t,\vec{x})|^2$. Therefore, a gauge transformation of the electromagnetic potential should amount to a change in the (unobservable) phase of the wave function. This is indeed what happens: the Schrödinger equation (4.52) is invariant under the gauge transformations (4.48) provided the phase of the wave function is transformed at the same time according to

$$\Psi(t,\vec{x}) \longrightarrow e^{-ie\,\epsilon(t,\vec{x})}\Psi(t,\vec{x}). \tag{4.53}$$

Aharonov–Bohm effect. This interplay between gauge transformations and the phase of the wave function gives rise to surprising phenomena. The first evidence of the rôle played by the electromagnetic potentials at the quantum level was pointed out by Yakir Aharonov and David Bohm [19]. Let us consider a double-slit experiment as shown in Fig. 7, where we have placed a shielded solenoid just behind the first screen. Although the magnetic field is confined to the interior of the solenoid, the vector potential is non-vanishing also outside. Of course the value of $\vec{A}$ outside the solenoid is a pure gauge, i.e., $\vec{\nabla}\times\vec{A} = \vec{0}$, however, because the region outside the solenoid is not simply connected the vector potential cannot be gauged to zero everywhere. If we denote by $\Psi_1^{(0)}$ and $\Psi_2^{(0)}$ the wave functions for each of the two electron beams in the absence of the solenoid, the total wave function once the magnetic field is switched on can be written as

$$\begin{aligned}\Psi &= e^{ie\int_{\Gamma_1}\vec{A}\cdot d\vec{x}}\Psi_1^{(0)} + e^{ie\int_{\Gamma_2}\vec{A}\cdot d\vec{x}}\Psi_2^{(0)} \\ &= e^{ie\int_{\Gamma_1}\vec{A}\cdot d\vec{x}}\left[\Psi_1^{(0)} + e^{ie\oint_{\Gamma}\vec{A}\cdot d\vec{x}}\Psi_2^{(0)}\right],\end{aligned} \tag{4.54}$$

where Γ_1 and Γ_2 are two curves surrounding the solenoid from different sides, and Γ is any closed loop surrounding it. Therefore the relative phase between the two beams gets an extra term depending on the value of the vector potential outside the solenoid as

$$U = \exp\left[ie\oint_\Gamma \vec{A}\cdot d\vec{x}\right]. \tag{4.55}$$

Because of the change in the relative phase of the electron wave functions, the presence of the vector potential becomes observable even if the electrons do not feel the magnetic field. If we perform the double-slit experiment when the magnetic field inside the solenoid is switched off we will observe the

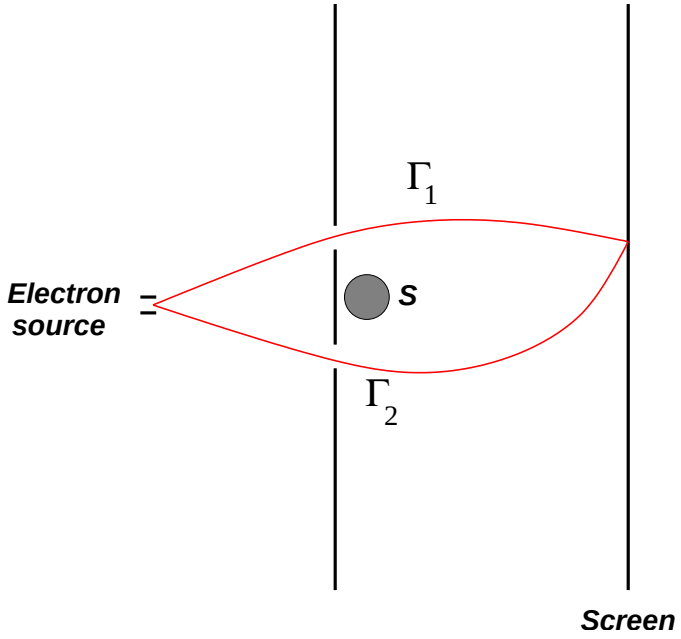


Fig. 7: Illustration of an interference experiment to show the Aharonov–Bohm effect. S represents the solenoid in whose interior the magnetic field is confined.

usual interference pattern on the second screen. However, if now the magnetic field is switched on, because of the phase (4.54), a change in the interference pattern will appear. This is the Aharonov–Bohm effect.

The first question that comes up is what happens with gauge invariance. Since we said that $\vec{A}$ can be changed by a gauge transformation it seems that the resulting interference patters might depend on the gauge used. Actually, the phase U in (4.55) is independent of the gauge although, unlike other gauge-invariant quantities like $\vec{E}$ and $\vec{B}$, non-local. Notice that, since $\vec{\nabla} \times \vec{A} = \vec{0}$ outside the solenoid, the value of U does not change under continuous deformations of the closed curve Γ, so long as it does not cross the solenoid.

The Dirac monopole. It is very easy to check that the vacuum Maxwell equations remain invariant under the transformation

$$\vec{E} - i\vec{B} \longrightarrow e^{i\theta}(\vec{E} - i\vec{B}), \qquad \theta \in [0, 2\pi] \tag{4.56}$$

which, in particular, for $\theta = \frac{\pi}{2}$ interchanges the electric and the magnetic fields: $\vec{E} \to \vec{B}$, $\vec{B} \to -\vec{E}$. This duality symmetry is, however, broken in the presence of electric sources. Nevertheless the Maxwell equations can be 'completed' by introducing sources for the magnetic field $(\rho_m, \vec{\jmath}_m)$ in such a way that the duality (4.56) is restored when supplemented by the transformation

$$\rho - i\rho_m \longrightarrow e^{i\theta}(\rho - i\rho_m), \qquad \vec{\jmath} - i\vec{\jmath}_m \longrightarrow e^{i\theta}(\vec{\jmath} - i\vec{\jmath}_m). \tag{4.57}$$

Again for $\theta = \pi/2$ the electric and magnetic sources are interchanged.

In 1931 Dirac [20] studied the possibility of finding solutions to the completed Maxwell equation with a magnetic monopole of charge g, i.e., solutions to

$$\vec{\nabla} \cdot \vec{B} = g\,\delta(\vec{x}). \tag{4.58}$$

Away from the position of the monopole, $\vec{\nabla} \cdot \vec{B} = 0$ and the magnetic field can still be derived locally from a vector potential $\vec{A}$ according to $\vec{B} = \vec{\nabla} \times \vec{A}$. However, the vector potential cannot be regular

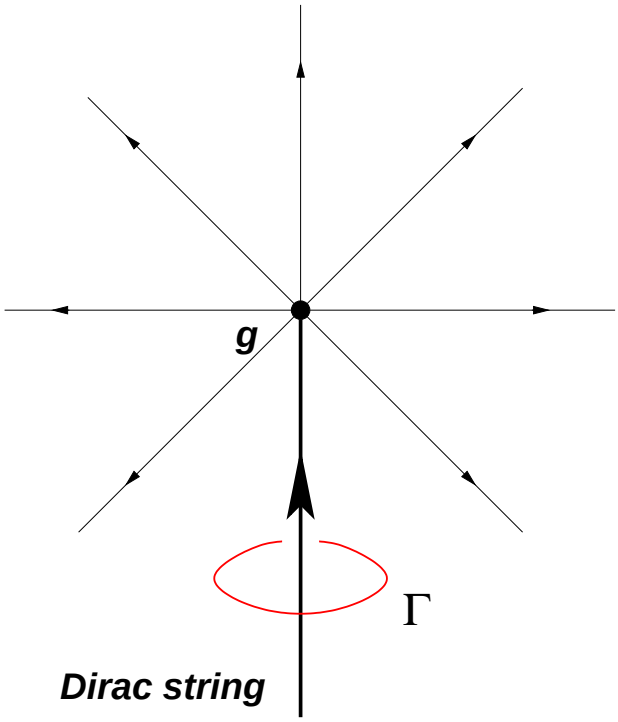


Fig. 8: The Dirac monopole

everywhere since otherwise Gauss's law would imply that the magnetic flux threading a closed surface around the monopole should vanish, contradicting (4.58).

We look now for solutions to Eq. (4.58). Working in spherical coordinates we find

$$B_r = \frac{g}{|\vec{x}|^2}, \qquad B_\varphi = B_\theta = 0. \tag{4.59}$$

Away from the position of the monopole ($\vec{x} \neq \vec{0}$) the magnetic field can be derived from the vector potential

$$A_\varphi = \frac{g}{|\vec{x}|} \tan\frac{\theta}{2}, \qquad A_r = A_\theta = 0. \tag{4.60}$$

As expected we find that this vector potential is actually singular around the half-line $\theta = \pi$ (see Fig. 8). This singular line starting at the position of the monopole is called the Dirac string and its position changes with a change of gauge but cannot be eliminated by any gauge transformation. Physically we can see it as an infinitely thin solenoid confining a magnetic flux entering into the magnetic monopole from infinity that equals the outgoing magnetic flux from the monopole.

Since the position of the Dirac string depends on the gauge that is chosen it seems that the presence of monopoles introduces an ambiguity. This would be rather strange, since Maxwell equations are gauge invariant also in the presence of magnetic sources. The solution to this apparent riddle lies in the fact that the Dirac string does not pose any consistency problem as long as it does not produce any physical effect, i.e., if its presence turns out to be undetectable. From our discussion of the Aharonov–Bohm effect we know that the wave function of charged particles picks up a phase (4.55) when surrounding a region where magnetic flux is confined (for example the solenoid in the Aharonov–Bohm experiment). As explained above, the Dirac string associated with the monopole can be seen as an infinitely thin solenoid. Therefore the Dirac string will be unobservable if the phase picked up by the wave function of a charged particle is equal to one. A simple calculation shows that this happens if

$$e^{i\,e\,g} = 1 \qquad \Longrightarrow \qquad e\,g = 2\pi n \;\text{ with }\; n \in \mathbb{Z}. \tag{4.61}$$

Interestingly, this discussion leads to the conclusion that the presence of a single magnetic monopole somewhere in the Universe implies for consistency the quantization of the electric charge in units of $\frac{2\pi}{g}$, where g is the magnetic charge of the monopole.

Quantization of the electromagnetic field. We now proceed to the quantization of the electromagnetic field in the absence of sources $\rho = 0$, $\vec{\jmath} = \vec{0}$. In this case the Maxwell equations (4.49) can be derived from the Lagrangian density

$$\mathcal{L}_{\rm Maxwell} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} = \frac{1}{2}\left(\vec{E}^{\,2} - \vec{B}^{\,2}\right). \tag{4.62}$$

Although in general the procedure to quantize the Maxwell Lagrangian is not very different from the one used for the Klein–Gordon or the Dirac field, here we need to deal with a new ingredient: gauge invariance. Unlike the cases studied so far, here the photon field A_μ is not unambiguously defined because the action and the equations of motion are insensitive to the gauge transformations $A_\mu \to A_\mu + \partial_\mu \varepsilon$. A first consequence of this symmetry is that the theory has less physical degrees of freedom than one would expect from the fact that we are dealing with a vector field.

The way to tackle the problem of gauge invariance is to fix the freedom in choosing the electromagnetic potential before quantization. This can be done in several ways, for example by imposing the Lorentz-gauge-fixing condition

$$\partial_\mu A^\mu = 0. \tag{4.63}$$

Notice that this condition does not fix completely the gauge freedom since Eq. (4.63) is left invariant by gauge transformations satisfying $\partial_\mu \partial^\mu \varepsilon = 0$. One of the advantages, however, of the Lorentz gauge is that it is covariant and therefore does not pose any danger to the Lorentz invariance of the quantum theory. Besides, applying it to the Maxwell equation $\partial_\mu F^{\mu\nu} = 0$ one finds

$$0 = \partial_\mu \partial^\mu A^\nu - \partial_\nu \left(\partial_\mu A^\mu\right) = \partial_\mu \partial^\mu A^\nu, \tag{4.64}$$

which means that since A_μ satisfies the massless Klein–Gordon equation the photon, the quantum of the electromagnetic field, has zero mass.

Once gauge invariance is fixed A_μ is expanded in a complete basis of solutions to (4.64) and the canonical commutation relations are imposed

$$\widehat{A}_\mu(t,\vec{x}) = \sum_{\lambda=\pm 1} \int \frac{d^3k}{(2\pi)^3} \frac{1}{2|\vec{k}|} \left[\epsilon_\mu(\vec{k},\lambda)\widehat{a}(\vec{k},\lambda) e^{-i|\vec{k}|t + i\vec{k}\cdot\vec{x}} + \epsilon_\mu(\vec{k},\lambda)^*\, \widehat{a}^\dagger(\vec{k},\lambda) e^{i|\vec{k}|t - i\vec{k}\cdot\vec{x}}\right] \tag{4.65}$$

where $\lambda = \pm 1$ represent the helicity of the photon, and $\epsilon_\mu(\vec{k},\lambda)$ are solutions to the equations of motion with well-defined momentum and helicity. Because of (4.63) the polarization vectors have to be orthogonal to k_μ

$$k^\mu \epsilon_\mu(\vec{k},\lambda) = k^\mu \epsilon_\mu(\vec{k},\lambda)^* = 0. \tag{4.66}$$

The canonical commutation relations imply that

$$\begin{aligned} [\widehat{a}(\vec{k},\lambda), \widehat{a}^\dagger(\vec{k}\,',\lambda')] &= i\delta(\vec{k} - \vec{k}\,')\delta_{\lambda\lambda'} \\ [\widehat{a}(\vec{k},\lambda), \widehat{a}(\vec{k}\,',\lambda')] &= [\widehat{a}^\dagger(\vec{k},\lambda), \widehat{a}^\dagger(\vec{k}\,',\lambda')] = 0. \end{aligned} \tag{4.67}$$

Therefore $\widehat{a}(\vec{k},\lambda)$, $\widehat{a}^\dagger(\vec{k},\lambda)$ form a set of creation–annihilation operators for photons with momentum $\vec{k}$ and helicity λ.

Behind the simple construction presented above there are a number of subtleties related to gauge invariance. In particular the gauge freedom seems to introduce states in the Hilbert space with negative

probability. A careful analysis shows that when gauge invariance is properly handled these spurious states decouple from physical states and can be eliminated. The details can be found in standard textbooks [1–10].

Coupling gauge fields to matter. Once we know how to quantize the electromagnetic field we consider theories containing electrically charged particles, for example electrons. To couple the Dirac Lagrangian to electromagnetism we use as guiding principle what we learned about the Schrödinger equation for a charged particle. There we saw that the gauge ambiguity of the electromagnetic potential is compensated with a U(1) phase shift in the wave function. In the case of the Dirac equation we know that the Lagrangian is invariant under $\psi \to e^{ie\varepsilon}\psi$, with ε a constant. However, this invariance is broken as soon as one identifies ε with the gauge transformation parameter of the electromagnetic field which depends on the position.

Looking at the Dirac Lagrangian (4.27) it is easy to see that in order to promote the global U(1) symmetry into a local one, $\psi \to e^{ie\varepsilon(x)}\psi$, it suffices to replace the ordinary derivative ∂_μ by a covariant one D_μ satisfying

$$D_\mu \left[e^{ie\varepsilon(x)}\psi \right] = e^{ie\varepsilon(x)} D_\mu \psi. \tag{4.68}$$

This covariant derivative can be constructed in terms of the gauge potential A_μ as

$$D_\mu = \partial_\mu - ieA_\mu. \tag{4.69}$$

The Lagrangian of a spin-$\frac{1}{2}$ field coupled to electromagnetism is written as

$$\mathcal{L}_{\rm QED} = -\frac{1}{4} F_{\mu\nu}F^{\mu\nu} + \overline{\psi}(i\not{D} - m)\psi, \tag{4.70}$$

invariant under the gauge transformations

$$\psi \longrightarrow e^{ie\varepsilon(x)}\psi, \qquad A_\mu \longrightarrow A_\mu + \partial_\mu \varepsilon(x). \tag{4.71}$$

Unlike the theories we have seen so far, the Lagrangian (4.70) describe an interacting theory. By plugging (4.69) into the Lagrangian we find that the interaction between fermions and photons is

$$\mathcal{L}^{(\rm int)}_{\rm QED} = -eA_\mu\, \overline{\psi}\gamma^\mu\psi. \tag{4.72}$$

As advertised above, in the Dirac theory the electric current four-vector is given by $j^\mu = e\overline{\psi}\gamma^\mu\psi$.

The quantization of interacting field theories poses new problems that we did not meet in the case of the free theories. In particular, in most cases it is not possible to solve the theory exactly. When this happens the physical observables have to be computed in perturbation theory in powers of the coupling constant. An added problem appears when computing quantum corrections to the classical result, since in that case the computation of observables is plagued with infinities that should be taken care of. We will go back to this problem in Section 8.

Non-Abelian gauge theories. Quantum electrodynamics (QED) is the simplest example of a gauge theory coupled to matter based on the Abelian gauge symmetry of local U(1) phase rotations. However, it is possible also to construct gauge theories based on non-Abelian groups. Actually, our knowledge of the strong and weak interactions is based on the use of such non-Abelian generalizations of QED.

Let us consider a gauge group G with generators T^a, $a = 1, \ldots, \dim G$ satisfying the Lie algebra[6]

$$[T^a, T^b] = if^{abc}T^c. \tag{4.73}$$

[6]Some basic facts about Lie groups have been summarized in Appendix A.

A gauge field taking values on the Lie algebra of $\mathcal{G}$ can be introduced $A_\mu \equiv A_\mu^a T^a$ which transforms under a gauge transformation as

$$A_\mu \longrightarrow \frac{1}{ig} U \partial_\mu U^{-1} + U A_\mu U^{-1}, \qquad U = e^{i\chi^a(x) T^a}, \tag{4.74}$$

where g is the coupling constant. The associated field strength is defined as

$$F_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a - g f^{abc} A_\mu^b A_\nu^c. \tag{4.75}$$

Notice that this definition of the $F_{\mu\nu}^a$ reduces to the one used in QED in the Abelian case when $f^{abc} = 0$. In general, however, unlike the case of QED the field strength is not gauge invariant. In terms of $F_{\mu\nu} = F_{\mu\nu}^a T^a$ it transforms as

$$F_{\mu\nu} \longrightarrow U F_{\mu\nu} U^{-1}. \tag{4.76}$$

The coupling of matter to a non-Abelian gauge field is done by introducing again a covariant derivative. For a field in a representation of $\mathcal{G}$

$$\Phi \longrightarrow U\Phi \tag{4.77}$$

the covariant derivative is given by

$$D_\mu \Phi = \partial_\mu \Phi - i g A_\mu^a T^a \Phi. \tag{4.78}$$

With the help of this we can write a generic Lagrangian for a non-Abelian gauge field coupled to scalars ϕ and spinors ψ as

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu}^a F^{\mu\nu\, a} + i\overline{\psi} \not{D} \psi + \overline{D_\mu \phi} D^\mu \phi - \overline{\psi} \left[M_1(\phi) + i\gamma_5 M_2(\phi) \right] \psi - V(\phi). \tag{4.79}$$

In order to keep the theory renormalizable we have to restrict $M_1(\phi)$ and $M_2(\phi)$ to be at most linear in ϕ whereas $V(\phi)$ has to be at most of quartic order. The Lagrangian of the Standard Model is of the form (4.79).

4.4 Understanding gauge symmetry

In classical mechanics the use of the Hamiltonian formalism starts with the replacement of generalized velocities by momenta

$$p_i \equiv \frac{\partial L}{\partial \dot{q}_i} \qquad \Longrightarrow \qquad \dot{q}_i = \dot{q}_i(q, p). \tag{4.80}$$

Most of the time there is no problem in inverting the relations $p_i = p_i(q, \dot{q})$. However, in some systems these relations might not be invertible and result in a number of constraints of the type

$$f_a(q, p) = 0, \qquad a = 1, \ldots, N_1. \tag{4.81}$$

These systems are called degenerate or constrained [21, 22].

The presence of constraints of the type (4.81) makes the formulation of the Hamiltonian formalism more involved. The first problem is related to the ambiguity in defining the Hamiltonian, since the addition of any linear combination of the constraints does not modify its value. Secondly, one has to make sure that the constraints are consistent with the time evolution in the system. In the language of Poisson brackets this means that further constraints have to be imposed in the form

$$\{f_a, H\} \approx 0. \tag{4.82}$$

Following [21] we use the symbol $\approx$ to indicate a 'weak' equality that holds when the constraints $f_a(q,p) = 0$ are satisfied. Notice however that since the computation of the Poisson brackets involves derivatives, the constraints can be used only after the bracket is computed. In principle the conditions (4.82) can give rise to a new set of constraints $g_b(q,p) = 0$, $b = 1, \ldots, N_2$. Again these constraints have to be consistent with time evolution and we have to repeat the procedure. Eventually this finishes when a set of constraints is found that does not require any further constraint to be preserved by the time evolution[7].

Once we find all the constraints of a degenerate system we consider the so-called first class constraints $\phi_a(q,p) = 0$, $a = 1, \ldots, M$, which are those whose Poisson bracket vanishes weakly

$$\{\phi_a, \phi_b\} = c_{abc}\phi_c \approx 0. \tag{4.83}$$

The constraints that do not satisfy this condition, called second class constraints, can be eliminated by modifying the Poisson bracket [21]. Then the total Hamiltonian of the theory is defined by

$$H_T = p_i q_i - L + \sum_{a=1}^{M} \lambda(t)\phi_a. \tag{4.84}$$

What has all this to do with gauge invariance? The interesting answer is that for a singular system the first class constraints ϕ_a generate gauge transformations. Indeed, because $\{\phi_a, \phi_b\} \approx 0 \approx \{\phi_a, H\}$ the transformations

$$\begin{aligned} q_i &\longrightarrow q_i + \sum_{a}^{M} \varepsilon_a(t)\{q_i, \phi_a\}, \\ p_i &\longrightarrow p_i + \sum_{a}^{M} \varepsilon_a(t)\{p_i, \phi_a\} \end{aligned} \tag{4.85}$$

leave invariant the state of the system. This ambiguity in the description of the system in terms of the generalized coordinates and momenta can be traced back to the equations of motion in Lagrangian language. Writing them in the form

$$\frac{\partial^2 L}{\partial \dot{q}_i \partial \dot{q}_j}\ddot{q}_j = -\frac{\partial^2 L}{\partial \dot{q}_i \partial q_j}\dot{q}_j + \frac{\partial L}{\partial q_i}, \tag{4.86}$$

we find that, in order to determine the accelerations in terms of the positions and velocities, the matrix $\frac{\partial^2 L}{\partial \dot{q}_i \partial \dot{q}_j}$ has to be invertible. However, the existence of constraints (4.81) precisely implies that the determinant of this matrix vanishes and therefore the time evolution is not uniquely determined in terms of the initial conditions.

Let us apply this to Maxwell electrodynamics described by the Lagrangian

$$L = -\frac{1}{4}\int d^3\, F_{\mu\nu}F^{\mu\nu}. \tag{4.87}$$

The generalized momentum conjugate to A_μ is given by

$$\pi^\mu = \frac{\delta L}{\delta(\partial_0 A_\mu)} = F^{0\mu}. \tag{4.88}$$

In particular for the time component we find the constraint $\pi^0 = 0$. The Hamiltonian is given by

$$H = \int d^3x \left[\pi^\mu \partial_0 A_\mu - \mathcal{L}\right] = \int d^3x \left[\frac{1}{2}\left(\vec{E}^{\,2} + \vec{B}^{\,2}\right) + \pi^0\partial_0 A_0 + A_0\vec{\nabla}\cdot\vec{E}\right]. \tag{4.89}$$

[7] In principle it is also possible that the procedure finishes because some kind of inconsistent identity is found. In this case the system itself is inconsistent as is the case with the Lagrangian $L(q,\dot{q}) = q$.

Requiring the consistency of the constraint $\pi^0 = 0$ we find a second constraint

$$\{\pi^0, H\} \approx \partial_0 \pi^0 + \vec{\nabla} \cdot \vec{E} = 0. \tag{4.90}$$

Together with the first constraint $\pi^0 = 0$ this one implies Gauss's law $\vec{\nabla} \cdot \vec{E} = 0$. These two constraints have vanishing Poisson bracket and therefore they are first class. Therefore the total Hamiltonian is given by

$$H_T = H + \int d^3x \left[\lambda_1(x) \pi^0 + \lambda_2(x) \vec{\nabla} \cdot \vec{E} \right], \tag{4.91}$$

where we have absorbed A_0 in the definition of the arbitrary functions $\lambda_1(x)$ and $\lambda_2(x)$. Actually, we can fix part of the ambiguity taking $\lambda_1 = 0$. Notice that, because A_0 has been included in the multipliers, fixing λ_1 amounts to fixing the value of A_0 and therefore it is equivalent to taking a temporal gauge. In this case the Hamiltonian is

$$H_T = \int d^3x \left[\frac{1}{2} \left(\vec{E}^{\,2} + \vec{B}^{\,2} \right) + \varepsilon(x) \vec{\nabla} \cdot \vec{E} \right] \tag{4.92}$$

and we are left just with Gauss's law as the only constraint. Using the canonical commutation relations

$$\{A_i(t, \vec{x}), E_j(t, \vec{x}\,')\} = \delta_{ij} \delta(\vec{x} - \vec{x}\,') \tag{4.93}$$

we find that the remaining gauge transformations are generated by Gauss's law

$$\delta A_i = \{A_i, \int d^3x' \, \varepsilon \, \vec{\nabla} \cdot \vec{E}\} = \partial_i \varepsilon, \tag{4.94}$$

while leaving A_0 invariant, so for consistency with the general gauge transformations the function $\varepsilon(x)$ should be independent of time. Notice that the constraint $\vec{\nabla} \cdot \vec{E} = 0$ can be implemented by demanding $\vec{\nabla} \cdot \vec{A} = 0$ which reduces the three degrees of freedom of $\vec{A}$ to the two physical degrees of freedom of the photon.

So much for the classical analysis. In the quantum theory the constraint $\vec{\nabla} \cdot \vec{E} = 0$ has to be imposed on the physical states $|\mathrm{phys}\rangle$. This is done by defining the following unitary operator on the Hilbert space

$$\mathcal{U}(\varepsilon) \equiv \exp\left(i \int d^3x \, \varepsilon(\vec{x}) \, \vec{\nabla} \cdot \vec{E} \right). \tag{4.95}$$

By definition, physical states should not change when a gauge transformations is performed. This is implemented by requiring that the operator $\mathcal{U}(\varepsilon)$ act trivially on a physical state

$$\mathcal{U}(\varepsilon)|\mathrm{phys}\rangle = |\mathrm{phys}\rangle \qquad \Longrightarrow \qquad (\vec{\nabla} \cdot \vec{E})|\mathrm{phys}\rangle = 0. \tag{4.96}$$

In the presence of charge density ρ, the condition that physical states are annihilated by Gauss's law changes to $(\vec{\nabla} \cdot \vec{E} - \rho)|\mathrm{phys}\rangle = 0$.

The role of gauge transformations in the quantum theory is very illuminating in understanding the real rôle of gauge invariance [23]. As we have learned, the existence of a gauge symmetry in a theory reflects a degree of redundancy in the description of physical states in terms of the degrees of freedom appearing in the Lagrangian. In classical mechanics, for example, the state of a system is usually determined by the value of the canonical coordinates (q_i, p_i). We know, however, that this is not the case for constrained Hamiltonian systems where the transformations generated by the first class constraints change the value of q_i and p_i without changing the physical state. In the case of Maxwell theory for every physical configuration determined by the gauge invariant quantities $\vec{E}$, $\vec{B}$ there is an infinite number of possible values of the vector potential that are related by gauge transformations $\delta A_\mu = \partial_\mu \varepsilon$.

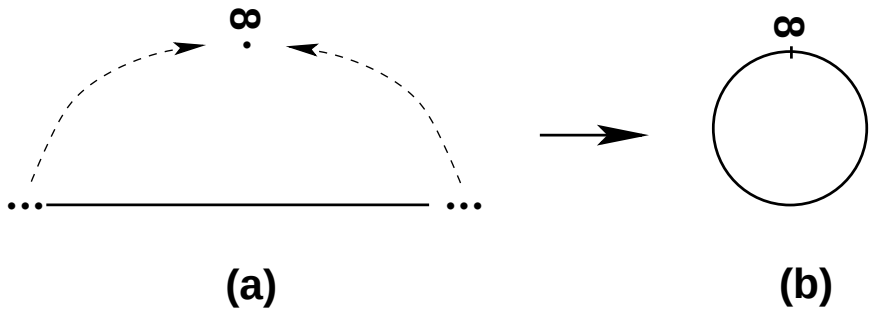


Fig. 9: Compactification of the real line (a) into the circumference S^1 (b) by adding the point at infinity

In the quantum theory this means that the Hilbert space of physical states is defined as the result of identifying all states related by the operator $\mathcal{U}(\varepsilon)$ with any gauge function $\varepsilon(x)$ into a single physical state $|\mathrm{phys}\rangle$. In other words, each physical state corresponds to a whole orbit of states that are transformed among themselves by gauge transformations.

This explains the necessity of gauge fixing. In order to avoid the redundancy in the states a further condition can be given that selects one single state on each orbit. In the case of Maxwell electrodynamics the conditions $A_0 = 0$, $\vec{\nabla} \cdot \vec{A} = 0$ selects a value of the gauge potential among all possible ones giving the same value for the electric and magnetic fields.

Since states have to be identified by gauge transformations the topology of the gauge group plays an important physical rôle. To illustrate the point let us first deal with a toy model of a U(1) gauge theory in 1+1 dimensions. Later we will be more general. In the Hamiltonian formalism gauge transformations $g(\vec{x})$ are functions defined on $\mathbb{R}$ with values on the gauge group U(1)

$$g : \mathbb{R} \longrightarrow U(1). \tag{4.97}$$

We assume that $g(x)$ is regular at infinity. In this case we can add to the real line $\mathbb{R}$ the point at infinity to compactify it into the circumference S^1 (see Fig. 9). Once this is done $g(x)$ are functions defined on S^1 with values on $U(1) = S^1$ that can be parametrized as

$$g : S^1 \longrightarrow U(1), \qquad g(x) = e^{i\alpha(x)}, \tag{4.98}$$

with $x \in [0, 2\pi]$.

Because S^1 does have a nontrivial topology, $g(x)$ can be divided into topological sectors. These sectors are labelled by an integer number $n \in \mathbb{Z}$ and are defined by

$$\alpha(2\pi) = \alpha(0) + 2\pi\, n \ . \tag{4.99}$$

Geometrically n gives the number of times that the spatial S^1 winds around the S^1 defining the gauge group U(1). This winding number can be written in a more sophisticated way as

$$\oint_{S^1} g(x)^{-1} dg(x) = 2\pi n \ , \tag{4.100}$$

where the integral is along the spatial S^1.

In $\mathbb{R}^3$ a similar situation happens with the gauge group[8] SU(2). If we demand $g(\vec{x}) \in$ SU(2) to be regular at infinity $|\vec{x}| \to \infty$ we can compactify $\mathbb{R}^3$ into a three-dimensional sphere S^3, exactly as we did in 1+1 dimensions. On the other hand, the function $g(\vec{x})$ can be written as

$$g(\vec{x}) = a^0(x)\mathbf{1} + \vec{a}(x) \cdot \vec{\sigma} \tag{4.101}$$

[8] Although we present for simplicity only the case of SU(2), similar arguments apply to any simple group.

and the conditions $g(x)^\dagger g(x) = \mathbf{1}$, $\det g = 1$ implies that $(a^0)^2 + \vec{a}^2 = 1$. Therefore SU(2) is a three-dimensional sphere and $g(x)$ defines a function

$$g : S^3 \longrightarrow S^3. \tag{4.102}$$

As was the case in 1+1 dimensions, here the gauge transformations $g(x)$ are also divided into topological sectors labelled this time by the winding number

$$n = \frac{1}{24\pi^2}\int_{S^3} d^3x\, \epsilon_{ijk}\mathrm{Tr}\left[\left(g^{-1}\partial_i g\right)\left(g^{-1}\partial_i g\right)\left(g^{-1}\partial_i g\right)\right] \in \mathbb{Z}. \tag{4.103}$$

In the two cases analysed we find that due to the nontrivial topology of the gauge group manifold the gauge transformations are divided into different sectors labelled by an integer n. Gauge transformations with different values of n cannot be smoothly deformed into each other. The sector with $n = 0$ corresponds to those gauge transformations that can be connected with the identity.

Now we can be a bit more formal. Let us consider a gauge theory in 3+1 dimensions with gauge group G and let us denote by $\mathcal{G}$ the set of all gauge transformations $\mathcal{G} = \{g : S^3 \to G\}$. At the same time we define $\mathcal{G}_0$ as the set of transformations in $\mathcal{G}$ that can be smoothly deformed into the identity. Our theory will have topological sectors if

$$\mathcal{G}/\mathcal{G}_0 \neq \mathbf{1}. \tag{4.104}$$

In the case of electromagnetism we have seen that Gauss's law annihilates physical states. For a non-Abelian theory the analysis is similar and leads to the condition

$$\mathcal{U}(g_0)|\mathrm{phys}\rangle \equiv \exp\left[i\int d^3x\, \chi^a(\vec{x})\vec{\nabla}\cdot\vec{E}^a\right]|\mathrm{phys}\rangle = |\mathrm{phys}\rangle, \tag{4.105}$$

where $g_0(\vec{x}) = e^{i\chi^a(\vec{x})T^a}$ is in the connected component of the identity $\mathcal{G}_0$. The important point to realize here is that only the elements of $\mathcal{G}_0$ can be written as exponentials of the infinitesimal generators. Since these generators annihilate the physical states this implies that $\mathcal{U}(g_0)|\mathrm{phys}\rangle = |\mathrm{phys}\rangle$ only when $g_0 \in \mathcal{G}_0$.

What happens then with the other topological sectors? If $g \in \mathcal{G}/\mathcal{G}_0$ there is still a unitary operator $\mathcal{U}(g)$ that realizes gauge transformations on the Hilbert space of the theory. However, since g is not in the connected component of the identity, it cannot be written as the exponential of Gauss's law. Still gauge invariance is preserved if $\mathcal{U}(g)$ only changes the overall global phase of the physical states. For example, if g_1 is a gauge transformation with winding number $n = 1$

$$\mathcal{U}(g_1)|\mathrm{phys}\rangle = e^{i\theta}|\mathrm{phys}\rangle. \tag{4.106}$$

It is easy to convince oneself that all transformations with winding number $n = 1$ have the same value of θ modulo 2π. This can be shown by noting that if $g(\vec{x})$ has winding number $n = 1$ then $g(\vec{x})^{-1}$ has opposite winding number $n = -1$. Since the winding number is additive, given two transformations g_1, g_2 with winding number 1, $g_1^{-1}g_2$ has winding number $n = 0$. This implies that

$$|\mathrm{phys}\rangle = \mathcal{U}(g_1^{-1}g_2)|\mathrm{phys}\rangle = \mathcal{U}(g_1)^\dagger\mathcal{U}(g_2)|\mathrm{phys}\rangle = e^{i(\theta_2-\theta_1)}|\mathrm{phys}\rangle \tag{4.107}$$

and we conclude that $\theta_1 = \theta_2 \bmod 2\pi$. Once we know this, it is straightforward to conclude that a gauge transformation $g_n(\vec{x})$ with winding number n has the following action on physical states

$$\mathcal{U}(g_n)|\mathrm{phys}\rangle = e^{in\theta}|\mathrm{phys}\rangle, \qquad n \in \mathbb{Z}. \tag{4.108}$$

To find a physical interpretation of this result we are going to look for similar things in other physical situations. One of them is borrowed from condensed matter physics and refers to the quantum

states of electrons in the periodic potential produced by the ion lattice in a solid. For simplicity we discuss the one-dimensional case where the minima of the potential are separated by a distance a. When the barrier between consecutive degenerate vacua is high enough we can neglect tunneling between different vacua and consider the ground state $|na\rangle$ of the potential near the minimum located at $x = na$ ($n \in \mathbb{Z}$) as possible vacua of the theory. This vacuum state is, however, not invariant under lattice translations

$$e^{ia\widehat{P}}|na\rangle = |(n+1)a\rangle. \tag{4.109}$$

However, it is possible to define a new vacuum state

$$|k\rangle = \sum_{n\in\mathbb{Z}} e^{-ikna}|na\rangle, \tag{4.110}$$

which under $e^{ia\widehat{P}}$ transforms by a global phase

$$e^{ia\widehat{P}}|k\rangle = \sum_{n\in\mathbb{Z}} e^{-ikna}|(n+1)a\rangle = e^{ika}|k\rangle. \tag{4.111}$$

This ground state is labelled by the momentum k and corresponds to the Bloch wave function.

This looks very much the same as what we found for non-Abelian gauge theories. The vacuum state labelled by θ plays a rôle similar to the Bloch wave function for the periodic potential with the identification of θ with the momentum k. To make this analogy more precise let us write the Hamiltonian for non-Abelian gauge theories

$$H = \frac{1}{2}\int d^3x \left(\vec{\pi}_a\cdot\vec{\pi}_a + \vec{B}_a\cdot\vec{B}_a\right) = \frac{1}{2}\int d^3x \left(\vec{E}_a\cdot\vec{E}_a + \vec{B}_a\cdot\vec{B}_a\right), \tag{4.112}$$

where we have used the expression of the canonical momenta π_a^i and we assume that the Gauss law constraint is satisfied. Looking at this Hamiltonian we can interpret the first term within the brackets as the kinetic energy $T = \frac{1}{2}\vec{\pi}_a\cdot\vec{\pi}_a$ and the second term as the potential energy $V = \frac{1}{2}\vec{B}_a\cdot\vec{B}_a$. Since $V \geq 0$ we can identify the vacua of the theory as those $\vec{A}$ for which $V = 0$, modulo gauge transformations. This happens wherever $\vec{A}$ is a pure gauge. However, since we know that the gauge transformations are labelled by the winding number we can have an infinite number of vacua which cannot be continuously connected with one another using trivial gauge transformations. Taking a representative gauge transformation $g_n(\vec{x})$ in the sector with winding number n, these vacua will be associated with the gauge potentials

$$\vec{A} = \frac{1}{ig}g_n(\vec{x})^{-1}\vec{\nabla}g_n(\vec{x}), \tag{4.113}$$

modulo topologically trivial gauge transformations. Therefore the theory is characterized by an infinite number of vacua $|n\rangle$ labelled by the winding number. These vacua are not gauge invariant. Indeed, a gauge transformation with $n = 1$ will change the winding number of the vacua in one unit

$$\mathcal{U}(g_1)|n\rangle = |n+1\rangle. \tag{4.114}$$

Nevertheless a gauge invariant vacuum can be defined as

$$|\theta\rangle = \sum_{n\in\mathbb{Z}} e^{-in\theta}|n\rangle, \qquad \text{with } \theta\in\mathbb{R} \tag{4.115}$$

satisfying

$$\mathcal{U}(g_1)|\theta\rangle = e^{i\theta}|\theta\rangle. \tag{4.116}$$

We have concluded that the non-trivial topology of the gauge group has very important physical consequences for quantum theory. In particular it implies an ambiguity in the definition of the vacuum. Actually, this can also be seen in a Lagrangian analysis. In constructing the Lagrangian for the non-Abelian version of Maxwell's theory we only consider the term $F^a_{\mu\nu}F^{\mu\nu\,a}$. However, this is not the only Lorentz- and gauge-invariant term that contains just two derivatives. We can write the more general Lagrangian

$$\mathcal{L} \;=\; -\frac{1}{4}F^a_{\mu\nu}F^{\mu\nu\,a} + \frac{\theta}{32\pi^2}F^a_{\mu\nu}\widetilde{F}^{\mu\nu\,a}, \tag{4.117}$$

where $\widetilde{F}^a_{\mu\nu}$ is the dual of the field strength defined by

$$\widetilde{F}^a_{\mu\nu} = \frac{1}{2}\epsilon_{\mu\nu\sigma\lambda}F^{\sigma\lambda}. \tag{4.118}$$

The extra term in (4.117), proportional to $\vec{E}^{\,a}\cdot\vec{B}^{\,a}$, is actually a total derivative and does not change the equations of motion or the quantum perturbation theory. Nevertheless it has several important physical consequences. One of them is that it violates both parity P and the combination of charge conjugation and parity CP. This means that since strong interactions are described by a non-Abelian gauge theory with group SU(3) there is an extra source of CP violation which puts a strong bound on the value of θ. One of the consequences of a term like (4.117) in the QCD Lagrangian is a non-vanishing electric dipole moment for the neutron [24]. The fact that this is not observed imposes a very strong bound on the value of the θ-parameter

$$|\theta| < 10^{-9}\,. \tag{4.119}$$

From a theoretical point of view it is still to be fully understood why θ either vanishes or has a very small value.

Finally, the θ-vacuum structure of gauge theories that we found in the Hamiltonian formalism can also be obtained using path integral techniques from the Lagrangian (4.117). The second term in Eq. (4.117) gives then a contribution that depends on the winding number of the corresponding gauge configuration.

5 Towards computational rules: Feynman diagrams

As the basic tool to describe the physics of elementary particles, the final aim of quantum field theory is the calculation of observables. Most of the information we have about the physics of subatomic particles comes from scattering experiments. Typically, these experiments consist of arranging two or more particles to collide with a certain energy and to set up an array of detectors, sufficiently far away from the region where the collision takes place, that register the outgoing products of the collision and their momenta (together with other relevant quantum numbers).

Next we discuss how these cross sections can be computed from quantum mechanical amplitudes and how these amplitudes themselves can be evaluated in perturbative quantum field theory. We keep our discussion rather heuristic and avoid technical details that can be found in standard texts [1–10]. The techniques described will be illustrated with the calculation of the cross section for Compton scattering at low energies.

5.1 Cross sections and S-matrix amplitudes

In order to fix ideas let us consider the simplest case of a collision experiment where two particles collide to produce again two particles in the final state. The aim of such an experiment is a direct measurement of the number of particles per unit time $\frac{dN}{dt}(\theta,\varphi)$ registered by the detector flying within a solid angle

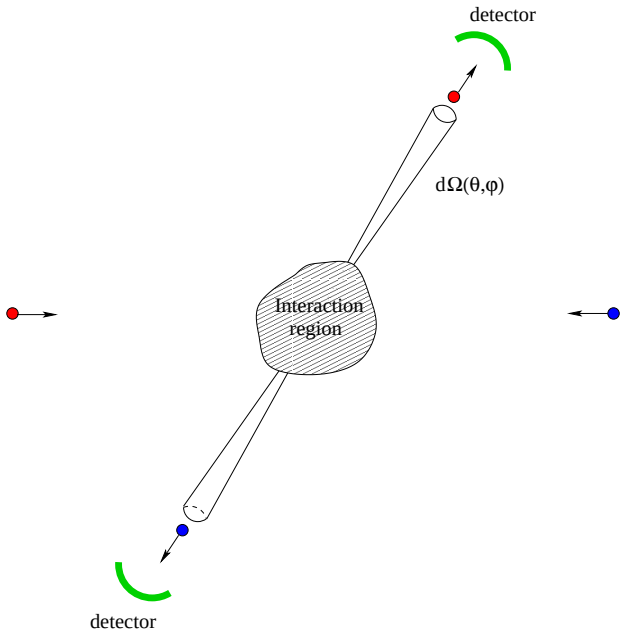


Fig. 10: Schematic setup of a two-to-two-particles single scattering event in the centre-of-mass reference frame

$d\Omega$ in the direction specified by the polar angles θ, φ (see Fig. 10). On general grounds we know that this quantity has to be proportional to the flux of incoming particles[9], $f_{\rm in}$. The proportionality constant defines the differential cross section

$$\frac{dN}{dt}(\theta,\varphi) = f_{\rm in}\frac{d\sigma}{d\Omega}(\theta,\varphi). \tag{5.1}$$

In natural units $f_{\rm in}$ has dimensions of (length)$^{-3}$, and then the differential cross section has dimensions of (length)^2. It depends, apart from the direction (θ,φ), on the parameters of the collision (energy, impact parameter, etc.) as well as on the masses and spins of the incoming particles.

Differential cross sections measure the angular distribution of the products of the collision. It is also physically interesting to quantify how effective the interaction between the particles is to produce a nontrivial dispersion. This is measured by the total cross section, which is obtained by integrating the differential cross section over all directions

$$\sigma = \int_{-1}^{1} d(\cos\theta)\int_{0}^{2\pi} d\varphi\,\frac{d\sigma}{d\Omega}(\theta,\varphi). \tag{5.2}$$

To get some physical intuition of the meaning of the total cross section we can think of the classical scattering of a point particle off a sphere of radius R. The particle undergoes a collision only when the impact parameter is smaller than the radius of the sphere and a calculation of the total cross section yields $\sigma = \pi R^2$. This is precisely the cross area that the sphere presents to incoming particles.

In quantum mechanics in general and in quantum field theory in particular, the starting point for the calculation of cross sections is the probability amplitude for the corresponding process. In a scattering experiment one prepares a system with a given number of particles with definite momenta $\vec{p}_1,\ldots,\vec{p}_n$. In the Heisenberg picture this is described by a time-independent state labelled by the incoming momenta of the particles (to keep things simple we consider spinless particles) that we denote by

$$|\vec{p}_1,\ldots,\vec{p}_n;{\rm in}\rangle. \tag{5.3}$$

[9]This is defined as the number of particles that enter the interaction region per unit time and per unit area perpendicular to the direction of the beam.

On the other hand, as a result of the scattering experiment a number k of particles with momenta $\vec{p}_1{}', \ldots, \vec{p}_k{}'$ are detected. Thus, the system is now in the 'out' Heisenberg picture state

$$|\vec{p}_1{}', \ldots, \vec{p}_k{}'; \text{out}\rangle \tag{5.4}$$

labelled by the momenta of the particles detected at late times. The probability amplitude of detecting k particles in the final state with momenta $\vec{p}_1{}', \ldots, \vec{p}_k{}'$ in the collision of n particles with initial momenta $\vec{p}_1, \ldots, \vec{p}_n$ defines the S-matrix amplitude

$$S(\text{in} \to \text{out}) = \langle \vec{p}_1{}', \ldots, \vec{p}_k{}'; \text{out} | \vec{p}_1, \ldots, \vec{p}_n; \text{in}\rangle. \tag{5.5}$$

It is very important to keep in mind that both the (5.3) and (5.4) are time-independent states in the Hilbert space of a very complicated interacting theory. However, since both at early and late times the incoming and outgoing particles are well apart from each other, the 'in' and 'out' states can be thought of as two states $|\vec{p}_1, \ldots, \vec{p}_n\rangle$ and $|\vec{p}_1{}', \ldots, \vec{p}_k{}'\rangle$ of the Fock space of the corresponding free theory in which the coupling constants are zero. Then, the overlaps (5.5) can be written in terms of the matrix elements of an S-matrix operator $\widehat{S}$ acting on the free Fock space

$$\langle \vec{p}_1{}', \ldots, \vec{p}_k{}'; \text{out} | \vec{p}_1, \ldots, \vec{p}_n; \text{in}\rangle = \langle \vec{p}_1{}', \ldots, \vec{p}_k{}' | \widehat{S} | \vec{p}_1, \ldots, \vec{p}_n\rangle. \tag{5.6}$$

The operator $\widehat{S}$ is unitary, $\widehat{S}^\dagger = \widehat{S}^{-1}$, and its matrix elements are analytic in the external momenta.

In any scattering experiment there is the possibility that the particles do not interact at all and the system is left in the same initial state. Then it is useful to write the S-matrix operator as

$$\widehat{S} = \mathbf{1} + i\widehat{T}, \tag{5.7}$$

where $\mathbf{1}$ represents the identity operator. In this way, all nontrivial interactions are encoded in the matrix elements of the T-operator $\langle \vec{p}_1{}', \ldots, \vec{p}_k{}' | i\widehat{T} | \vec{p}_1, \ldots, \vec{p}_n\rangle$. Since momentum has to be conserved, a global delta function can be factored out from these matrix elements to define the invariant scattering amplitude $i\mathcal{M}$

$$\langle \vec{p}_1{}', \ldots, \vec{p}_k{}' | i\widehat{T} | \vec{p}_1, \ldots, \vec{p}_n\rangle = (2\pi)^4 \delta^{(4)}\left(\sum_{\text{initial}} p_i - \sum_{\text{final}} p'_f \right) i\mathcal{M}(\vec{p}_1, \ldots, \vec{p}_n; \vec{p}_1{}', \ldots, \vec{p}_k{}') \,. \tag{5.8}$$

Total and differential cross sections can be now computed from the invariant amplitudes. Here we consider the most common situation in which two particles with momenta $\vec{p}_1$ and $\vec{p}_2$ collide to produce a number of particles in the final state with momenta $\vec{p}_i{}'$. In this case the total cross section is given by

$$\sigma = \frac{1}{(2\omega_{p_1})(2\omega_{p_2})|\vec{v}_{12}|} \int \left[\prod_{\substack{\text{final} \\ \text{states}}} \frac{d^3 p'_i}{(2\pi)^3} \frac{1}{2\omega_{p'_i}} \right] \left| \mathcal{M}_{i\to f} \right|^2 (2\pi)^4 \delta^{(4)}\left(p_1 + p_2 - \sum_{\substack{\text{final} \\ \text{states}}} p'_i \right), \tag{5.9}$$

where $\vec{v}_{12}$ is the relative velocity of the two scattering particles. The corresponding differential cross section can be computed by dropping the integration over the directions of the final momenta. We will use this expression later in Section 5.3 to evaluate the cross section of Compton scattering.

We see how particle cross sections are determined by the invariant amplitude for the corresponding process, i.e., S-matrix amplitudes. In general, in quantum field theory it is not possible to compute exactly these amplitudes. However, in many physical situations it can be argued that interactions are weak enough to allow for a perturbative evaluation. In what follows we will describe how S-matrix elements can be computed in perturbation theory using Feynman diagrams and rules. These are very convenient book-keeping techniques allowing both to keep track of all contributions to a process at a given order in perturbation theory, and to compute the different contributions.

5.2 Feynman rules

The basic quantities to be computed in quantum field theory are vacuum expectation values of products of the operators of the theory. Particularly useful are time-ordered Green functions,

$$\langle\Omega|T\Big[\mathcal{O}_1(x_1)\dots\mathcal{O}_n(x_n)\Big]|\Omega\rangle, \tag{5.10}$$

where $|\Omega\rangle$ is the the ground state of the theory and the time-ordered product is defined

$$T\Big[\mathcal{O}_i(x)\mathcal{O}_j(y)\Big] = \theta(x^0-y^0)\mathcal{O}_i(x)\mathcal{O}_j(y) + \theta(y^0-x^0)\mathcal{O}_j(y)\mathcal{O}_i(x). \tag{5.11}$$

The generalization to products with more than two operators is straightforward: operators are always multiplied in time order, those evaluated at earlier times always to the right. The interest of these kinds of correlation functions lies in the fact that they can be related to S-matrix amplitudes through the so-called reduction formula. To keep our discussion as simple as possible we will not derive it or even write it down in full detail. Its form for different theories can be found in any textbook. Here suffice it to say that the reduction formula simply states that any S-matrix amplitude can be written in terms of the Fourier transform of a time-ordered correlation function. Morally speaking

$$\begin{gathered}\langle\vec{p}_1{}',\dots,\vec{p}_m{}';\mathrm{out}|\vec{p}_1,\dots,\vec{p}_n;\mathrm{in}\rangle \\ \Downarrow \\ \int d^4x_1\dots\int d^4y_n\langle\Omega|T\Big[\phi(x_1)^\dagger\dots\phi(x_m)^\dagger\phi(y_1)\dots\phi(y_n)\Big]|\Omega\rangle\, e^{ip_1{}'\cdot x_1}\dots e^{-ip_n\cdot y_n},\end{gathered} \tag{5.12}$$

where $\phi(x)$ is the field whose elementary excitations are the particles involved in the scattering.

The reduction formula reduces the problem of computing S-matrix amplitudes to that of evaluating time-ordered correlation functions of field operators. These quantities are easy to compute exactly in the free theory. For an interacting theory the situation is more complicated, however. Using path integrals, the vacuum expectation value of the time-ordered product of a number of operators can be expressed as

$$\langle\Omega|T\Big[\mathcal{O}_1(x_1)\dots\mathcal{O}_n(x_n)\Big]|\Omega\rangle = \frac{\displaystyle\int\mathscr{D}\phi\mathscr{D}\phi^\dagger\,\mathcal{O}_1(x_1)\dots\mathcal{O}_n(x_n)\,e^{iS[\phi,\phi^\dagger]}}{\displaystyle\int\mathscr{D}\phi\mathscr{D}\phi^\dagger\,e^{iS[\phi,\phi^\dagger]}}. \tag{5.13}$$

For a theory with interactions, the path integral neither in the numerator nor in the denominator is Gaussian and they cannot be calculated exactly. However, Eq. (5.13) is still very useful. The action $S[\phi,\phi^\dagger]$ can be split into the free (quadratic) piece and the interaction part

$$S[\phi,\phi^\dagger] = S_0[\phi,\phi^\dagger] + S_{\mathrm{int}}[\phi,\phi^\dagger]. \tag{5.14}$$

All dependence on the coupling constants of the theory comes from the second piece. Expanding now $\exp[iS_{\mathrm{int}}]$ in power series of the coupling constant we find that each term in the series expansion of both the numerator and the denominator has the structure

$$\int\mathscr{D}\phi\mathscr{D}\phi^\dagger\Big[\dots\Big]e^{iS_0[\phi,\phi^\dagger]}, \tag{5.15}$$

where "..." denotes certain monomial of fields. The important point is that now the integration measure only involves the free action, and the path integral in (5.15) is Gaussian and therefore can be computed exactly. The same conclusion can be reached using the operator formalism. In this case the correlation function (5.10) can be expressed in terms of correlation functions of operators in the interaction picture. The advantage of using this picture is that the fields satisfy the free equations of motion and therefore

can be expanded in creation–annihilation operators. The correlations functions are then easily computed using Wick's theorem.

Putting together all the previous ingredients we can calculate S-matrix amplitudes in a perturbative series in the coupling constants of the field theory. This can be done using Feynman diagrams and rules, a very economical way to compute each term in the perturbative expansion of the S-matrix amplitude for a given process. We will not detail the the construction of Feynman rules but just present them heuristically.

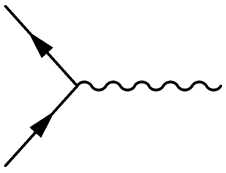
For the sake of concreteness we focus on the case of QED first. Going back to Eq. (4.70) we expand the covariant derivative to write the action

$$S_{\rm QED} = \int d^4x \left[-\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \overline{\psi}(i\not{\partial} - m)\psi + e\overline{\psi}\gamma^\mu \psi A_\mu \right] . \tag{5.16}$$

The action contains two types of particle, fermions and photons, that we represent by straight and wavy lines respectively

The arrow in the fermion line does not represent the direction of the momentum but the flux of (negative) charge. This distinguishes particles from antiparticles: if the fermion propagates from left to right (i.e., in the direction of the charge flux) it represents a particle, whereas when it goes from right to left it corresponds to an antiparticle. Photons are not charged and therefore wavy lines do not have orientation.

Next we turn to the interaction part of the action containing a photon field, a spinor and its conjugate. In a Feynman diagram this corresponds to the vertex



Now, in order to compute an S-matrix amplitude to a given order in the coupling constant e for a process with a certain number of incoming and outgoing asymptotic states one only has to draw all possible diagrams with as many vertices as the order in perturbation theory, and the corresponding number and type of external legs. It is very important to keep in mind that in joining the fermion lines among the different building blocks of the diagram one has to respect their orientation. This reflects the conservation of the electric charge. In addition one should only consider diagrams that are topologically non-equivalent, i.e., they cannot be smoothly deformed into one another keeping the external legs fixed[10].

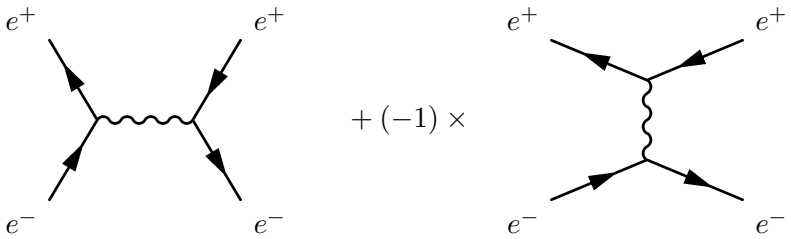
To show in a practical way how Feynman diagrams are drawn, we consider Bhabha scattering, i.e., the elastic dispersion of an electron and a positron:

$$e^+ + e^- \longrightarrow e^+ + e^- .$$

Our problem is to compute the S-matrix amplitude to the leading order in the electric charge. Because the QED vertex contains a photon line and our process has photons neither in the initial nor the final states

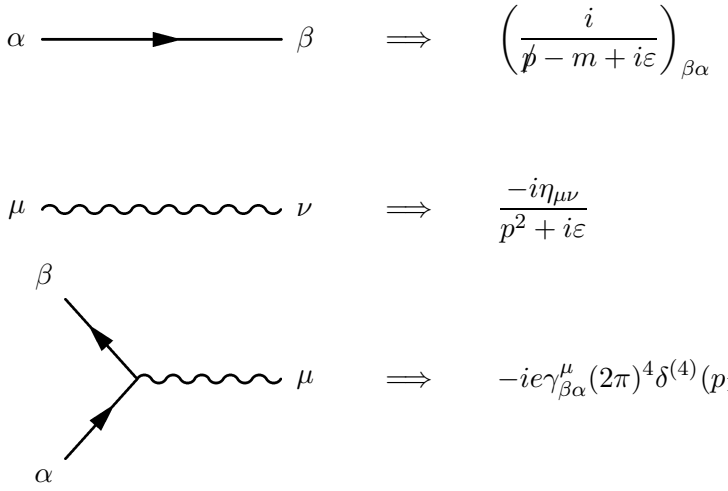
[10]From the point of view of the operator formalism, the requirement of considering only diagrams that are topologically non-equivalent comes from the fact that each diagram represents a certain Wick contraction in the correlation function of interaction-picture operators.

we find that drawing a Feynman diagram requires at least two vertices. In fact, the leading contribution is of order e^2 and comes from the following two diagrams, each containing two vertices:



Incoming and outgoing particles appear respectively on the left and the right of this diagram. Notice how the identification of electrons and positrons is done comparing the direction of the charge flux with the direction of propagation. For electrons the flux of charges goes in the direction of propagation, whereas for positrons the two directions are opposite. These are the only two diagrams that can be drawn at this order in perturbation theory. It is important to include a relative minus sign between the two contributions. To understand the origin of this sign we have to remember that in the operator formalism Feynman diagrams are just a way to encode a particular Wick contraction of field operators in the interaction picture. The factor of -1 reflects the relative sign in Wick contractions represented by the two diagrams, due to the fermionic character of the Dirac field.

We have learned how to draw Feynman diagrams in QED. Now one needs to compute the contribution of each one to the corresponding amplitude using the so-called Feynman rules. The idea is simple: given a diagram, each of its building blocks (vertices as well as external and internal lines) has an associated contribution that allows the calculation of the corresponding diagram. In the case of QED in the Feynman gauge, we have the following correspondence for vertices and internal propagators:



A change in the gauge would be reflected in an extra piece in the photon propagator. The delta function implementing conservation of momenta is written using the convention that all momenta are entering the vertex. In addition, one has to perform an integration over all momenta running in internal lines with the measure

$$\int \frac{d^d p}{(2\pi)^4}, \tag{5.17}$$

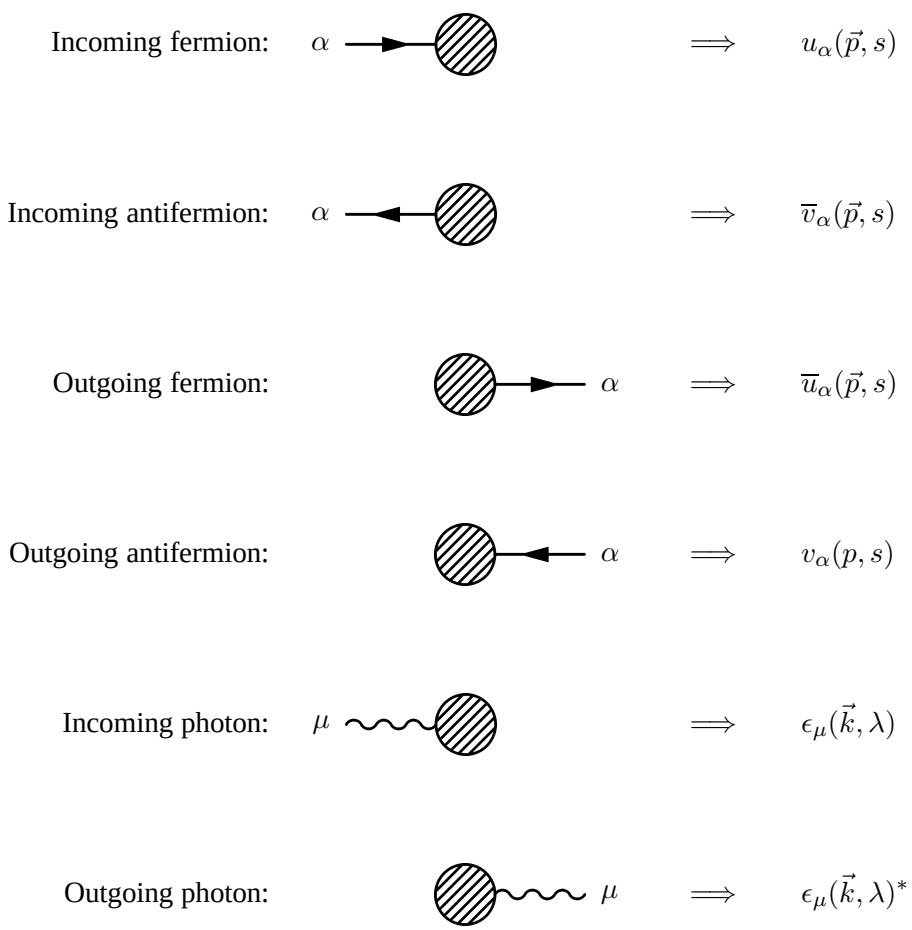
and introduce a factor of -1 for each fermion loop in the diagram[11].

[11]The contribution of each diagram comes also multiplied by a degeneracy factor that takes into account in how many ways a given Wick contraction can be done. In QED, however, these factors are equal to 1 for many diagrams.

In fact, some of the integrations over internal momenta can actually be done using the delta function at the vertices, leaving just a global delta function implementing the total momentum conservation in the diagram [cf. Eq. (5.8)]. It is even possible that all integrations can be eliminated in this way. This is the case when we have tree level diagrams, i.e., those without closed loops. In the case of diagrams with loops there will be as many remaining integrations as the number of independent loops in the diagram.

The need to perform integrations over internal momenta in loop diagrams has important consequences in quantum field theory. The reason is that in many cases the resulting integrals are ill-defined, i.e., are divergent either at small or large values of the loop momenta. In the first case one speaks of *infrared divergences* and usually they cancel once all contributions to a given process are added together. More profound, however, are the divergences appearing at large internal momenta. These *ultraviolet divergences* cannot be cancelled and have to be dealt with through the renormalization procedure. We will discuss this problem in some detail in Section 8.

Were we computing time-ordered (amputated) correlation function of operators, this would be all. However, in the case of S-matrix amplitudes this is not the whole story. In addition to the previous rules here one needs to attach contributions also to the external legs in the diagram. These are the wave functions of the corresponding asymptotic states containing information about the spin and momenta of the incoming and outgoing particles. In the case of QED these contributions are:

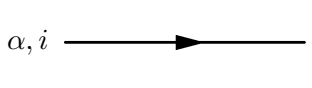


Here we have assumed that the momenta for incoming (outgoing) particles are entering (leaving) the diagram. It is important also to keep in mind that in the computation of S-matrix amplitudes all external

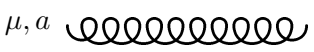
states are on-shell. In Section 5.3 we illustrate the use of Feynman rules for QED with the case of Compton scattering.

The application of Feynman diagrams to carry out computations in perturbation theory is extremely convenient. It provides a very useful book-keeping technique to account for all contributions to a process at a given order in the coupling constant. This does not mean that the calculation of Feynman diagrams is an easy task. The number of diagrams contributing to the process grows very fast with the order in perturbation theory, and the integrals that appear in calculating loop diagrams also get very complicated. This means that, generically, the calculation of Feynman diagrams beyond the first few orders very often requires the use of computers.

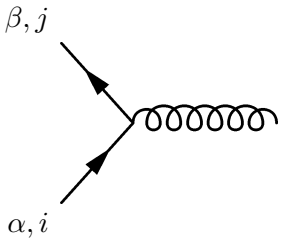
Above we have illustrated the Feynman rules with the case of QED. Similar rules can be computed for other interacting quantum field theories with scalar, vector, or spinor fields. In the case of the non-Abelian gauge theories introduced in Section 4.3 we have:



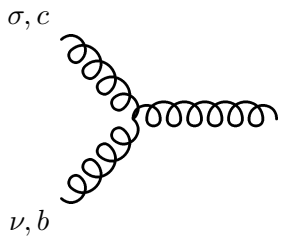
$$\Longrightarrow \qquad \left(\frac{i}{\not{p} - m + i\varepsilon}\right)_{\beta\alpha} \delta_{ij}$$



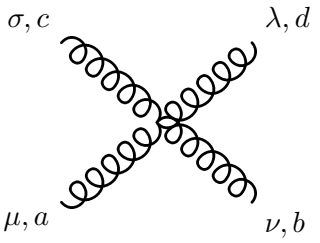
$$\Longrightarrow \qquad \frac{-i\eta_{\mu\nu}}{p^2 + i\varepsilon}\delta^{ab}$$



$$\Longrightarrow \qquad -ig\gamma^{\mu}_{\beta\alpha} t^a_{ij}$$



$$\Longrightarrow \qquad g\, f^{abc}\Big[\eta^{\mu\nu}(p_1^{\sigma} - p_2^{\sigma}) + \text{permutations}\Big]$$



$$\Longrightarrow \qquad -ig^2\Big[f^{abe}f^{cde}\Big(\eta^{\mu\sigma}\eta^{\nu\lambda} - \eta^{\mu\lambda}\eta^{\nu\sigma}\Big) + \text{permutations}\Big]$$

It is not our aim here to give a full and detailed description of the Feynman rules for non-Abelian gauge theories. Suffice it to point out that, unlike the case of QED, here the gauge fields can interact among themselves. Indeed, the three- and four-gauge field vertices are a consequence of the cubic and

quartic terms in the action

$$S = -\frac{1}{4}\int d^4x\, F^a_{\mu\nu}F^{\mu\nu\,a}, \tag{5.18}$$

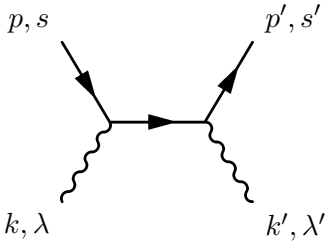
where the non-Abelian gauge field strength $F^a_{\mu\nu}$ is given in Eq. (4.75). The self-interaction of the non-Abelian gauge fields has crucial dynamical consequences and it is at the very heart of its success in describing the physics of elementary particles.

5.3 An example: Compton scattering

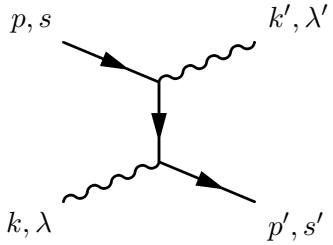
To illustrate the use of Feynman diagrams and Feynman rules we compute the cross section for the dispersion of photons by free electrons, the so-called Compton scattering:

$$\gamma(k,\lambda) + e^-(p,s) \longrightarrow \gamma(k',\lambda') + e^-(p',s').$$

In brackets we have indicated the momenta for the different particles, as well as the polarizations and spins of the incoming and outgoing photon and electrons respectively. The first step is to identify all the diagrams contributing to the process at leading order. Taking into account that the vertex of QED contains two fermion and one photon leg, it is straightforward to realize that any diagram contributing to the process at hand must contain at least two vertices. Hence the leading contribution is of order e^2. A first diagram we can draw is:

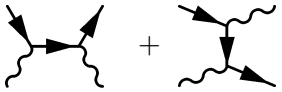


This is, however, not the only possibility. Indeed, there is a second possible diagram:



It is important to stress that these two diagrams are topologically nonequivalent, since deforming one into the other would require changing the label of the external legs. Therefore the leading $\mathcal{O}(e^2)$ amplitude has to be computed adding the contributions from both of them.

Using the Feynman rules of QED we find



$$\begin{aligned} &= (ie)^2\overline{u}(\vec{p}\,',s')\not{\epsilon}\,'(\vec{k}\,',\lambda')^*\frac{\not{p}+\not{k}+m_e}{(p+k)^2-m_e^2}\not{\epsilon}(\vec{k},\lambda)u(\vec{p},s) \\ &+ (ie)^2\overline{u}(\vec{p}\,',s')\not{\epsilon}(\vec{k},\lambda)\frac{\not{p}-\not{k}'+m_e}{(p-k')^2-m_e^2}\not{\epsilon}\,'(\vec{k}\,',\lambda')^*u(\vec{p},s). \end{aligned} \tag{5.19}$$

Because the leading order contributions involve only tree-level diagrams, there is no integration over internal momenta and therefore we are left with a purely algebraic expression for the amplitude. To get

an explicit expression we begin by simplifying the numerators. The following simple identity turns out to be very useful for this task

$$\not{a}\not{b} = -\not{b}\not{a} + 2(a \cdot b)\mathbf{1}. \tag{5.20}$$

Indeed, looking at the first term in Eq. (5.19) we have

$$\begin{aligned}(\not{p} + \not{k} + m_e)\not{\epsilon}(\vec{k},\lambda)u(\vec{p},s) &= -\not{\epsilon}(\vec{k},\lambda)(\not{p} - m_e)u(\vec{p},s) + \not{k}\not{\epsilon}(\vec{k},\lambda)u(\vec{p},s) \\ &+ 2p\cdot\epsilon(\vec{k},\lambda)u(\vec{p},s),\end{aligned} \tag{5.21}$$

where we have applied the identity (5.20) on the first term inside the parenthesis. The first term on the right-hand side of this equation vanishes identically because of Eq. (4.35). The expression can be further simplified if we restrict our attention to the Compton scattering at low energy when electrons are nonrelativistic. This means that all spatial momenta are much smaller than the electron mass

$$|\vec{p}|, |\vec{k}|, |\vec{p}\,'|, |\vec{k}\,'| \ll m_e. \tag{5.22}$$

In this approximation we have that $p^\mu, p'^\mu \approx (m_e, \vec{0})$ and therefore

$$p\cdot\epsilon(\vec{k},\lambda) = 0. \tag{5.23}$$

This follows from the absence of temporal photon polarization. Then we conclude that at low energies

$$(\not{p} + \not{k} + m_e)\not{\epsilon}(\vec{k},\lambda)u(\vec{p},s) = \not{k}\not{\epsilon}(\vec{k},\lambda)u(\vec{p},s) \tag{5.24}$$

and similarly for the second term in Eq. (5.19)

$$(\not{p} - \not{k}' + m_e)\not{\epsilon}'(\vec{k}',\lambda')^* u(\vec{p},s) = -\not{k}'\not{\epsilon}'(\vec{k}',\lambda')^* u(\vec{p},s). \tag{5.25}$$

Next, we turn to the denominators in Eq. (5.19). As it was explained in Section 5.2, in computing scattering amplitudes incoming and outgoing particles should have on-shell momenta,

$$p^2 = m_e^2 = p'^2 \qquad \text{and} \qquad k^2 = 0 = k'^2. \tag{5.26}$$

Then, the two denominators in Eq. (5.19) simplify respectively to

$$(p+k)^2 - m_e^2 = p^2 + k^2 + 2p\cdot k - m_e^2 = 2p\cdot k = 2\omega_p|\vec{k}| - 2\vec{p}\cdot\vec{k} \tag{5.27}$$

and

$$(p-k')^2 - m_e^2 = p^2 + k'^2 + 2p\cdot k' - m_e^2 = -2p\cdot k' = -2\omega_p|\vec{k}\,'| + 2\vec{p}\cdot\vec{k}\,'. \tag{5.28}$$

Working again in the low-energy approximation (5.22) these two expressions simplify to

$$(p+k)^2 - m_e^2 \approx 2m_e|\vec{k}|, \qquad (p-k')^2 - m_e^2 \approx -2m_e|\vec{k}\,'|. \tag{5.29}$$

Putting together all these expressions we find that at low energies

$$\approx \frac{(ie)^2}{2m_e}\overline{u}(\vec{p}\,',s')\left[\not{\epsilon}'(\vec{k}\,'\lambda')^*\frac{\not{k}}{|\vec{k}|}\epsilon(\vec{k},\lambda) + \epsilon(\vec{k},\lambda)\frac{\not{k}'}{|\vec{k}\,'|}\not{\epsilon}'(\vec{k}\,'\lambda')^*\right]u(\vec{p},s). \tag{5.30}$$

Using now again the identity (5.20) a number of times as well as the transversality condition of the polarization vectors (4.66) we end up with a handier equation

$$\approx \frac{e^2}{m_e}\Big[\epsilon(\vec{k},\lambda)\cdot\epsilon'(\vec{k}\,',\lambda')^*\Big]\overline{u}(\vec{p}\,',s')\frac{\not{k}}{|\vec{k}|}u(\vec{p},s)$$
$$+ \frac{e^2}{2m_e}\overline{u}(\vec{p}\,',s')\not{\epsilon}(\vec{k},\lambda)\not{\epsilon}\,'(\vec{k}\,',\lambda')^*\left(\frac{\not{k}}{|\vec{k}|}-\frac{\not{k}'}{|\vec{k}\,'|}\right)u(\vec{p},s). \qquad (5.31)$$

With a little bit of effort we can show that the second term on the right-hand side vanishes. First we notice that in the low-energy limit $|\vec{k}|\approx|\vec{k}\,'|$. If in addition we make use of the conservation of momentum $k-k'=p'-p$ and the identity (4.35)

$$\overline{u}(\vec{p}\,',s')\not{\epsilon}(\vec{k},\lambda)\not{\epsilon}\,'(\vec{k}\,',\lambda')^*\left(\frac{\not{k}}{|\vec{k}|}-\frac{\not{k}'}{|\vec{k}\,'|}\right)u(\vec{p},s)$$
$$\approx \frac{1}{|\vec{k}|}\overline{u}(\vec{p}\,',s')\not{\epsilon}(\vec{k},\lambda)\not{\epsilon}\,'(\vec{k}\,',\lambda')^*(\not{p}'-m_e)u(\vec{p},s). \qquad (5.32)$$

Next we use the identity (5.20) to take the term $(\not{p}'-m_e)$ to the right. Taking into account that in the low-energy limit the electron four-momenta are orthogonal to the photon polarization vectors [see Eq. (5.23)] we conclude that

$$\overline{u}(\vec{p}\,',s')\not{\epsilon}(\vec{k},\lambda)\not{\epsilon}\,'(\vec{k}\,',\lambda')^*(\not{p}'-m_e)u(\vec{p},s)$$
$$=\overline{u}(\vec{p}\,',s')(\not{p}'-m_e)\not{\epsilon}(\vec{k},\lambda)\not{\epsilon}\,'(\vec{k}\,',\lambda')^*u(\vec{p},s)=0 \qquad (5.33)$$

where the last identity follows from the equation satisfied by the conjugate positive-energy spinor, $\overline{u}(\vec{p}\,',s')(\not{p}'-m_e)=0$.

After all these lengthy manipulations we have finally arrived at the expression of the invariant amplitude for Compton scattering at low energies

$$i\mathcal{M}=\frac{e^2}{m_e}\Big[\epsilon(\vec{k},\lambda)\cdot\epsilon'(\vec{k}\,',\lambda')^*\Big]\overline{u}(\vec{p}\,',s')\frac{\not{k}}{|\vec{k}|}u(\vec{p},s). \qquad (5.34)$$

The calculation of the cross section involves computing the modulus squared of this quantity. For many physical applications, however, one is interested in the dispersion of photons with a given polarization by electrons that are not polarized, i.e., whose spins are randomly distributed. In addition, in many situations either we are not interested, or there is no way to measure the final polarization of the outgoing electron. This is, for example, the situation in cosmology, where we do not have any information about the polarization of the free electrons in the primordial plasma before or after the scattering with photons (although we have ways to measure the polarization of the scattered photons).

To describe this physical situation we have to average over initial electron polarizations (since we do not know them) and sum over all possible final electron polarizations (because our detector is blind to this quantum number),

$$\overline{|i\mathcal{M}|^2}=\frac{1}{2}\left(\frac{e^2}{m_e|\vec{k}|}\right)^2\Big|\epsilon(\vec{k},\lambda)\cdot\epsilon'(\vec{k}\,',\lambda')^*\Big|^2\sum_{s=\pm\frac{1}{2}}\sum_{s'=\pm\frac{1}{2}}\Big|\overline{u}(\vec{p}\,',s')\not{k}u(\vec{p},s)\Big|^2. \qquad (5.35)$$

The factor of $\frac{1}{2}$ comes from averaging over the two possible polarizations of the incoming electrons. The sums in this expression can be calculated without much difficulty. Expanding the absolute value explicitly

$$\sum_{s=\pm\frac{1}{2}}\sum_{s'=\pm\frac{1}{2}}\Big|\overline{u}(\vec{p}\,',s')\not{k}u(\vec{p},s)\Big|^2=\sum_{s=\pm\frac{1}{2}}\sum_{s'=\pm\frac{1}{2}}\Big[u(\vec{p},s)^\dagger\not{k}^\dagger\overline{u}(\vec{p}\,',s')^\dagger\Big]\Big[\overline{u}(\vec{p}\,',s')\not{k}u(\vec{p},s)\Big], \qquad (5.36)$$

using that $\gamma^{\mu\dagger} = \gamma^0\gamma^\mu\gamma^0$ and after some manipulation one finds that

$$\sum_{s=\pm\frac{1}{2}}\sum_{s'=\pm\frac{1}{2}}\left|\overline{u}(\vec{p}',s')\not{k}u(\vec{p},s)\right|^2 = \left[\sum_{s=\pm\frac{1}{2}}u_\alpha(\vec{p},s)\overline{u}_\beta(\vec{p},s)\right](\not{k})_{\beta\sigma}\left[\sum_{s'=\pm\frac{1}{2}}u_\sigma(\vec{p}',s')\overline{u}_\rho(\vec{p}',s')\right](\not{k})_{\rho\alpha}$$
$$= \mathrm{Tr}\left[(\not{p}+m_e)\not{k}(\not{p}'+m_e)\not{k}\right], \tag{5.37}$$

where the final expression has been computed using the completeness relations in Eq. (4.38). The final evaluation of the trace can be done using the standard Dirac matrices identities. Here we compute it applying again the relation (5.20) to commute $\not{p}'$ and $\not{k}$. Using that $k^2 = 0$ and that we are working in the low-energy limit we have[12]

$$\mathrm{Tr}\left[(\not{p}+m_e)\not{k}(\not{p}'+m_e)\not{k}\right] = 2(p\cdot k)(p'\cdot k)\mathrm{Tr}\,\mathbf{1} \approx 8m_e^2|\vec{k}|^2. \tag{5.38}$$

This gives the following value for the invariant amplitude

$$\overline{|i\mathcal{M}|^2} = 4e^4\left|\epsilon(\vec{k},\lambda)\cdot\epsilon'(\vec{k}',\lambda')^*\right|^2 \tag{5.39}$$

Plugging $\overline{|i\mathcal{M}|^2}$ into the formula for the differential cross section we get

$$\frac{d\sigma}{d\Omega} = \frac{1}{64\pi^2m_e^2}\overline{|i\mathcal{M}|^2} = \left(\frac{e^2}{4\pi m_e}\right)^2\left|\epsilon(\vec{k},\lambda)\cdot\epsilon'(\vec{k}',\lambda')^*\right|^2. \tag{5.40}$$

The prefactor of the last equation is precisely the square of the so-called classical electron radius r_{cl}. In fact, the previous differential cross section can be rewritten as

$$\frac{d\sigma}{d\Omega} = \frac{3}{8\pi}\sigma_T\left|\epsilon(\vec{k},\lambda)\cdot\epsilon'(\vec{k}',\lambda')^*\right|^2, \tag{5.41}$$

where σ_T is the total Thomson cross section

$$\sigma_T = \frac{e^4}{6\pi m_e^2} = \frac{8\pi}{3}r_{\mathrm{cl}}^2. \tag{5.42}$$

The result (5.41) is relevant in many areas of Physics, but its importance is paramount in the study of the cosmological microwave background (CMB). Just before recombination the universe is filled by a plasma of electrons interacting with photons via Compton scattering, with temperatures of the order of 1 keV. Electrons are then non-relativistic ($m_e \sim 0.5$ MeV) and the approximations leading to Eq. (5.41) are fully valid. Because we do not know the polarization state of the photons before being scattered by electrons we have to consider the cross section averaged over incoming photon polarizations. From Eq. (5.41) we see that this is proportional to

$$\frac{1}{2}\sum_{\lambda=1,2}\left|\epsilon(\vec{k},\lambda)\cdot\epsilon'(\vec{k}',\lambda')^*\right|^2 = \left[\frac{1}{2}\sum_{\lambda=1,2}\epsilon_i(\vec{k},\lambda)\epsilon_j(\vec{k},\lambda)^*\right]\epsilon_j(\vec{k}',\lambda')\epsilon_i(\vec{k}',\lambda')^*. \tag{5.43}$$

The sum inside the brackets can be computed using the normalization condition of the polarization vectors, $|\vec{\epsilon}(\vec{k},\lambda)|^2 = 1$, and the transversality condition $\vec{k}\cdot\vec{\epsilon}(\vec{k},\lambda) = 0$

$$\frac{1}{2}\sum_{\lambda=1,2}\left|\epsilon(\vec{k},\lambda)\cdot\epsilon'(\vec{k}',\lambda')^*\right|^2 = \frac{1}{2}\left(\delta_{ij}-\frac{k_ik_j}{|\vec{k}|^2}\right)\epsilon'_j(\vec{k}',\lambda')\epsilon'_i(\vec{k}',\lambda')^*$$

[12] We also use the fact that the trace of the product of an odd number of Dirac matrices is always zero.

$$= \frac{1}{2}\Big[1 - |\vec{\ell}\cdot\vec{\epsilon}\,'(\vec{k}\,',\lambda')|^2\Big], \tag{5.44}$$

where $\vec{\ell} = \frac{\vec{k}}{|\vec{k}|}$ is the unit vector in the direction of the incoming photon.

From the last equation we conclude that Thomson scattering suppresses all polarizations parallel to the direction of the incoming photon $\vec{\ell}$, whereas the differential cross section reaches the maximum in the plane normal to $\vec{\ell}$. If photons would collide with the electrons in the plasma with the same intensity from all directions, the result would be an unpolarized CMB radiation. The fact that polarization is actually measured in the CMB carries crucial information about the physics of the plasma before recombination and, as a consequence, about the very early universe (see for example [25] for a thorough discussion).

6 Symmetries

6.1 Noether's theorem

In classical mechanics and classical field theory there is a basic result that relates symmetries and conserved charges. This is called Noether's theorem and states that for each continuous symmetry of the system there is conserved current. In its simplest version in classical mechanics it can be easily proved. Let us consider a Lagrangian $L(q_i, \dot{q}_i)$ which is invariant under a transformation $q_i(t) \to q_i'(t,\epsilon)$ labelled by a parameter ϵ. This means that $L(q', \dot{q}') = L(q, \dot{q})$ without using the equations of motion[13]. If $\epsilon \ll 1$ we can consider an infinitesimal variation of the coordinates $\delta_\epsilon q_i(t)$ and the invariance of the Lagrangian implies

$$0 = \delta_\epsilon L(q_i,\dot{q}_i) = \frac{\partial L}{\partial q_i}\delta_\epsilon q_i + \frac{\partial L}{\partial \dot{q}_i}\delta_\epsilon \dot{q}_i = \left[\frac{\partial L}{\partial q_i} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i}\right]\delta_\epsilon q_i + \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\delta_\epsilon q_i\right). \tag{6.1}$$

When $\delta_\epsilon q_i$ is applied on a solution to the equations of motion the term inside the square brackets vanishes and we conclude that there is a conserved quantity

$$\dot{Q} = 0 \quad \text{with} \quad Q \equiv \frac{\partial L}{\partial \dot{q}_i}\delta_\epsilon q_i. \tag{6.2}$$

Notice that in this derivation it is crucial that the symmetry depends on a continuous parameter since otherwise the infinitesimal variation of the Lagrangian in Eq. (6.1) does not make sense.

In classical field theory a similar result holds. Let us consider for simplicity a theory of a single field $\phi(x)$. We say that the variations $\delta_\epsilon\phi$ depending on a continuous parameter ϵ are a symmetry of the theory if, without using the equations of motion, the Lagrangian density changes by

$$\delta_\epsilon \mathcal{L} = \partial_\mu K^\mu. \tag{6.3}$$

If this happens then the action remains invariant and so do the equations of motion. Working out now the variation of $\mathcal{L}$ under $\delta_\epsilon\phi$ we find

$$\partial_\mu K^\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\mu\phi)}\partial_\mu\delta_\epsilon\phi + \frac{\partial\mathcal{L}}{\partial\phi}\delta_\epsilon\phi = \partial_\mu\left(\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\delta_\epsilon\phi\right) + \left[\frac{\partial\mathcal{L}}{\partial\phi} - \partial_\mu\left(\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\right)\right]\delta_\epsilon\phi. \tag{6.4}$$

If $\phi(x)$ is a solution to the equations of motion, the last terms disappears and we find that there is a conserved current

$$\partial_\mu J^\mu = 0 \quad \text{with} \quad J^\mu = \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\delta_\epsilon\phi - K^\mu. \tag{6.5}$$

[13] The following result can also be derived in more general situations where the Lagrangian changes by a total time derivative.

Actually a conserved current implies the existence of a charge

$$Q \equiv \int d^3x\, J^0(t,\vec{x}) \tag{6.6}$$

which is conserved

$$\frac{dQ}{dt} = \int d^3x\, \partial_0 J^0(t,\vec{x}) = -\int d^3x\, \partial_i J^i(t,\vec{x}) = 0, \tag{6.7}$$

provided the fields vanish at infinity fast enough. Moreover, the conserved charge Q is a Lorentz scalar. After canonical quantization the charge Q defined by Eq. (6.6) is promoted to an operator that generates the symmetry on the fields

$$\delta\phi = i[\phi, Q]. \tag{6.8}$$

As an example we can consider a scalar field $\phi(x)$ which under a coordinate transformation $x \to x'$ changes as $\phi'(x') = \phi(x)$. In particular performing a space-time translation $x^{\mu\,'} = x^\mu + a^\mu$ we have

$$\phi'(x) - \phi(x) = -a^\mu \partial_\mu \phi + \mathcal{O}(a^2) \qquad \Longrightarrow \qquad \delta\phi = -a^\mu \partial_\mu \phi. \tag{6.9}$$

Since the Lagrangian density is also a scalar quantity, it transforms under translations as

$$\delta\mathcal{L} = -a^\mu \partial_\mu \mathcal{L}. \tag{6.10}$$

Therefore the corresponding conserved charge is

$$J^\mu = -\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)} a^\nu \partial_\nu \phi + a^\mu \mathcal{L} \equiv -a_\nu T^{\mu\nu}, \tag{6.11}$$

where we introduced the energy-momentum tensor

$$T^{\mu\nu} = \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)} \partial^\nu \phi - \eta^{\mu\nu}\mathcal{L}. \tag{6.12}$$

We find that associated with the invariance of the theory with respect to space-time translations there are four conserved currents defined by $T^{\mu\nu}$ with $\nu = 0, \ldots, 3$, each one associated with the translation along a space-time direction. These four currents form a rank-two tensor under Lorentz transformations satisfying

$$\partial_\mu T^{\mu\nu} = 0. \tag{6.13}$$

The associated conserved charges are given by

$$P^\nu = \int d^3x\, T^{0\nu} \tag{6.14}$$

and correspond to the total energy-momentum content of the field configuration. Therefore the energy density of the field is given by T^{00} while T^{0i} is the momentum density. In the quantum theory the P^μ are the generators of space-time translations.

Another example of a symmetry related with a physically relevant conserved charge is the global phase invariance of the Dirac Lagrangian (4.27), $\psi \to e^{i\theta}\psi$. For small θ this corresponds to variations $\delta_\theta\psi = i\theta\psi$, $\delta_\theta\overline{\psi} = -i\theta\overline{\psi}$ which by Noether's theorem result in the conserved charge

$$j^\mu = \overline{\psi}\gamma^\mu\psi, \qquad \partial_\mu j^\mu = 0. \tag{6.15}$$

Thus implying the existence of a conserved charge

$$Q = \int d^3x \overline{\psi}\gamma^0\psi = \int d^3x \psi^\dagger\psi. \tag{6.16}$$

In physics there are several instances of global U(1) symmetries that act as phase shifts on spinors. This is the case, for example, of the baryon and lepton number conservation in the Standard Model. A more familiar case is the U(1) local symmetry associated with electromagnetism. Notice that although in this case we are dealing with a local symmetry $\theta \to e\alpha(x)$, the invariance of the Lagrangian holds in particular for global transformations and therefore there is a conserved current $j^\mu = e\overline{\psi}\gamma^\mu\psi$. In Eq. (4.72) we saw that the spinor is coupled to the photon field precisely through this current. Its time component is the electric charge density ρ, while the spatial components are the current density vector $\vec{j}$.

This analysis can be carried over also to non-Abelian unitary global symmetries acting as

$$\psi_i \longrightarrow U_{ij}\psi_j, \qquad U^\dagger U = \mathbf{1} \tag{6.17}$$

and leaving invariant the Dirac Lagrangian when we have several fermions. If we write the matrix U in terms of the hermitian group generators T^a as

$$U = \exp\left(i\alpha_a T^a\right), \qquad (T^a)^\dagger = T^a, \tag{6.18}$$

we find the conserved current

$$j^{\mu\,a} = \overline{\psi}_i T^a_{ij}\gamma^\mu\psi_j, \qquad \partial_\mu j^\mu = 0. \tag{6.19}$$

This is the case, for example of the approximate flavor symmetries in hadron physics. The simplest example is the isospin symmetry that mixes the quarks u and d

$$\begin{pmatrix} u \\ d \end{pmatrix} \longrightarrow M \begin{pmatrix} u \\ d \end{pmatrix}, \qquad M \in \mathrm{SU}(2). \tag{6.20}$$

Since the proton is a bound state of two quarks u and one quark d while the neutron is made out of one quark u and two quarks d, this isospin symmetry reduces at low energies to the well-known isospin transformations of nuclear physics that mixes protons and neutrons.

6.2 Symmetries in the quantum theory

We have seen that in canonical quantization the conserved charges Q^a associated to symmetries by Noether's theorem are operators implementing the symmetry at the quantum level. Since the charges are conserved they must commute with the Hamiltonian

$$[Q^a, H] = 0. \tag{6.21}$$

There are several possibilities in the quantum mechanical realization of a symmetry:

Wigner–Weyl realization. In this case the ground state of the theory $|0\rangle$ is invariant under the symmetry. Since the symmetry is generated by Q^a this means that

$$\mathcal{U}(\alpha)|0\rangle \equiv e^{i\alpha_a Q^a}|0\rangle = |0\rangle \qquad \Longrightarrow \qquad Q^a|0\rangle = 0. \tag{6.22}$$

At the same time the fields of the theory have to transform according to some irreducible representation of the group generated by the Q^a. From Eq. (6.8) it is easy to prove that

$$\mathcal{U}(\alpha)\phi_i\mathcal{U}(\alpha)^{-1} = U_{ij}(\alpha)\phi_j, \tag{6.23}$$

where $U_{ij}(\alpha)$ is an element of the representation in which the field ϕ_i transforms. If we consider now the quantum state associated with the operator ϕ_i

$$|i\rangle = \phi_i|0\rangle \tag{6.24}$$

we find that because of the invariance of the vacuum (6.22) the states $|i\rangle$ transform in the same representation as ϕ_i

$$\mathcal{U}(\alpha)|i\rangle = \mathcal{U}(\alpha)\phi_i\mathcal{U}(\alpha)^{-1}\mathcal{U}(\alpha)|0\rangle = U_{ij}(\alpha)\phi_j|0\rangle = U_{ij}(\alpha)|j\rangle. \tag{6.25}$$

Therefore the spectrum of the theory is classified in multiplets of the symmetry group. In addition, since $[H,\mathcal{U}(\alpha)] = 0$ all states in the same multiplet have the same energy. If we consider one-particle states, then going to the rest frame we conclude that all states in the same multiplet have exactly the same mass.

Nambu–Goldstone realization. In our previous discussion the result that the spectrum of the theory is classified according to multiplets of the symmetry group depended crucially on the invariance of the ground state. However this condition is not mandatory and one can relax it to consider theories where the vacuum state is not left invariant by the symmetry

$$e^{i\alpha_a Q^a}|0\rangle \neq |0\rangle \qquad \Longrightarrow \qquad Q^a|0\rangle \neq 0. \tag{6.26}$$

In this case it is also said that the symmetry is spontaneously broken by the vacuum.

To illustrate the consequences of (6.26) we consider the example of a number of scalar fields φ^i ($i = 1,\ldots,N$) whose dynamics is governed by the Lagrangian

$$\mathcal{L} = \frac{1}{2}\partial_\mu\varphi^i\partial^\mu\varphi^i - V(\varphi), \tag{6.27}$$

where we assume that $V(\phi)$ is bounded from below. This theory is globally invariant under the transformations

$$\delta\varphi^i = \epsilon^a(T^a)^i_j\varphi^j, \tag{6.28}$$

with T^a, $a = 1,\ldots,\frac{1}{2}N(N-1)$ the generators of the group SO(N).

To analyse the structure of vacua of the theory we construct the Hamiltonian

$$H = \int d^3x\left[\frac{1}{2}\pi^i\pi^i + \frac{1}{2}\vec{\nabla}\varphi^i\cdot\vec{\nabla}\varphi^i + V(\varphi)\right] \tag{6.29}$$

and look for the minimum of

$$\mathcal{V}(\varphi) = \int d^3x\left[\frac{1}{2}\vec{\nabla}\varphi^i\cdot\vec{\nabla}\varphi^i + V(\varphi)\right]. \tag{6.30}$$

Since we are interested in finding constant field configurations, $\vec{\nabla}\varphi = \vec{0}$ to preserve translational invariance, the vacua of the potential $\mathcal{V}(\varphi)$ coincides with the vacua of $V(\varphi)$. Therefore the minima of the potential correspond to the vacuum expectation values[14]

$$\langle\varphi^i\rangle: \qquad V(\langle\varphi^i\rangle) = 0, \qquad \left.\frac{\partial V}{\partial\varphi^i}\right|_{\varphi^i=\langle\varphi^i\rangle} = 0. \tag{6.31}$$

We divide the generators T^a of SO(N) into two groups: Those denoted by H^α ($\alpha = 1,\ldots,h$) that satisfy

$$(H^\alpha)^i_j\langle\varphi^j\rangle = 0. \tag{6.32}$$

[14]For simplicity we consider that the minima of $V(\phi)$ occur at zero potential.

This means that the vacuum configuration $\langle\varphi^i\rangle$ is left invariant by the transformation generated by H^α. For this reason we call them *unbroken generators*. Notice that the commutator of two unbroken generators also annihilates the vacuum expectation value, $[H^\alpha, H^\beta]_{ij}\langle\varphi^j\rangle = 0$. Therefore the generators $\{H^\alpha\}$ form a subalgebra of the algebra of the generators of SO(N). The subgroup of the symmetry group generated by them is realized à la Wigner–Weyl.

The remaining generators K^A, with $A = 1, \ldots, \frac{1}{2}N(N-1) - h$, by definition do not preserve the vacuum expectation value of the field

$$(K^A)^i_j\langle\varphi^j\rangle \neq 0. \tag{6.33}$$

These will be called the *broken generators*. Next we prove a very important result concerning the broken generators known as the Goldstone theorem: for each generator broken by the vacuum expectation value there is a massless excitation.

The mass matrix of the excitations around the vacuum $\langle\varphi^i\rangle$ is determined by the quadratic part of the potential. Since we assumed that $V(\langle\varphi\rangle) = 0$ and we are expanding around a minimum, the first term in the expansion of the potential $V(\varphi)$ around the vacuum expectation values is given by

$$V(\varphi) = \left.\frac{\partial^2 V}{\partial\varphi^i\partial\varphi^j}\right|_{\varphi=\langle\varphi\rangle}(\varphi^i - \langle\varphi^i\rangle)(\varphi^j - \langle\varphi^j\rangle) + \mathcal{O}\left[(\varphi - \langle\varphi\rangle)^3\right] \tag{6.34}$$

and the mass matrix is

$$M^2_{ij} \equiv \left.\frac{\partial^2 V}{\partial\varphi^i\partial\varphi^j}\right|_{\varphi=\langle\varphi\rangle}. \tag{6.35}$$

In order to avoid a cumbersome notation we do not show explicitly the dependence of the mass matrix on the vacuum expectation values $\langle\varphi^i\rangle$.

To extract some information about the possible zero modes of the mass matrix, we write down the conditions that follow from the invariance of the potential under $\delta\varphi^i = \epsilon^a(T^a)^i_j\varphi^j$. At first order in ϵ^a

$$\delta V(\varphi) = \epsilon^a\frac{\partial V}{\partial\varphi^i}(T^a)^i_j\varphi^j = 0. \tag{6.36}$$

Differentiating this expression with respect to φ^k we arrive at

$$\frac{\partial^2 V}{\partial\varphi^i\partial\varphi^k}(T^a)^i_j\varphi^j + \frac{\partial V}{\partial\varphi^i}(T^a)^i_k = 0. \tag{6.37}$$

Now we evaluate this expression in the vacuum $\varphi^i = \langle\varphi^i\rangle$. Then the derivative in the second term cancels while the second derivative in the first one gives the mass matrix. Hence we find

$$M^2_{ik}(T^a)^i_j\langle\varphi^j\rangle = 0. \tag{6.38}$$

Now we can write this expression for both broken and unbroken generators. For the unbroken ones, since $(H^\alpha)^i_j\langle\varphi^j\rangle = 0$, we find a trivial identity $0 = 0$. On the other hand for the broken generators we have

$$M^2_{ik}(K^A)^i_j\langle\varphi^j\rangle = 0. \tag{6.39}$$

Since $(K^A)^i_j\langle\varphi^j\rangle \neq 0$ this equation implies that the mass matrix has as many zero modes as broken generators. Therefore we have proven Goldstone's theorem: associated with each broken symmetry there is a massless mode in the theory. Here we have presented a classical proof of the theorem. In the quantum theory the proof follows the same lines as the one presented here but one has to consider the effective action containing the effects of the quantum corrections to the classical Lagrangian.

As an example to illustrate this theorem, we consider a SO(3) invariant scalar field theory with a 'mexican hat' potential

$$V(\vec{\varphi}) = \frac{\lambda}{4}\left(\vec{\varphi}^{\,2} - a^2\right)^2 . \tag{6.40}$$

The vacua of the theory correspond to the configurations satisfying $\langle\vec{\varphi}\rangle^{\,2} = a^2$. In field space this equation describes a two-dimensional sphere and each solution is just a point in that sphere. Geometrically it is easy to visualize that a given vacuum field configuration, i.e., a point in the sphere, is preserved by SO(2) rotations around the axis of the sphere that passes through that point. Hence the vacuum expectation value of the scalar field breaks the symmetry according to

$$\langle\vec{\varphi}\rangle : \qquad \mathrm{SO}(3) \longrightarrow \mathrm{SO}(2). \tag{6.41}$$

Since SO(3) has three generators and SO(2) only one, we see that two generators are broken and therefore there are two massless Goldstone bosons. Physically these massless modes can be thought of as corresponding to excitations along the surface of the sphere $\langle\vec{\varphi}\rangle^{\,2} = a^2$.

Once a minimum of the potential has been chosen we can proceed to quantize the excitations around it. Since the vacuum only leaves invariant a SO(2) subgroup of the original SO(3) symmetry group it seems that the fact that we are expanding around a particular vacuum expectation value of the scalar field has resulted in a loss of symmetry. This is, however, not the case. The full quantum theory is symmetric under the whole symmetry group SO(3). This is reflected in the fact that the physical properties of the theory do not depend on the particular point of the sphere $\langle\vec{\varphi}\rangle^{\,2} = a^2$ that we have chosen. Different vacua are related by the full SO(3) symmetry and therefore should give the same physics.

It is very important to realize that given a theory with a vacuum determined by $\langle\vec{\varphi}\rangle$ all other possible vacua of the theory are unaccessible in the infinite volume limit. This means that two vacuum states $|0_1\rangle$, $|0_2\rangle$ corresponding to different vacuum expectation values of the scalar field are orthogonal $\langle 0_1|0_2\rangle = 0$ and cannot be connected by any local observable $\Phi(x)$, $\langle 0_1|\Phi(x)|0_2\rangle = 0$. Heuristically this can be understood by noticing that in the infinite volume limit switching from one vacuum into another one requires changing the vacuum expectation value of the field everywhere in space at the same time, something that cannot be done by any local operator. Notice that this is radically different from our expectations based on the quantum mechanics of a system with a finite number of degrees of freedom.

In high-energy physics the typical example of a Goldstone boson is the pion, associated with the spontaneous breaking of the global chiral isospin $\mathrm{SU}(2)_L \times \mathrm{SU}(2)_R$ symmetry. This symmetry acts independently in the left- and right-handed spinors as

$$\begin{pmatrix} u_{L,R} \\ d_{L,R} \end{pmatrix} \longrightarrow M_{L,R} \begin{pmatrix} u_{L,R} \\ d_{L,R} \end{pmatrix}, \qquad M_{L,R} \in \mathrm{SU}(2)_{L,R} . \tag{6.42}$$

Presumably since the quarks are confined at low energies this symmetry is spontaneously broken down to the diagonal SU(2) acting in the same way on the left- and right-handed components of the spinors. Associated with this symmetry breaking there is a Goldstone mode which is identified as the pion. Notice, nevertheless, that the $\mathrm{SU}(2)_L \times \mathrm{SU}(2)_R$ would be an exact global symmetry of the QCD Lagrangian only in the limit when the masses of the quarks are zero $m_u, m_d \to 0$. Since these quarks have non-zero masses the chiral symmetry is only approximate and as a consequence the corresponding Goldstone boson is not massless. That is why pions have masses, although they are the lightest particle among the hadrons.

Symmetry breaking appears also in many places in condensed matter. For example, when a solid crystallizes from a liquid the translational invariance that is present in the liquid phase is broken to a discrete group of translations that represent the crystal lattice. This symmetry breaking has Goldstone

bosons associated which are identified with phonons which are the quantum excitation modes of the vibrational degrees of freedom of the lattice.

The Higgs mechanism. Gauge symmetry seems to prevent a vector field from having a mass. This is obvious once we realize that a term in the Lagrangian like $m^2 A_\mu A^\mu$ is incompatible with gauge invariance.

However, certain physical situations seem to require massive vector fields. This happened for example during the 1960s in the study of weak interactions. The Glashow model gave a common description of both electromagnetic and weak interactions based on a gauge theory with group SU(2)×U(1) but, in order to reproduce Fermi's four-fermion theory of the β-decay it was necessary that two of the vector fields involved be massive. Also in condensed matter physics massive vector fields are required to describe certain systems, most notably in superconductivity.

The way out of this situation is found in the concept of spontaneous symmetry breaking discussed previously. The consistency of the quantum theory requires gauge invariance, but this invariance can be realized à la Nambu–Goldstone. When this is the case the full gauge symmetry is not explicitly present in the effective action constructed around the particular vacuum chosen by the theory. This makes possible the existence of mass terms for gauge fields without jeopardizing the consistency of the full theory, which is still invariant under the whole gauge group.

To illustrate the Higgs mechanism we study the simplest example, the Abelian Higgs model: a U(1) gauge field coupled to a self-interacting charged complex scalar field Φ with Lagrangian

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \overline{D_\mu \Phi} D^\mu \Phi - \frac{\lambda}{4}\left(\overline{\Phi}\Phi - \mu^2\right)^2, \tag{6.43}$$

where the covariant derivative is given by Eq. (4.69). This theory is invariant under the gauge transformations

$$\Phi \to e^{i\alpha(x)}\Phi, \qquad A_\mu \to A_\mu + \partial_\mu \alpha(x). \tag{6.44}$$

The minimum of the potential is defined by the equation $|\Phi| = \mu$. We have a continuum of different vacua labelled by the phase of the scalar field. None of these vacua, however, is invariant under the gauge symmetry

$$\langle \Phi \rangle = \mu e^{i\vartheta_0} \to \mu e^{i\vartheta_0 + i\alpha(x)} \tag{6.45}$$

and therefore the symmetry is spontaneously broken. Let us study now the theory around one of these vacua, for example $\langle \Phi \rangle = \mu$, by writing the field Φ in terms of the excitations around this particular vacuum

$$\Phi(x) = \left[\mu + \frac{1}{\sqrt{2}}\sigma(x)\right] e^{i\vartheta(x)}. \tag{6.46}$$

Independently of whether we are expanding around a particular vacuum for the scalar field we should keep in mind that the whole Lagrangian is still gauge invariant under (6.44). This means that performing a gauge transformation with parameter $\alpha(x) = -\vartheta(x)$ we can get rid of the phase in Eq. (6.46). Substituting then $\Phi(x) = \mu + \frac{1}{\sqrt{2}}\sigma(x)$ in the Lagrangian we find

$$\begin{aligned} \mathcal{L} &= -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + e^2\mu^2 A_\mu A^\mu + \frac{1}{2}\partial_\mu\sigma\partial^\mu\sigma - \frac{1}{2}\lambda\mu^2\sigma^2 \\ &- \lambda\mu\sigma^3 - \frac{\lambda}{4}\sigma^4 + e^2\mu A_\mu A^\mu \sigma + e^2 A_\mu A^\mu \sigma^2. \end{aligned} \tag{6.47}$$

What are the excitations of the theory around the vacuum $\langle \Phi \rangle = \mu$? First we find a massive real scalar field $\sigma(x)$. The important point, however, is that the vector field A_μ now has a mass given by

$$m_\gamma^2 = 2e^2\mu^2. \tag{6.48}$$

The remarkable thing about this way of giving a mass to the photon is that at no point have we given up gauge invariance. The symmetry is only hidden. Therefore in quantizing the theory we can still enjoy all the advantages of having a gauge theory but at the same time we have managed to generate a mass for the gauge field.

It is surprising, however, that in the Lagrangian (6.47) we did not find any massless mode. Since the vacuum chosen by the scalar field breaks the $U(1)$ generator of U(1) we would have expected one massless particle from Goldstone's theorem. To understand the fate of the missing Goldstone boson we have to revisit the calculation leading to Eq. (6.47). Were we dealing with a global U(1) theory, the Goldstone boson would correspond to excitation of the scalar field along the valley of the potential and the phase $\vartheta(x)$ would be the massless Goldstone boson. However, we have to keep in mind that in computing the Lagrangian we managed to get rid of $\vartheta(x)$ by shifting it into A_μ using a gauge transformation. Actually, by identifying the gauge parameter with the Goldstone excitation we have completely fixed the gauge, and the Lagrangian (6.47) does not have any gauge symmetry left.

A massive vector field has three polarizations: two transverse ones $\vec{k}\cdot\vec{\epsilon}(\vec{k},\pm 1)=0$ plus a longitudinal one $\vec{\epsilon}_L(\vec{k})\sim\vec{k}$. In gauging away the massless Goldstone boson $\vartheta(x)$ we have transformed it into the longitudinal polarization of the massive vector field. In the literature this is usually expressed saying that the Goldstone mode is 'eaten up' by the longitudinal component of the gauge field. It is important to realize that in spite of the fact that the Lagrangian (6.47) looks pretty different from the one we started with, we have not lost any degrees of freedom. We started with the two polarizations of the photon plus the two degrees of freedom associated with the real and imaginary components of the complex scalar field. After symmetry breaking we end up with the three polarizations of the massive vector field and the degree of freedom of the real scalar field $\sigma(x)$.

We can also understand the Higgs mechanism in the light of our discussion of gauge symmetry in Section 4.4. In the Higgs mechanism the invariance of the theory under infinitesimal gauge transformations is not explicitly broken, and this implies that Gauss's law is satisfied quantum mechanically, $\vec{\nabla}\cdot\vec{E}_a|\text{phys}\rangle=0$. The theory remains invariant under gauge transformations in the connected component of the identity $\mathcal{G}_0$, the ones generated by Gauss's law. This does not pose any restriction on the possible breaking of the invariance of the theory with respect to transformations that cannot be continuously deformed to the identity. Hence in the Higgs mechanism the invariance under gauge transformation that is not in the connected component of the identity, $\mathcal{G}/\mathcal{G}_0$, can be broken. Let us try to put it in more precise terms. As we learned in Section 4.4, in the Hamiltonian formulation of the theory, finite energy gauge field configurations tend to a pure gauge at spatial infinity

$$\vec{A}_\mu(\vec{x})\longrightarrow\frac{1}{ig}g(\vec{x})^{-1}\vec{\nabla}g(\vec{x}),\qquad |\vec{x}|\to\infty\,. \tag{6.49}$$

The set transformations $g_0(\vec{x})\in\mathcal{G}_0$ that tend to the identity at infinity are the ones generated by Gauss's law. However, one can also consider in general gauge transformations $g(\vec{x})$ which, as $|\vec{x}|\to\infty$, approach any other element $g\in G$. The quotient $\mathcal{G}_\infty\equiv\mathcal{G}/\mathcal{G}_0$ gives a copy of the gauge group at infinity. There is no reason, however, why this group should not be broken, and in general it is if the gauge symmetry is spontaneously broken. Notice that this is not a threat to the consistency of the theory. Properties like the decoupling of unphysical states are guaranteed by the fact that Gauss's law is satisfied quantum mechanically and are not affected by the breaking of $\mathcal{G}_\infty$.

The Abelian Higgs model discussed here can be regarded as a toy model of the Higgs mechanism responsible for giving mass to the $W^\pm$ and Z^0 gauge bosons in the Standard Model. In condensed matter physics the symmetry breaking described by the nonrelativistic version of the Abelian Higgs model can be used to characterize the onset of a superconducting phase in the BCS theory, where the complex scalar field Φ is associated with the Cooper pairs. In this case the parameter μ^2 depends on the temperature. Above the critical temperature T_c, $\mu^2(T)>0$ and there is only a symmetric vacuum $\langle\Phi\rangle=0$. When, on the other hand, $T<T_c$ then $\mu^2(T)<0$ and symmetry breaking takes place. The onset of a non-zero

mass of the photon (6.48) below the critical temperature explains the Meissner effect: the magnetic fields cannot penetrate inside superconductors beyond a distance of the order $1/m_\gamma$.

7 Anomalies

So far we did not worry too much about how classical symmetries of a theory are carried over to the quantum theory. We have implicitly assumed that classical symmetries are preserved in the process of quantization, so they are also realized in the quantum theory.

This, however, does not necessarily have to be the case. Quantizing an interacting field theory is a very involved process that requires regularization and renormalization and sometimes, it does not matter how hard we try, there is no way for a classical symmetry to survive quantization. When this happens one says that the theory has an *anomaly* (for a review see Ref. [26]). It is important to avoid here the misconception that anomalies appear due to a bad choice of the way a theory is regularized in the process of quantization. When we talk about anomalies we mean a classical symmetry that *cannot* be realized in the quantum theory, no matter how smart we are in choosing the regularization procedure.

In the following we analyse some examples of anomalies associated with global and local symmetries of the classical theory. In Section 8 we will encounter yet another example of an anomaly, this time associated with the breaking of classical scale invariance in the quantum theory.

7.1 Axial anomaly

Probably the best known examples of anomalies appear when we consider axial symmetries. If we consider a theory of two Weyl spinors $u_\pm$

$$\mathcal{L} = i\overline{\psi}\partial\!\!\!/\psi = iu_+^\dagger\sigma_+^\mu\partial_\mu u_+ + iu_-^\dagger\sigma_-^\mu\partial_\mu u_- \quad \text{with} \qquad \psi = \begin{pmatrix} u_+ \\ u_- \end{pmatrix} \tag{7.1}$$

the Lagrangian is invariant under two types of global U(1) transformations. In the first one both helicities transform with the same phase, this is a *vector* transformation:

$$\text{U(1)}_V : u_\pm \longrightarrow e^{i\alpha}u_\pm, \tag{7.2}$$

whereas in the second one, the axial $U(1)$, the signs of the phases are different for the two chiralities

$$\text{U(1)}_A : u_\pm \longrightarrow e^{\pm i\alpha}u_\pm. \tag{7.3}$$

Using Noether's theorem, there are two conserved currents, a vector current

$$J_V^\mu = \overline{\psi}\gamma^\mu\psi = u_+^\dagger\sigma_+^\mu u_+ + u_-^\dagger\sigma_-^\mu u_- \quad \Longrightarrow \quad \partial_\mu J_V^\mu = 0 \tag{7.4}$$

and an axial vector current

$$J_A^\mu = \overline{\psi}\gamma^\mu\gamma_5\psi = u_+^\dagger\sigma_+^\mu u_+ - u_-^\dagger\sigma_-^\mu u_- \quad \Longrightarrow \quad \partial_\mu J_A^\mu = 0. \tag{7.5}$$

The theory described by the Lagrangian (7.1) can be coupled to the electromagnetic field. The resulting classical theory is still invariant under the vector and axial U(1) symmetries (7.2) and (7.3). Surprisingly, upon quantization it turns out that the conservation of the axial current (7.5) is spoiled by quantum effects

$$\partial_\mu J_A^\mu \sim \hbar\,\vec{E}\cdot\vec{B}. \tag{7.6}$$

To understand more clearly how this result comes about we study first a simple model in two dimensions that captures the relevant physics involved in the four-dimensional case [27]. We work in

Minkowski space in two dimensions with coordinates $(x^0, x^1) \equiv (t, x)$ and where the spatial direction is compactified to a circle S^1. In this setup we consider a fermion coupled to the electromagnetic field. Notice that since we are living in two dimensions the field strength $F_{\mu\nu}$ has only one independent component that corresponds to the electric field along the spatial direction, $F^{01} \equiv \mathcal{E}$ (in two dimensions there are no magnetic fields!).

To write the Lagrangian for the spinor field we need to find a representation of the algebra of γ-matrices

$$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu} \quad \text{with} \quad \eta = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{7.7}$$

In two dimensions the dimension of the representation of the γ-matrices is $2^{[\frac{2}{2}]} = 2$. Here take

$$\gamma^0 \equiv \sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \gamma^1 \equiv i\sigma^2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{7.8}$$

This is a chiral representation since the matrix γ_5 is diagonal[15]

$$\gamma_5 \equiv -\gamma^0\gamma^1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{7.9}$$

Writing the two-component spinor ψ as

$$\psi = \begin{pmatrix} u_+ \\ u_- \end{pmatrix} \tag{7.10}$$

and defining as usual the projectors $P_\pm = \frac{1}{2}(\mathbf{1} \pm \gamma_5)$ we find that the components $u_\pm$ of ψ are respectively a right- and left-handed Weyl spinor in two dimensions.

Once we have a representation of the γ-matrices we can write the Dirac equation. Expressing it in terms of the components $u_\pm$ of the Dirac spinor we find

$$(\partial_0 - \partial_1)u_+ = 0, \qquad (\partial_0 + \partial_1)u_- = 0. \tag{7.11}$$

The general solution to these equations can immediately be written as

$$u_+ = u_+(x^0 + x^1), \qquad u_- = u_-(x^0 - x^1). \tag{7.12}$$

Hence $u_\pm$ are two wave packets moving along the spatial dimension respectively to the left (u_+) and to the right (u_-). Notice that according to our convention the left-moving u_+ is a right-handed spinor (positive helicity) whereas the right-moving u_- is a left-handed spinor (negative helicity).

If we want to interpret (7.11) as the wave equation for two-dimensional Weyl spinors we have the following wave functions for free particles with well-defined momentum $p^\mu = (E, p)$.

$$u_\pm^{(E)}(x^0 \pm x^1) = \frac{1}{\sqrt{L}} e^{-iE(x^0 \pm x^1)} \quad \text{with} \quad p = \mp E. \tag{7.13}$$

As is always the case with the Dirac equation we have both positive and negative energy solutions. For u_+, since $E = -p$, we see that the solutions with positive energy are those with negative momentum $p < 0$, whereas the negative energy solutions are plane waves with $p > 0$. For the left-handed spinor u_- the situation is reversed. Besides, since the spatial direction is compact with length L the momentum p is quantized according to

$$p = \frac{2\pi n}{L}, \qquad n \in \mathbb{Z}. \tag{7.14}$$

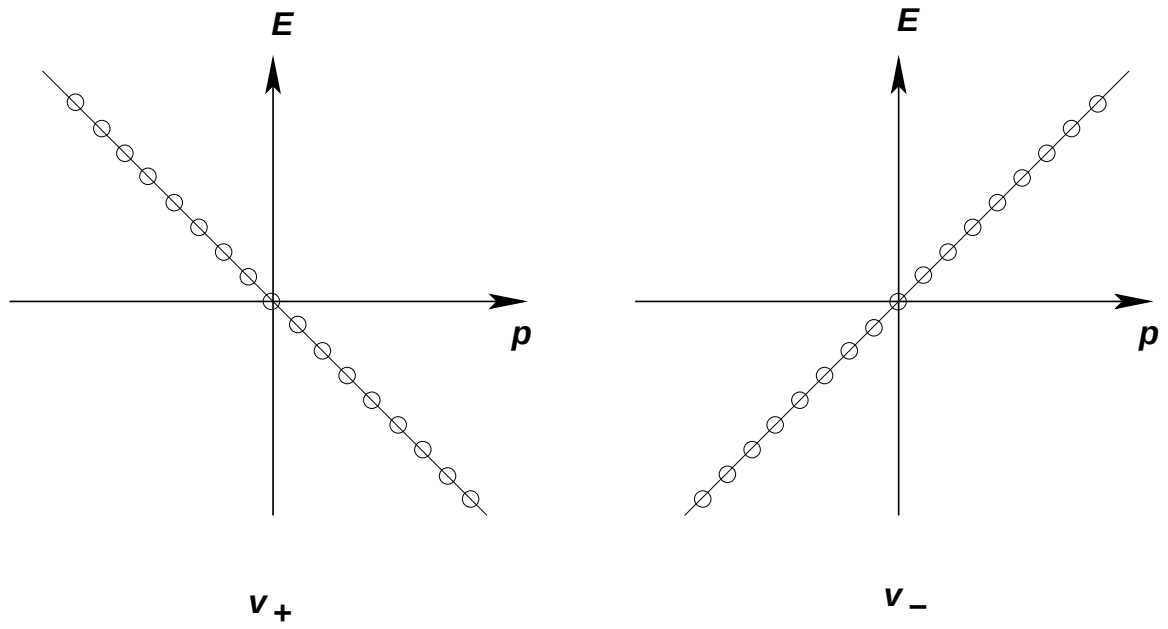


Fig. 11: Spectrum of the massless two-dimensional Dirac field

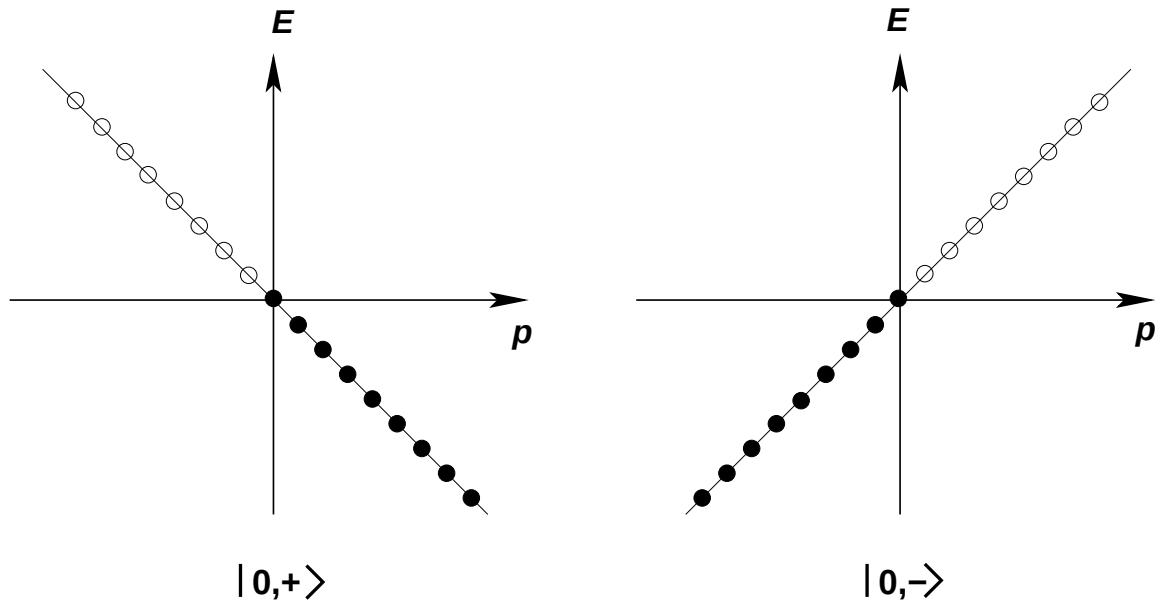


Fig. 12: Vacuum of the theory

The spectrum of the theory is represented in Fig. 11.

Once we have the spectrum of the theory the next step is to obtain the vacuum. As with the Dirac equation in four dimensions we fill all the states with $E \leq 0$ (Fig. 12). Exciting of a particle in the Dirac sea produces a positive energy fermion plus a hole that is interpreted as an antiparticle. This gives us the clue on how to quantize the theory. In the expansion of the operator $u_{\pm}$ in terms of the modes (7.13) we associate positive energy states with annihilation operators whereas the states with negative energy are

[15] In any even number of dimensions γ_5 is defined to satisfy the conditions $\gamma_5^2 = \mathbf{1}$ and $\{\gamma_5, \gamma^\mu\} = 0$.

associated with creation operators for the corresponding antiparticle

$$u_\pm(x) = \sum_{E>0} \left[a_\pm(E) v_\pm^{(E)}(x) + b_\pm^\dagger(E) v_\pm^{(E)}(x)^* \right]. \tag{7.15}$$

The operator $a_\pm(E)$ acting on the vacuum $|0,\pm\rangle$ annihilates a particle with positive energy E and momentum $\mp E$. In the same way $b_\pm^\dagger(E)$ creates out of the vacuum an antiparticle with positive energy E and spatial momentum $\mp E$. In the Dirac sea picture the operator $b_\pm(E)^\dagger$ is originally an annihilation operator for a state of the sea with negative energy $-E$. As in the four-dimensional case the problem of the negative energy states is solved by interpreting annihilation operators for negative energy states as creation operators for the corresponding antiparticle with positive energy (and vice versa). The operators appearing in the expansion of $u_\pm$ in Eq. (7.15) satisfy the usual algebra

$$\{a_\lambda(E), a_{\lambda'}^\dagger(E')\} = \{b_\lambda(E), b_{\lambda'}^\dagger(E')\} = \delta_{E,E'}\delta_{\lambda\lambda'}, \tag{7.16}$$

where we have introduced the label $\lambda, \lambda' = \pm$. Also, $a_\lambda(E)$, $a_\lambda^\dagger(E)$ anticommute with $b_{\lambda'}(E')$, $b_{\lambda'}^\dagger(E')$.

The Lagrangian of the theory

$$\mathcal{L} = i u_+^\dagger(\partial_0 + \partial_1) u_+ + i u_-^\dagger(\partial_0 - \partial_1) u_- \tag{7.17}$$

is invariant under both U(1)$_V$, Eq. (7.2), and U(1)$_A$, Eq. (7.3). The associated Noether currents are in this case

$$J_V^\mu = \begin{pmatrix} u_+^\dagger u_+ + u_-^\dagger u_- \\ -u_+^\dagger u_+ + u_-^\dagger u_- \end{pmatrix}, \qquad J_A^\mu = \begin{pmatrix} u_+^\dagger u_+ - u_-^\dagger u_- \\ -u_+^\dagger u_+ - u_-^\dagger u_- \end{pmatrix}. \tag{7.18}$$

The associated conserved charges are given, for the vector current by

$$Q_V = \int_0^L dx^1 \left(u_+^\dagger u_+ + u_-^\dagger u_- \right) \tag{7.19}$$

and for the axial current

$$Q_A = \int_0^L dx^1 \left(u_+^\dagger u_+ - u_-^\dagger u_- \right). \tag{7.20}$$

Using the orthonormality relations for the modes $v_\pm^{(E)}(x)$

$$\int_0^L dx^1 \, v_\pm^{(E)}(x) \, v_\pm^{(E')}(x) = \delta_{E,E'} \tag{7.21}$$

we find for the conserved charges:

$$\begin{aligned} Q_V &= \sum_{E>0} \left[a_+^\dagger(E) a_+(E) - b_+^\dagger(E) b_+(E) + a_-^\dagger(E) a_-(E) - b_-^\dagger(E) b_-(E) \right], \\ Q_A &= \sum_{E>0} \left[a_+^\dagger(E) a_+(E) - b_+^\dagger(E) b_+(E) - a_-^\dagger(E) a_-(E) + b_-^\dagger(E) b_-(E) \right]. \end{aligned} \tag{7.22}$$

We see that Q_V counts the net number (particles minus antiparticles) of positive helicity states plus the net number of states with negative helicity. The axial charge, on the other hand, counts the net number of positive helicity states minus the number of negative helicity ones. In the case of the vector current we have subtracted a formally divergent vacuum contribution to the charge (the 'charge of the Dirac sea').

In the free theory there is of course no problem with the conservation of either Q_V or Q_A, since the occupation numbers do not change. What we want to study is the effect of coupling the theory to electric

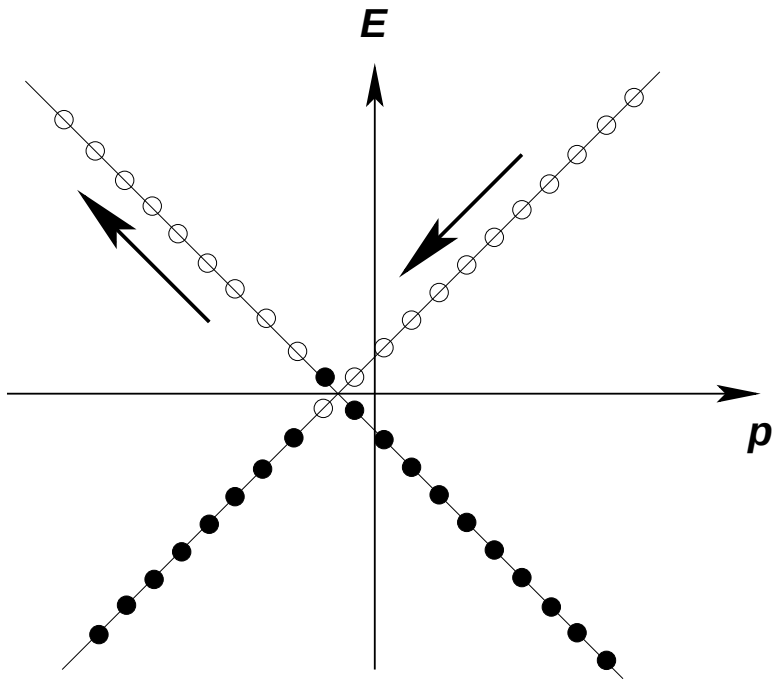


Fig. 13: Effect of the electric field

field $\mathcal{E}$. We work in the gauge $A_0 = 0$. Instead of solving the problem exactly we are going to simulate the electric field by adiabatically varying in a long time τ_0 the vector potential A_1 from zero value to $-\mathcal{E}\tau_0$. From our discussion in section 4.3 we know that the effect of the electromagnetic coupling in the theory is a shift in the momentum according to

$$p \longrightarrow p - eA_1, \tag{7.23}$$

where e is the charge of the fermions. Since we assumed that the vector potential varies adiabatically, we can assume it to be approximately constant at each time.

Then, we have to understand what is the effect of (7.23) on the vacuum depicted in Fig. 12. What we find is that the two branches move as shown in Fig. 13 resulting in some of the negative energy states of the v_+ branch acquiring positive energy while the same number of the empty positive energy states of the other branch v_- will become empty negative energy states. Physically this means that the external electric field $\mathcal{E}$ creates a number of particle–antiparticle pairs out of the vacuum. Denoting by $N \sim e\mathcal{E}$ the number of such pairs created by the electric field per unit time, the final values of the charges Q_V and Q_A are

$$\begin{aligned} Q_A(\tau_0) &= (N-0) + (0-N) = 0, \\ Q_V(\tau_0) &= (N-0) - (0-N) = 2N. \end{aligned} \tag{7.24}$$

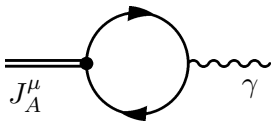
Therefore we conclude that the coupling to the electric field produces a violation in the conservation of the axial charge per unit time given by $\Delta Q_A \sim e\mathcal{E}$. This implies that

$$\partial_\mu J^\mu_A \sim e\hbar\mathcal{E}, \tag{7.25}$$

where we have restored $\hbar$ to make clear that the violation in the conservation of the axial current is a quantum effect. At the same time $\Delta Q_V = 0$ guarantees that the vector current remains conserved also quantum mechanically, $\partial_\mu J^\mu_V = 0$.

We have just studied a two-dimensional example of the Adler–Bell–Jackiw axial anomaly [28]. The heuristic analysis presented here can be made more precise by computing the quantity

$$C^{\mu\nu} = \langle 0|T\left[J_A^\mu(x) J_V^\nu(0)\right]|0\rangle = \quad .$$



(7.26)

The anomaly is given then by $\partial_\mu C^{\mu\nu}$. A careful calculation yields the numerical prefactor missing in Eq. (7.25) leading to the result

$$\partial_\mu J_A^\mu = \frac{e\hbar}{2\pi}\varepsilon^{\nu\sigma}F_{\nu\sigma}, \tag{7.27}$$

with $\varepsilon^{01} = -\varepsilon^{10} = 1$.

The existence of an anomaly in the axial symmetry that we have illustrated in two dimensions is present in all even dimensional space-times. In particular in four dimensions the axial anomaly is given by

$$\partial_\mu J_A^\mu = -\frac{e^2}{16\pi^2}\varepsilon^{\mu\nu\sigma\lambda}F_{\mu\nu}F_{\sigma\lambda}. \tag{7.28}$$

This result has very important consequences in the physics of strong interactions as we will see in what follows

7.2 Chiral symmetry in QCD

Our knowledge of the physics of strong interactions is based on the theory of Quantum Chromodynamics (QCD) [29]. This is a non-Abelian gauge theory with gauge group SU(N_c) coupled to a number N_f of quarks. These are spin-$\frac{1}{2}$ particles $Q^{i\,f}$ labelled by two quantum numbers: color $i = 1, \ldots, N_c$ and flavor $f = 1, \ldots, N_f$. The interaction between them is mediated by the $N_c^2 - 1$ gauge bosons, the gluons A_μ^a, $a = 1, \ldots, N_c^2 - 1$. In the real world $N_c = 3$ and the number of flavors is six, corresponding to the number of different quarks: up (u), down (d), charm (c), strange (s), top (t) and bottom (b).

For the time being we are going to study a general theory of QCD with N_c colors and N_f flavors. Also, for reasons that will be clear later we are going to work in the limit of vanishing quark masses, $m_f \to 0$. In this cases the Lagrangian is given by

$$\mathcal{L}_{\rm QCD} = -\frac{1}{4}F^a_{\mu\nu}F^{a\,\mu\nu} + \sum_{f=1}^{N_f}\left[i\overline{Q}^f_L \not{D}\, Q^f_L + i\overline{Q}^f_R \not{D}\, Q^f_R\right], \tag{7.29}$$

where the subscripts L and R indicate respectively left- and right-handed spinors, $Q^f_{L,R} \equiv P_\pm Q^f$, and the field strength $F^a_{\mu\nu}$ and the covariant derivative D_μ are respectively defined in Eqs. (4.75) and (4.78). Apart from the gauge symmetry, this Lagrangian is also invariant under a global U(N_f)$_L\times$U(N_f)$_R$ acting on the flavor indices and defined by

$$\mathrm{U}(N_f)_L : \begin{cases} Q^f_L \;\to\; \sum_{f'}(U_L)_{ff'}Q^{f'}_L \\ Q^f_R \;\to\; Q^f_R \end{cases} \qquad \mathrm{U}(N_f)_R : \begin{cases} Q^f_L \;\to\; Q^f_L \\ Q^r_R \;\to\; \sum_{f'}(U_R)_{ff'}Q^{f'}_R \end{cases} \tag{7.30}$$

with $U_L, U_R \in \mathrm{U}(N_f)$. Actually, since U($N$)=U(1)×SU($N$) this global symmetry group can be written as $\mathrm{SU}(N_f)_L \times \mathrm{SU}(N_f)_R \times \mathrm{U}(1)_L \times \mathrm{U}(1)_R$. The Abelian subgroup $\mathrm{U}(1)_L \times \mathrm{U}(1)_R$ can now be decomposed into their vector $\mathrm{U}(1)_B$ and axial $\mathrm{U}(1)_A$ subgroups defined by the transformations

$$\mathrm{U}(1)_B : \begin{cases} Q_L^f & \to & e^{i\alpha} Q_L^f \\ \\ Q_R^f & \to & e^{i\alpha} Q_R^f \end{cases} \qquad \mathrm{U}(1)_A : \begin{cases} Q_L^f & \to & e^{i\alpha} Q_L^f \\ \\ Q_R^f & \to & e^{-i\alpha} Q_R^f \end{cases} \tag{7.31}$$

According to Noether's theorem, associated with these two Abelian symmetries we have two conserved currents:

$$J_V^\mu = \sum_{f=1}^{N_f} \overline{Q}^f \gamma^\mu \, Q^f, \qquad J_A^\mu = \sum_{f=1}^{N_f} \overline{Q}^f \gamma^\mu \gamma_5 \, Q^f. \tag{7.32}$$

The conserved charge associated with vector charge J_V^μ is actually the baryon number defined as the number of quarks minus number of antiquarks.

The non-Abelian part of the global symmetry group $\mathrm{SU}(N_f)_L \times \mathrm{SU}(N_f)_R$ can also be decomposed into its vector and axial subgroups, $\mathrm{SU}(N_f)_V \times \mathrm{SU}(N_f)_A$, defined by the following transformations of the quarks fields

$$\mathrm{SU}(N_f)_V : \begin{cases} Q_L^f & \to & \sum_{f'} (U_L)_{ff'} Q_L^{f'} \\ \\ Q_R^f & \to & \sum_{f'} (U_L)_{ff'} Q_R^{f'} \end{cases} \qquad \mathrm{SU}(N_f)_A : \begin{cases} Q_L^f & \to & \sum_{f'} (U_L)_{ff'} Q_L^{f'} \\ \\ Q_R^f & \to & \sum_{f'} (U_R^{-1})_{ff'} Q_R^{f'} \end{cases} \tag{7.33}$$

Again, the application of Noether's theorem shows the existence of the following non-Abelian conserved charges

$$J_V^{I\,\mu} \equiv \sum_{f,f'=1}^{N_f} \overline{Q}^f \gamma^\mu (T^I)_{ff'} Q^{f'}, \qquad J_A^{I\,\mu} \equiv \sum_{f,f'=1}^{N_f} \overline{Q}^f \gamma^\mu \gamma_5 (T^I)_{ff'} Q^{f'}. \tag{7.34}$$

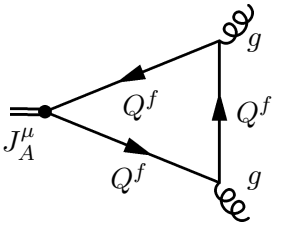
To summarize, we have shown that the initial chiral symmetry of the QCD Lagrangian (7.29) can be decomposed into its chiral and vector subgroups according to

$$\mathrm{U}(N_f)_L \times \mathrm{U}(N_f)_R = \mathrm{SU}(N_f)_V \times \mathrm{SU}(N_f)_A \times \mathrm{U}(1)_B \times \mathrm{U}(1)_A. \tag{7.35}$$

The question to address now is which part of the classical global symmetry is preserved by the quantum theory.

As argued in Section 7.1, the conservation of the axial currents J_A^μ and $J_A^{a\,\mu}$ can in principle be spoiled due to the presence of an anomaly. In the case of the Abelian axial current J_A^μ the relevant quantity is the correlation function

$$C^{\mu\nu\sigma} \equiv \langle 0 | T \left[J_A^\mu(x) j_{\mathrm{gauge}}^{a\,\nu}(x') j_{\mathrm{gauge}}^{b\,\sigma}(0) \right] | 0 \rangle = \sum_{f=1}^{N_f} \Bigg[\quad \Bigg]_{\mathrm{symmetric}} \tag{7.36}$$



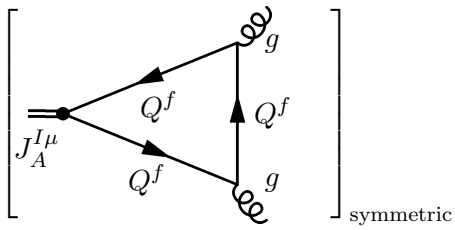
Here $j_{\mathrm{gauge}}^{a\,\mu}$ is the non-Abelian conserved current coupling to the gluon field

$$j_{\mathrm{gauge}}^{a\,\mu} \equiv \sum_{f=1}^{N_f} \overline{Q}^f \gamma^\mu \tau^a Q^f, \tag{7.37}$$

where, to avoid confusion with the generators of the global symmetry, we have denoted by τ^a the generators of the gauge group SU(N_c). The anomaly can be read now from $\partial_\mu C^{\mu\nu\sigma}$. If we impose Bose symmetry with respect to the interchange of the two outgoing gluons and gauge invariance of the whole expression, $\partial_\nu C^{\mu\nu\sigma} = 0 = \partial_\sigma C^{\mu\nu\sigma}$, we find that the axial Abelian global current has an anomaly given by[16]

$$\partial_\mu J_A^\mu = -\frac{g^2 N_f}{32\pi^2}\varepsilon^{\mu\nu\sigma\lambda} F^a_{\mu\nu} F^{a\,\mu\nu}. \tag{7.38}$$

In the case of the non-Abelian axial global symmetry SU(N_f)$_A$ the calculation of the anomaly is made as above. The result, however, is quite different since in this case we conclude that the non-Abelian axial current $J_A^{a\,\mu}$ is not anomalous. This can easily be seen by noticing that associated with the axial current vertex we have a generator T^I of SU(N_f), whereas for the two gluon vertices we have the generators τ^a of the gauge group SU(N_c). Therefore, the triangle diagram is proportional to the group-theoretic factor

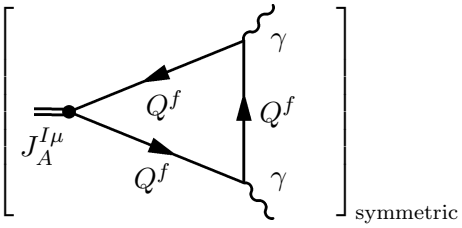


$$\sim \mathrm{tr}\, T^I \ \mathrm{tr}\,\{\tau^a, \tau^b\} = 0 \tag{7.39}$$

which vanishes because the generators of SU(N_f) are traceless.

From here we would conclude that the non-Abelian axial symmetry SU(N_f)$_A$ is nonanomalous. However, this is not the whole story since quarks are charged particles that also couple to photons. Hence there is a second potential source of an anomaly coming from the the one-loop triangle diagram coupling $J_A^{I\,\mu}$ to two photons

$$\langle 0|T\left[J_A^{I\,\mu}(x) j^\nu_{\rm em}(x') j^\sigma_{\rm em}(0)\right]|0\rangle = \sum_{f=1}^{N_f}$$



(7.40)

where $j^\mu_{\rm em}$ is the electromagnetic current

$$j^\mu_{\rm em} = \sum_{f=1}^{N_f} q_f\, \overline{Q}^f \gamma^\mu Q^f, \tag{7.41}$$

with q_f the electric charge of the f-th quark flavor. A calculation of the diagram in (7.40) shows the existence of an Adler–Bell–Jackiw anomaly given by

$$\partial_\mu J_A^{I\,\mu} = -\frac{N_c}{16\pi^2}\left[\sum_{f=1}^{N_f} (T^I)_{ff}\, q_f^2\right] \varepsilon^{\mu\nu\sigma\lambda} F_{\mu\nu} F_{\sigma\lambda}, \tag{7.42}$$

where $F_{\mu\nu}$ is the field strength of the electromagnetic field coupling to the quarks. The only chance for the anomaly to cancel is that the factor between brackets in this equation be identically zero.

[16]The normalization of the generators T^I of the global SU(N_f) is given by $\mathrm{tr}\,(T^I T^J) = \frac{1}{2}\delta^{IJ}$.

Before proceeding let us summarize the results found so far. Because of the presence of anomalies the axial part of the global chiral symmetry $\mathrm{SU}(N_f)_A$ and $\mathrm{U}(1)_A$ are not realized quantum mechanically in general. We found that $\mathrm{U}(1)_A$ is always affected by an anomaly. However, because the right-hand side of the anomaly equation (7.38) is a total derivative, the anomalous character of J_A^μ does not explain the absence of $\mathrm{U}(1)_A$ multiplets in the hadron spectrum, since a new current can be constructed which is conserved. In addition, the nonexistence of candidates for a Goldstone boson associated with the right quantum numbers indicates that $\mathrm{U}(1)_A$ is not spontaneously broken either, so it has be explicitly broken somehow. This is the so-called U(1)-problem which was solved by 't Hooft [30] who showed how the contribution of quantum transitions between vacua with topologically nontrivial gauge field configurations (instantons) results in an explicit breaking of this symmetry.

Owing to the dynamics of the $\mathrm{SU}(N_c)$ gauge theory the axial non-Abelian symmetry is spontaneously broken due to the presence at low energies of a vacuum expectation value for the fermion bilinear $\overline{Q}^f Q^f$

$$\langle 0|\overline{Q}^f Q^f|0\rangle \neq 0 \qquad \text{(No summation in } f\text{!)}. \tag{7.43}$$

This nonvanishing vacuum expectation value for the quark bilinear actually breaks chiral invariance spontaneously to the vector subgroup $\mathrm{SU}(N_f)_V$, so the only subgroup of the original global symmetry that is realized by the full theory at low energy is

$$\mathrm{U}(N_f)_L \times \mathrm{U}(N_f)_R \longrightarrow \mathrm{SU}(N_f)_V \times \mathrm{U}(1)_B. \tag{7.44}$$

Associated with this breaking, a Goldstone boson should appear with the quantum numbers of the broken non-Abelian current. For example, in the case of QCD the Goldstone bosons associated with the spontaneously symmetry breaking induced by the vacuum expectation values $\langle\overline{u}u\rangle$, $\langle\overline{d}d\rangle$ and $\langle(\overline{u}d-\overline{d}u)\rangle$ have been identified as the pions π^0, $\pi^\pm$. These bosons are not exactly massless because of the nonvanishing mass of the u and d quarks. Since the global chiral symmetry is already slightly broken by mass terms in the Lagrangian, the associated Goldstone bosons also have masses although they are very light compared to the masses of other hadrons.

In order to have a better physical understanding of the role of anomalies in the physics of strong interactions we particularize now our analysis of the case of real QCD. Since the u and d quarks are much lighter than the other four flavors, QCD at low energies can be well described by including only these two flavors and ignoring heavier quarks. In this approximation, from our previous discussion we know that the low-energy global symmetry of the theory is $\mathrm{SU}(2)_V \times \mathrm{U}(1)_B$, where now the vector group $\mathrm{SU}(2)_V$ is the well-known isospin symmetry. The axial $\mathrm{U}(1)_A$ current is anomalous due to Eq. (7.38) with $N_f = 2$. In the case of the non-Abelian axial symmetry $\mathrm{SU}(2)_A$, taking into account that $q_u = \frac{2}{3}e$ and $q_d = -\frac{1}{3}e$ and that the three generators of SU(2) can be written in terms of the Pauli matrices as $T^K = \frac{1}{2}\sigma^K$ we find

$$\sum_{f=u,d} (T^1)_{ff}\, q_f^2 = \sum_{f=u,d} (T^1)_{ff}\, q_f^2 = 0, \qquad \sum_{f=u,d} (T^3)_{ff}\, q_f^2 = \frac{e^2}{6}. \tag{7.45}$$

Therefore $J_A^{3\,\mu}$ is anomalous.

Physically, the anomaly in the axial current $J_A^{3\,\mu}$ has an important consequence. In the quark model, the wave function of the neutral pion π^0 is given in terms of those for the u and d quark by

$$|\pi^0\rangle = \frac{1}{\sqrt{2}}\left(|\bar{u}\rangle|u\rangle - |\bar{d}\rangle|d\rangle\right). \tag{7.46}$$

The isospin quantum numbers of $|\pi^0\rangle$ are those of the generator T^3. Actually the analogy goes further since $\partial_\mu J_A^{3\,\mu}$ is the operator creating a pion π^0 out of the vacuum

$$|\pi^0\rangle \sim \partial_\mu J_A^{3\,\mu}|0\rangle. \tag{7.47}$$

This leads to the physical interpretation of the triangle diagram (7.40) with $J_A^{3\,\mu}$ as the one-loop contribution to the decay of a neutral pion into two photons

$$\pi^0 \longrightarrow 2\gamma\,. \tag{7.48}$$

This is an interesting piece of physics. In 1967 Sutherland and Veltman [31] presented a calculation, using current algebra techniques, according to which the decay of the pion into two photons should be suppressed. This however contradicted the experimental evidence that showed the existence of such a decay. The way out of this paradox, as pointed out in [28], is the axial anomaly. What happens is that the current algebra analysis overlooks the ambiguities associated with the regularization of divergences in quantum field theory. A QED evaluation of the triangle diagram leads to a divergent integral that has to be regularized somehow. It is in this process that the Adler–Bell–Jackiw axial anomaly appears resulting in a nonvanishing value for the $\pi^0 \to 2\gamma$ amplitude[17].

The existence of anomalies associated with global currents does not necessarily mean difficulties for the theory. On the contrary, as we saw in the case of the axial anomaly it is its existence that allows for a solution of the Sutherland–Veltman paradox and an explanation of the electromagnetic decay of the pion. The situation, however, is very different if we deal with local symmetries. A quantum mechanical violation of gauge symmetry leads to all kinds of problems, from lack of renormalizability to nondecoupling of negative norm states. This is because the presence of an anomaly in the theory implies that the Gauss law constraint $\vec{\nabla}\cdot\vec{E}_a = \rho_a$ cannot be consistently implemented in the quantum theory. As a consequence states that classically are eliminated by the gauge symmetry become propagating fields in the quantum theory, thus spoiling the consistency of the theory.

Anomalies in a gauge symmetry can be expected only in chiral theories where left- and right-handed fermions transform in different representations of the gauge group. Physically, the most interesting example of such theories is the electroweak sector of the Standard Model where, for example, left-handed fermions transform as doublets under SU(2) whereas right-handed fermions are singlets. On the other hand, QCD is free of gauge anomalies since both left- and right-handed quarks transform in the fundamental representation of SU(3).

We consider the Lagrangian

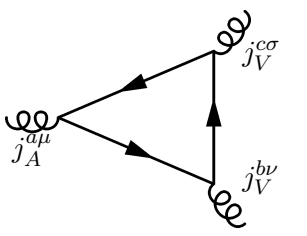
$$\mathcal{L} = -\frac{1}{4}F^{a\,\mu\nu}F^a_{\mu\nu} + i\sum_{i=1}^{N_+}\overline{\psi}^i_+ \not{D}^{\,(+)}\psi^i_+ + i\sum_{j=1}^{N_-}\overline{\psi}^j_- \not{D}^{\,(-)}\psi^j_-, \tag{7.49}$$

where the chiral fermions $\psi^i_\pm$ transform according to the representations $\tau^a_{i,\pm}$ of the gauge group G ($a = 1,\ldots,\dim G$). The covariant derivatives $D^{(\pm)}_\mu$ are then defined by

$$D^{(\pm)}_\mu\psi^i_\pm = \partial_\mu\psi^i_\pm + igA^K_\mu\tau^K_{i,\pm}\psi^i_\pm. \tag{7.50}$$

As for global symmetries, anomalies in the gauge symmetry appear in the triangle diagram with one axial and two vector gauge current vertices

$$\langle 0|T\left[j_A^{a\,\mu}(x)j_V^{b\,\nu}(x')j_V^{c\,\sigma}(0)\right]|0\rangle = \left[\quad\right]_{\text{symmetric}} \tag{7.51}$$



[17] An early computation of the triangle diagram for the electromagnetic decay of the pion was made by Steinberger [32].

where gauge vector and axial currents $j_V^{a\,\mu}$, $j_A^{a\,\mu}$ are given by

$$\begin{aligned} j_V^{a\mu} &= \sum_{i=1}^{N_+} \overline{\psi}_+^i \tau_+^a \gamma^\mu \psi_+^i + \sum_{j=1}^{N_-} \overline{\psi}_-^j \tau_-^a \gamma^\mu \psi_-^j, \\ j_A^{a\mu} &= \sum_{i=1}^{N_+} \overline{\psi}_+^i \tau_+^a \gamma^\mu \psi_+^i - \sum_{i=1}^{N_-} \overline{\psi}_-^j \tau_-^a \gamma^\mu \psi_-^j. \end{aligned} \tag{7.52}$$

Luckily, we do not have to compute the whole diagram in order to find an anomaly cancellation condition, it is enough that we calculate the overall group theoretical factor. In the case of the diagram in Eq. (7.51) for every fermion species running in the loop this factor is equal to

$$\mathrm{tr}\,\left[\tau_{i,\pm}^a \{\tau_{i,\pm}^b, \tau_{i,\pm}^c\}\right], \tag{7.53}$$

where the sign $\pm$ corresponds respectively to the generators of the representation of the gauge group for the left- and right-handed fermions. Hence the anomaly cancellation condition reads

$$\sum_{i=1}^{N_+} \mathrm{tr}\,\left[\tau_{i,+}^a \{\tau_{i,+}^b, \tau_{i,+}^c\}\right] - \sum_{j=1}^{N_-} \mathrm{tr}\,\left[\tau_{j,-}^a \{\tau_{j,-}^b, \tau_{j,-}^c\}\right] = 0. \tag{7.54}$$

Knowing this we can proceed to check the anomaly cancellation in the Standard Model SU(3)×SU(2)× Left-handed fermions (both leptons and quarks) transform as doublets with respect to the SU(2) factor whereas the right-handed components are singlets. The charge with respect to the U(1) part, the hypercharge Y, is determined by the Gell-Mann–Nishijima formula

$$Q = T_3 + Y, \tag{7.55}$$

where Q is the electric charge of the corresponding particle and T_3 is the eigenvalue with respect to the third generator of the SU(2) group in the corresponding representation: $T_3 = \frac{1}{2}\sigma^3$ for the doublets and $T_3 = 0$ for the singlets. For the first family of quarks (u, d) and leptons (e, ν_e) we have the following field content

$$\begin{array}{lccc} \text{quarks:} & \begin{pmatrix} u^\alpha \\ d^\alpha \end{pmatrix}_{L,\frac{1}{6}} & u^\alpha_{R,\frac{2}{3}} & d^\alpha_{R,\frac{2}{3}} \\ \text{leptons:} & \begin{pmatrix} \nu_e \\ e \end{pmatrix}_{L,-\frac{1}{2}} & e_{R,-1} & \end{array} \tag{7.56}$$

where $\alpha = 1, 2, 3$ labels the color quantum number and the subscript indicates the value of the weak hypercharge Y. Denoting the representations of SU(3)×SU(2)×U(1) by $(n_c, n_w)_Y$, with n_c and n_w the representations of SU(3) and SU(2) respectively and Y the hypercharge, the matter content of the Standard Model consists of a three-family replication of the representations:

$$\begin{array}{lcccc} \text{left-handed fermions:} & (3,2)^L_{\frac{1}{6}} & (1,2)^L_{-\frac{1}{2}} & & \\ \text{right-handed fermions:} & (3,1)^R_{\frac{2}{3}} & (3,1)^R_{-\frac{1}{3}} & (1,1)^R_{-1}. & \end{array} \tag{7.57}$$

In computing the triangle diagram we have 10 possibilities depending on which factor of the gauge group

SU(3)×SU(2)×U(1) couples to each vertex:

$\mathrm{SU}(3)^3$	$\mathrm{SU}(2)^3$	$\mathrm{U}(1)^3$
$\mathrm{SU}(3)^2\,\mathrm{SU}(2)$	$\mathrm{SU}(2)\,\mathrm{U}(1)$	
$\mathrm{SU}(3)^2\,\mathrm{U}(1)$	$\mathrm{SU}(2)\,\mathrm{U}(1)^2$	
$\mathrm{SU}(3)\,\mathrm{SU}(2)^2$		
$\mathrm{SU}(3)\,\mathrm{SU}(2)\,\mathrm{U}(1)$		
$\mathrm{SU}(3)\,\mathrm{U}(1)^2$		

It is easy to check that some of them do not give rise to anomalies. For example, the anomaly for the $\mathrm{SU}(3)^3$ case cancels because left- and right-handed quarks transform in the same representation. In the case of $\mathrm{SU}(2)^3$ the cancellation happens term by term because of the Pauli matrices identity $\sigma^a\sigma^b = \delta^{ab} + i\varepsilon^{abc}\sigma^c$ that leads to

$$\mathrm{tr}\,\left[\sigma^a\{\sigma^b,\sigma^c\}\right] = 2\,(\mathrm{tr}\,\sigma^a)\,\delta^{bc} = 0. \tag{7.58}$$

However, the hardest anomaly cancellation condition to satisfy is the one with three U(1)'s. In this case the absence of anomalies within a single family is guaranteed by the nontrivial identity

$$\begin{aligned}\sum_{\text{left}} Y_+^3 - \sum_{\text{right}} Y_-^3 &= 3\times 2\times\left(\frac{1}{6}\right)^3 + 2\times\left(-\frac{1}{2}\right)^3 - 3\times\left(\frac{2}{3}\right)^3 - 3\times\left(-\frac{1}{3}\right)^3 - (-1)^3 \\ &= \left(-\frac{3}{4}\right) + \left(\frac{3}{4}\right) = 0.\end{aligned} \tag{7.59}$$

It is remarkable that the anomaly exactly cancels between leptons and quarks. Notice that this result holds even if a right-handed sterile neutrino is added since such a particle is a singlet under the whole Standard Model gauge group and therefore does not contribute to the triangle diagram. Therefore we see how the matter content of the Standard Model conspires to yield a consistent quantum field theory.

In all our discussion of anomalies we considered the computation of one-loop diagrams only. It may happen that higher loop orders impose additional conditions. Fortunately this is not so: the Adler–Bardeen theorem [33] guarantees that the axial anomaly receives contributions only from one-loop diagrams. Therefore, once anomalies are cancelled (if possible) at one-loop we know that there will be no new conditions coming from higher-loop diagrams in perturbation theory.

The Adler–Bardeen theorem, however, only applies in perturbation theory. It is nonetheless possible that nonperturbative effects can result in the quantum violation of a gauge symmetry. This is precisely the case pointed out by Witten [34] with respect to the SU(2) gauge symmetry of the Standard Model. In this case the problem lies in the nontrivial topology of the gauge group SU(2). The invariance of the theory with respect to gauge transformations which are not in the connected component of the identity makes all correlation functions equal to zero. Only when the number of left-handed SU(2) fermion doublets is even does gauge invariance allow for a nontrivial theory. It is again remarkable that the family structure of the Standard Model makes this anomaly cancel

$$3\times\begin{pmatrix} u \\ d \end{pmatrix}_L + 1\times\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L = 4\ \text{SU(2)-doublets}, \tag{7.60}$$

where the factor of 3 comes from the number of colors.

8 Renormalization

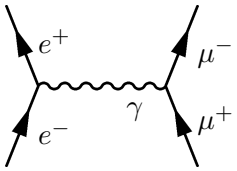
8.1 Removing infinities

From its very early stages, quantum field theory was faced with infinities. They emerged in the calculation of most physical quantities, such as the correction to the charge of the electron due to the interactions with the radiation field. The way these divergences were handled in the 1940s, starting with Kramers, was physically very much in the spirit of the quantum theory emphasis in observable quantities: since the observed magnitude of physical quantities (such as the charge of the electron) is finite, this number should arise from the addition of a 'bare' (unobservable) value and the quantum corrections. The fact that both of these quantities were divergent was not a problem physically, since only its finite sum was an observable quantity. To make thing mathematically sound, the handling of infinities requires the introduction of some regularization procedure which cuts the divergent integrals off at some momentum scale Λ. Morally speaking, the physical value of an observable $\mathcal{O}_{\rm physical}$ is given by

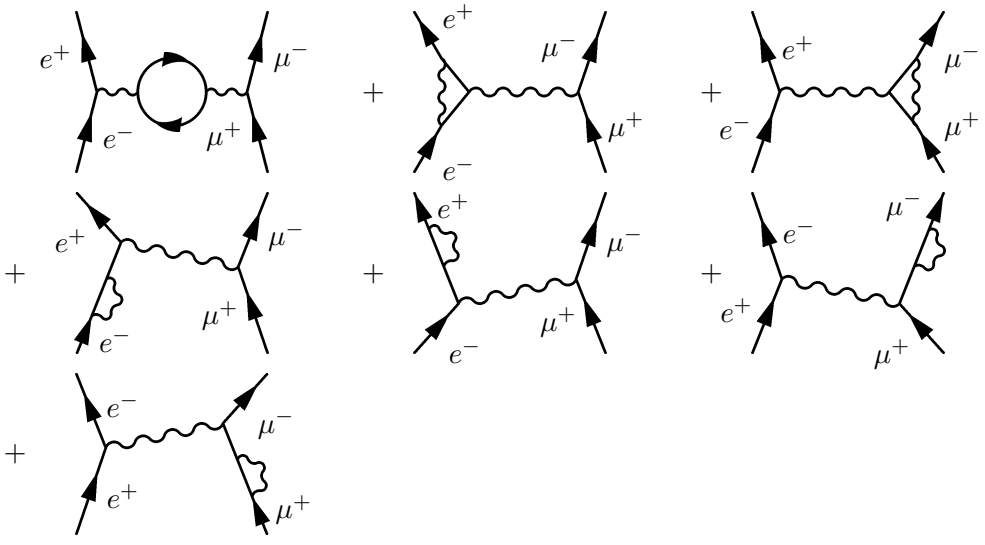
$$\mathcal{O}_{\rm physical} = \lim_{\Lambda\to\infty} \left[\mathcal{O}(\Lambda)_{\rm bare} + \Delta\mathcal{O}(\Lambda)_\hbar\right], \tag{8.1}$$

where $\Delta\mathcal{O}(\Lambda)_\hbar$ represents the regularized quantum corrections.

To make this qualitative discussion more precise we compute the corrections to the electric charge in quantum electrodynamics. We consider the process of annihilation of an electron–positron pair to create a muon–antimuon pair $e^-e^+ \to \mu^+\mu^-$. To lowest order in the electric charge e the only diagram contributing is

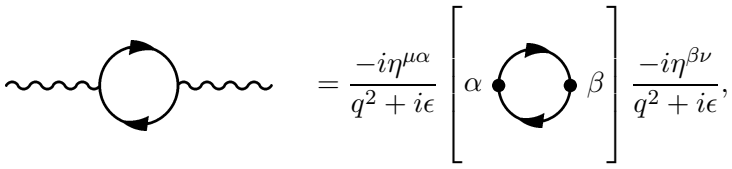


However, the corrections at order e^4 to this result requires the calculation of seven more diagrams



In order to compute the renormalization of the charge we consider the first diagram which takes into account the first correction to the propagator of the virtual photon interchanged between the pairs

due to vacuum polarization. We begin by evaluating



$$= \frac{-i\eta^{\mu\alpha}}{q^2+i\epsilon}\left[\alpha \quad \beta\right]\frac{-i\eta^{\beta\nu}}{q^2+i\epsilon}, \tag{8.2}$$

where the diagram between brackets is given by

$$\equiv \Pi^{\alpha\beta}(q) = i^2(-ie)^2(-1)\int \frac{d^4k}{(2\pi)^4}\frac{\mathrm{Tr}\,(\not{k}+m_e)\gamma^\alpha(\not{k}+\not{q}+m_e)\gamma^\beta}{[k^2-m_e^2+i\epsilon]\,[(k+q)^2-m_e^2+i\epsilon]}. \tag{8.3}$$

Physically this diagram includes the correction to the propagator due to the polarization of the vacuum, i.e., the creation of virtual electron–positron pairs by the propagating photon. The momentum q is the total momentum of the electron–positron pair in the intermediate channel.

It is instructive to look at this diagram from the point of view of perturbation theory in nonrelativistic quantum mechanics. In each vertex the interaction consists of the annihilation (creation) of a photon and the creation (annihilation) of an electron–positron pair. This can be implemented by the interaction Hamiltonian

$$H_{\rm int} = e\int d^3x\,\overline{\psi}\gamma^\mu\psi A_\mu. \tag{8.4}$$

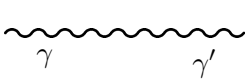
All fields inside the integral can be expressed in terms of the corresponding creation-annihilation operators for photons, electrons and positrons. In quantum mechanics, the change in the wave function at first order in the perturbation $H_{\rm int}$ is given by

$$|\gamma,{\rm in}\rangle = |\gamma,{\rm in}\rangle_0 + \sum_n \frac{\langle n|H_{\rm int}|\gamma,{\rm in}\rangle_0}{E_{\rm in}-E_n}|n\rangle \tag{8.5}$$

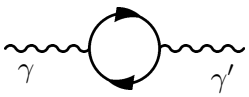
and similarly for $|\gamma,{\rm out}\rangle$, where we have denoted symbolically by $|n\rangle$ all the possible states of the electron–positron pair. Since these states are orthogonal to $|\gamma,{\rm in}\rangle_0$, $|\gamma,{\rm out}\rangle_0$, we find to order e^2

$$\langle\gamma,{\rm in}|\gamma',{\rm out}\rangle = {}_0\langle\gamma,{\rm in}|\gamma',{\rm out}\rangle_0 + \sum_n \frac{{}_0\langle\gamma,{\rm in}|H_{\rm int}|n\rangle\,\langle n|H_{\rm int}|\gamma',{\rm out}\rangle_0}{(E_{\rm in}-E_n)(E_{\rm out}-E_n)} + \mathcal{O}(e^4). \tag{8.6}$$

Hence, we see that the diagram of Eq. (8.2) really corresponds to the order-e^2 correction to the photon propagator $\langle\gamma,{\rm in}|\gamma',{\rm out}\rangle$



$$\longrightarrow \quad {}_0\langle\gamma,{\rm in}|\gamma',{\rm out}\rangle_0$$



$$\longrightarrow \quad \sum_n \frac{\langle\gamma,{\rm in}|H_{\rm int}|n\rangle\,\langle n|H_{\rm int}|\gamma',{\rm out}\rangle}{(E_{\rm in}-E_n)(E_{\rm out}-E_n)}. \tag{8.7}$$

Once we have understood the physical meaning of the Feynman diagram to be computed we proceed to its evaluation. In principle there is no problem in computing the integral in Eq. (8.2) for nonzero values of the electron mass. However, since here we are going to be mostly interested in seeing how the divergence of the integral results in a scale-dependent renormalization of the electric charge, we will set $m_e = 0$. This is something safe to do, since in the case of this diagram we are not inducing new infrared divergences in taking the electron as massless. Doing some γ-matrices gymnastics it is not complicated to show that the polarization tensor $\Pi_{\mu\nu}(q)$ defined in Eq. (8.3) can be written as

$$\Pi_{\mu\nu}(q) = \left(q^2\eta_{\mu\nu} - q_\mu q_\nu\right)\Pi(q^2) \tag{8.8}$$

with

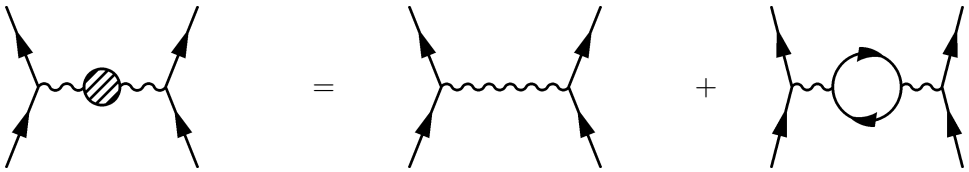
$$\Pi(q^2) = \frac{4e^2}{3q^2}\int \frac{d^4k}{(2\pi)^4}\frac{k^2 + k\cdot q}{[k^2+i\epsilon]\,[(k+q)^2+i\epsilon]}. \tag{8.9}$$

Although by naïve power counting we could conclude that the previous integral is quadratically divergent, it can be seen that the quadratic divergence actually cancels leaving behind only a logarithmic one. In order to handle this divergent integral we have to figure out some procedure to render it finite. This can be done in several ways, but here we choose to cut the integrals off at a high energy scale Λ, where new physics might be at work, $|p| < \Lambda$. This gives the result

$$\Pi(q^2) \simeq \frac{e^2}{12\pi^2}\log\left(\frac{q^2}{\Lambda^2}\right) + \text{finite terms}. \tag{8.10}$$

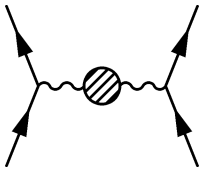
If we were to send the cutoff to infinity $\Lambda \to \infty$ the divergence blows up and something has to be done about it.

If we want to make sense out of this, we have to go back to the physical question that led us to compute Eq. (8.2). Our primordial motivation was to compute the corrections to the annihilation of two electrons into two muons. Including the correction to the propagator of the virtual photon we have



$$\begin{aligned} &= \eta_{\alpha\beta}\left(\overline{v}_e\gamma^\alpha u_e\right)\frac{e^2}{4\pi q^2}\left(\overline{v}_\mu\gamma^\beta u_\mu\right) + \eta_{\alpha\beta}\left(\overline{v}_e\gamma^\alpha u_e\right)\frac{e^2}{4\pi q^2}\Pi(q^2)\left(\overline{v}_\mu\gamma^\beta u_\mu\right) \\ &= \eta_{\alpha\beta}\left(\overline{v}_e\gamma^\alpha u_e\right)\left\{\frac{e^2}{4\pi q^2}\left[1 + \frac{e^2}{12\pi^2}\log\left(\frac{q^2}{\Lambda^2}\right)\right]\right\}\left(\overline{v}_\mu\gamma^\beta u_\mu\right). \end{aligned} \tag{8.11}$$

Now let us imagine that we are performing a $e^-\,e^+ \to \mu^-\mu^+$ with a centre-of-mass energy μ. From the previous result we can identify the effective charge of the particles at this energy scale $e(\mu)$ as



$$= \eta_{\alpha\beta}\left(\overline{v}_e\gamma^\alpha u_e\right)\left[\frac{e(\mu)^2}{4\pi q^2}\right]\left(\overline{v}_\mu\gamma^\beta u_\mu\right). \tag{8.12}$$

This charge, $e(\mu)$, is the quantity that is physically measurable in our experiment. Now we can make sense of the formally divergent result (8.11) by assuming that the charge appearing in the classical Lagrangian of QED is just a 'bare' value that depends on the scale Λ at which we cut off the theory,

$e \equiv e(\Lambda)_{\rm bare}$. In order to reconcile (8.11) with the physical results (8.12) we must assume that the dependence of the bare (unobservable) charge $e(\Lambda)_{\rm bare}$ on the cutoff Λ is determined by the identity

$$e(\mu)^2 = e(\Lambda)^2_{\rm bare}\left[1 + \frac{e(\Lambda)^2_{\rm bare}}{12\pi^2}\log\left(\frac{\mu^2}{\Lambda^2}\right)\right]. \tag{8.13}$$

If we still insist in removing the cutoff $\Lambda \to \infty$ we have to send the bare charge to zero $e(\Lambda)_{\rm bare} \to 0$ in such a way that the effective coupling has the finite value given by the experiment at the energy scale μ. It is not a problem, however, that the bare charge is small for large values of the cutoff, since the only measurable quantity is the effective charge that remains finite. Therefore all observable quantities should be expressed in perturbation theory as a power series in the physical coupling $e(\mu)^2$ and not in the unphysical bare coupling $e(\Lambda)_{\rm bare}$.

8.2 The beta-function and asymptotic freedom

We can look at the previous discussion, and in particular Eq. (8.13), from a different point of view. In order to remove the ambiguities associated with infinities we have been forced to introduce a dependence of the coupling constant on the energy scale at which a process takes place. From the expression of the physical coupling in terms of the bare charge (8.13) we can actually eliminate the cutoff Λ, whose value after all should not affect the value of physical quantities. Taking into account that we are working in perturbation theory in $e(\mu)^2$, we can express the bare charge $e(\Lambda)^2_{\rm bare}$ in terms of $e(\mu)^2$ as

$$e(\Lambda)^2 = e(\mu)^2\left[1 + \frac{e(\mu)^2}{12\pi^2}\log\left(\frac{\mu^2}{\Lambda^2}\right)\right] + \mathcal{O}[e(\mu)^6]. \tag{8.14}$$

This expression allows us to eliminate all dependence in the cutoff in the expression of the effective charge at a scale μ by replacing $e(\Lambda)_{\rm bare}$ in Eq. (8.13) by the one computed using (8.14) at a given reference energy scale μ_0

$$e(\mu)^2 = e(\mu_0)^2\left[1 + \frac{e(\mu_0)^2}{12\pi^2}\log\left(\frac{\mu^2}{\mu_0^2}\right)\right]. \tag{8.15}$$

From this equation we can compute, at this order in perturbation theory, the effective value of the coupling constant at an energy μ, once we know its value at some reference energy scale μ_0. In the case of the electron charge we can use as a reference Thomson scattering at energies of the order of the electron mass $m_e \simeq 0.5$ MeV, where the value of the electron charge is given by the well-known value

$$e(m_e)^2 \simeq \frac{1}{137}. \tag{8.16}$$

With this we can compute $e(\mu)^2$ at any other energy scale applying Eq. (8.15), for example at the electron mass $\mu = m_e \simeq 0.5\,$MeV. However, in computing the electromagnetic coupling constant at any other scale we must take into account the fact that other charged particles can run in the loop in Eq. (8.11). Suppose, for example, that we want to calculate the fine structure constant at the mass of the Z^0-boson $\mu = M_Z \equiv 92$ GeV. Then we should include in Eq. (8.15) the effect of other fermionic Standard Model fields with masses below M_Z. Doing this, we find[18]

$$e(M_Z)^2 = e(m_e)^2\left[1 + \frac{e(m_e)^2}{12\pi^2}\left(\sum_i q_i^2\right)\log\left(\frac{M_Z^2}{m_e^2}\right)\right], \tag{8.17}$$

[18] In the first version of these notes the argument used to show the growing of the electromagnetic coupling constant could have led to confusion to some readers. To avoid this potential problem we include in the equation for the running coupling $e(\mu)^2$ the contribution of all fermions with masses below M_Z. We thank Lubos Motl for bringing this issue to our attention.

where q_i is the charge in units of the electron charge of the i-th fermionic species running in the loop and we sum over all fermions with masses below the mass of the Z^0 boson. This expression shows how the electromagnetic coupling grows with energy. However, in order to compare with the experimental value of $e(M_Z)^2$ it is not enough to include the effect of fermionic fields, since the $W^\pm$ bosons also can run in the loop ($M_W < M_Z$). Taking this into account, as well as threshold effects, the value of the electron charge at the scale M_Z is found to be [35]

$$e(M_Z)^2 \simeq \frac{1}{128.9} \ . \tag{8.18}$$

This growing of the effective fine structure constant with energy can be understood heuristically by remembering that the effect of the polarization of the vacuum shown in the diagram of Eq. (8.2) amounts to the creation of a plethora of electron–positron pairs around the location of the charge. These virtual pairs behave as dipoles that, as in a dielectric medium, tend to screen this charge and decrease its value at long distances (i.e. lower energies).

The variation of the coupling constant with energy is usually encoded in quantum field theory in the *beta function* defined by

$$\beta(g) = \mu \frac{dg}{d\mu}. \tag{8.19}$$

In the case of QED the beta function can be computed from Eq. (8.15) with the result

$$\beta(e)_{\rm QED} = \frac{e^3}{12\pi^2}. \tag{8.20}$$

The fact that the coefficient of the leading term in the beta function is positive $\beta_0 \equiv \frac{1}{6\pi} > 0$ gives us the overall behavior of the coupling as we change the scale. Equation (8.20) means that, if we start at an energy where the electric coupling is small enough for our perturbative treatment to be valid, the effective charge grows with the energy scale. This growing of the effective coupling constant with energy means that QED is infrared safe, since the perturbative approximation gives better and better results as we go to lower energies. Actually, because the electron is the lighter electrically charged particle and has a finite nonvanishing mass, the running of the fine structure constant stops at the scale m_e in the well-known value $\frac{1}{137}$. Were other charged fermions with masses below m_e present in Nature, the effective value of the fine structure constant in the interaction between these particles would run further to lower values at energies below the electron mass.

On the other hand, if we increase the energy scale, $e(\mu)^2$ grows until at some scale the coupling is of order one and the perturbative approximation breaks down. In QED this is known as the problem of the Landau pole but in fact it does not pose any serious threat to the reliability of QED perturbation theory: a simple calculation shows that the energy scale at which the theory would become strongly coupled is $\Lambda_{\rm Landau} \simeq 10^{277}$ GeV. However, we know that QED does not live that long! At much lower scales we expect electromagnetism to be unified with other interactions, and even if this is not the case we will enter the uncharted territory of quantum gravity at energies of the order of 10^{19} GeV.

So much for QED. The next question that one may ask at this stage is whether it is possible to find quantum field theories with a behavior opposite to that of QED, i.e., such that they become weakly coupled at high energies. This is not a purely academic question. In the late 1960s a series of deep-inelastic scattering experiments carried out at SLAC showed that the quarks behave essentially as free particles inside hadrons. The apparent problem was that no theory was known at that time that would become free at very short distances: the example set by QED seemed to be followed by all the theories that were studied. This posed a very serious problem for quantum field theory as a way to describe subnuclear physics, since it seemed that its predictive power was restricted to electrodynamics but failed miserably when applied to describe strong interactions.

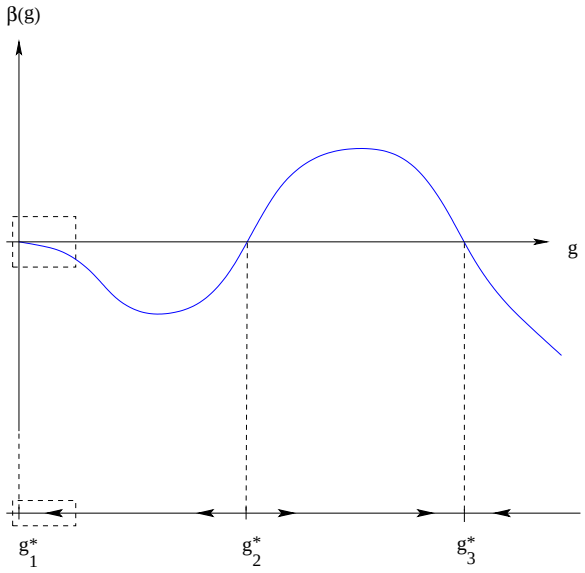


Fig. 14: Beta function for a hypothetical theory with three fixed points g_1^*, g_2^* and g_3^*. A perturbative analysis would capture only the regions shown in the boxes.

Nevertheless, this critical time for quantum field theory turned out to be its finest hour. In 1973 David Gross and Frank Wilczek [36] and David Politzer [37] showed that non-Abelian gauge theories can actually display the required behavior. For the QCD Lagrangian in Eq. (7.29) the beta function is given by[19]

$$\beta(g) = -\frac{g^3}{16\pi^2}\left[\frac{11}{3}N_c - \frac{2}{3}N_f\right]. \tag{8.21}$$

In particular, for real QCD ($N_C = 3$, $N_f = 6$) we have $\beta(g) = -\frac{7g^3}{16\pi^2} < 0$. This means that for a theory that is weakly coupled at an energy scale μ_0 the coupling constant decreases as the energy increases $\mu \to \infty$. This explain the apparent freedom of quarks inside the hadrons: when the quarks are very close together their effective color charge tends to zero. This phenomenon is called *asymptotic freedom*.

Asymptotic free theories display a behavior that is opposite to that found above in QED. At high energies their coupling constant approaches zero whereas at low energies they become strongly coupled (infrared slavery). These features are at the heart of the success of QCD as a theory of strong interactions, since this is exactly the type of behavior found in quarks: they are quasi-free particles inside the hadrons but the interaction potential between them increases at large distances.

Although asymptotic free theories can be handled in the ultraviolet, they become extremely complicated in the infrared. In the case of QCD it is still to be understood (at least analytically) how the theory confines color charges and generates the spectrum of hadrons, as well as the breaking of the chiral symmetry (7.43).

In general, the ultraviolet and infrared properties of a theory are controlled by the fixed points of the beta function, i.e., those values of the coupling constant g for which it vanishes

$$\beta(g^*) = 0. \tag{8.22}$$

Using perturbation theory we have seen that for both QED and QCD one of such fixed points occurs at zero coupling, $g^* = 0$. However, our analysis also showed that the two theories present radically

[19] The expression of the beta function of QCD was also known to 't Hooft [38]. There are even earlier computations in the Russian literature [39].

different behavior at high and low energies. From the point of view of the beta function, the difference lies in the energy regime at which the coupling constant approaches its critical value. This is in fact governed by the sign of the beta function around the critical coupling.

We have seen above that when the beta function is negative close to the fixed point (the case of QCD) the coupling tends to its critical value, $g^* = 0$, as the energy is increased. This means that the critical point is *ultraviolet stable*, i.e., it is an attractor as we evolve towards higher energies. If, on the contrary, the beta function is positive (as happens in QED) the coupling constant approaches the critical value as the energy decreases. This is the case of an *infrared stable* fixed point.

This analysis that we have motivated with the examples of QED and QCD is completely general and can be carried out for any quantum field theory. In Fig. 14 we have represented the beta function for a hypothetical theory with three fixed points located at couplings g_1^*, g_2^* and g_3^*. The arrows in the line below the plot represent the evolution of the coupling constant as the energy increases. From the analysis presented above we see that $g_1^* = 0$ and g_3^* are ultraviolet stable fixed points, while the fixed point g_2^* is infrared stable.

In order to understand the high- and low-energy behavior of a quantum field theory it is then crucial to know the structure of the beta functions associated with its couplings. This can be a very difficult task, since perturbation theory only allows the study of the theory around 'trivial' fixed points, i.e., those that occur at zero coupling like the case of g_1^* in Fig. 14. On the other hand, any 'nontrivial' fixed point occurring in a theory (like g_2^* and g_3^*) cannot be captured in perturbation theory and requires a full nonperturbative analysis.

The moral to be learned from our discussion above is that dealing with the ultraviolet divergences in a quantum field theory has the consequence, among others, of introducing an energy dependence in the measured value of the coupling constants of the theory (for example the electric charge in QED). This happens even in the case of renormalizable theories without mass terms. These theories are scale invariant at the classical level because the action does not contain any dimensionful parameter. In this case the running of the coupling constants can be seen as resulting from a quantum breaking of classical scale invariance: different energy scales in the theory are distinguished by different values of the coupling constants. Remembering what we learned in Section 7, we conclude that classical scale invariance is an anomalous symmetry. One heuristic way to see how the conformal anomaly comes about is to notice that the regularization of an otherwise scale-invariant field theory requires the introduction of an energy scale (e.g., a cutoff). This breaking of scale invariance cannot be restored after renormalization.

Nevertheless, scale invariance is not lost forever in the quantum theory. It is recovered at the fixed points of the beta function where, by definition, the coupling does not run. To understand how this happens we go back to a scale-invariant classical field theory whose field $\phi(x)$ transforms under coordinate rescalings as

$$x^\mu \longrightarrow \lambda x^\mu, \qquad \phi(x) \longrightarrow \lambda^{-\Delta}\phi(\lambda^{-1}x), \tag{8.23}$$

where Δ is called the canonical scaling dimension of the field. An example of such a theory is a massless ϕ^4 theory in four dimensions

$$\mathcal{L} = \frac{1}{2}\partial_\mu\phi\,\partial^\mu\phi - \frac{g}{4!}\phi^4, \tag{8.24}$$

where the scalar field has canonical scaling dimension $\Delta = 1$. The Lagrangian density transforms as

$$\mathcal{L} \longrightarrow \lambda^{-4}\mathcal{L}[\phi] \tag{8.25}$$

and the classical action remains invariant[20].

[20] In a D-dimensional theory the canonical scaling dimensions of the fields coincide with its engineering dimension: $\Delta = \frac{D-2}{2}$ for bosonic fields and $\Delta = \frac{D-1}{2}$ for fermionic ones. For a Lagrangian with no dimensionful parameters classical scale invariance follows then from dimensional analysis.

We look at the free theory $g = 0$ for a moment. Now there are no divergences and all correlation functions can be exactly computed. In particular we consider the momentum space n-point correlation function

$$G_0(p_1, \ldots, p_n)(2\pi)^4 \delta^{(4)}(p_1 + \ldots + p_n) = \int d^4x_1 \ldots d^4x_n \, e^{ip_1 \cdot x_1 + \ldots + ip_n \cdot x_n} \langle 0 | T\Big[\phi_0(x_1) \ldots \phi_0(x_n)\Big] | 0 \rangle, \tag{8.26}$$

where by $\phi_0(x)$ we denote the free field operator. Applying the rescaling (8.23) we find the following transformation for the correlation function

$$G_0(p_1, \ldots, p_n) \longrightarrow \lambda^{4(n-1)-n\Delta} G_0(\lambda p_1, \ldots, \lambda p_n). \tag{8.27}$$

For the free theory the only relevant correlation function is the two-point function, where we have (remember that we are dealing with a massless theory)

$$G_0(p^2) = \frac{i}{p^2} \longrightarrow \lambda^2 G_0(\lambda^2 p^2) = \frac{i}{p^2}. \tag{8.28}$$

The transformation of any other correlation function follows from this result and Wick's theorem, that allows to write any the $2n$-correlation function as sum of products of n 2-point correlation functions (correlation functions with an odd number of fields are identically zero).

We turn to the interacting theory. Things now get much more complicated, since correlation functions cannot be exactly computed in general. However, when the theory sits at the critical coupling we can use a few useful facts. For example, since the critical theory is scale invariant, it should either contain only massless one-particle states or have continuous spectrum. To keep the argument simple, we consider the first possibility. Hence, the exact two-point function should have a pole at $p^2 = 0$, and close to this pole the correlation function has the form

$$G(p^2; \mu) \approx \frac{iZ(\mu)}{p^2}, \tag{8.29}$$

where $Z(\mu)$, called the field renormalization, depends on the scale. The *anomalous dimension* $\gamma(g)$ is then defined by the equation

$$\gamma(g) = \frac{1}{2} \mu \frac{d}{d\mu} \log Z. \tag{8.30}$$

This new function is the analog of the beta function (8.19) for the field renormalization $Z(\mu)$. Moreover, at the critical point $g(\mu) = g^*$ and the anomalous dimension is independent of the energy, $\gamma^* = \gamma(g^*)$. In this case Eq. (8.30) can be integrated to give

$$Z(\mu) = Z_0 \left(\frac{\mu}{\mu_0} \right)^{2\gamma^*}, \tag{8.31}$$

where Z_0 and μ_0 are some reference values. Then, we find that the two-point function at the critical point is invariant under the rescaling

$$G(p^2; \mu) \longrightarrow \lambda^{2(1-\gamma^*)} G(\lambda^2 p^2; \lambda\mu). \tag{8.32}$$

Here we have presented a rather sketchy and heuristic argument. A more thorough analysis (using for example the Callan–Symanzik equation [1–10]) shows that at the critical point all n-point correlation functions are invariant under the rescaling

$$G(p_1, \ldots, p_n; \mu) \longrightarrow \lambda^{4(n-1)-n(\Delta+\gamma^*)} G(\lambda p_1, \ldots, \lambda p_n; \lambda\mu). \tag{8.33}$$

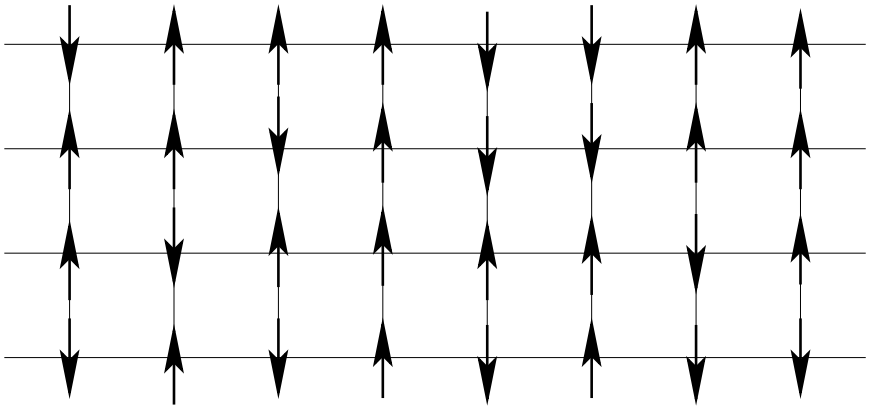


Fig. 15: Systems of spins in a two-dimensional square lattice

Comparing (8.32) and (8.33) with (8.27) we see that this invariance is analogous to that of the free (scale-invariant) theory. Now, however, the fields transform under rescalings with an anomalous scaling dimension given by

$$\Delta_{\rm anom} = \Delta + \gamma^*, \tag{8.34}$$

with Δ the canonical scaling dimension of the corresponding field. This justifies the name given to the function $\gamma(g)$ defined in Eq. (8.30). Notice, however, that strictly speaking $\gamma(g)$ represents an anomalous dimension for the theory only at the critical coupling g^*.

The previous discussion clarifies a little bit the high-energy properties of an asymptotically free theory like QCD. The fact that the fixed point occurs at zero coupling might give the wrong impression that the theory at the critical point is just the one obtained by setting $g = 0$ in the action. Life, however, is more complicated than that. What we have seen above shows that although the critical theory is a free scale-invariant field theory, the fields have anomalous scaling dimensions which are different from the ones of the 'naive' free theory. These anomalous dimensions carry the dynamical information about the high-energy behavior of the asymptotically free theory.

8.3 The renormalization group

In spite of its successes, the renormalization procedure presented above can be seen as some kind of prescription or recipe to get rid of the divergences in an ordered way. This discomfort about renormalization was expressed on occasion by comparing it with "sweeping the infinities under the rug". However, thanks to Ken Wilson to a large extent [40], the process of renormalization is now understood in a very profound way as a procedure to incorporate the effects of physics at high energies by modifying the value of the parameters that appear in the Lagrangian.

Statistical mechanics. Wilson's ideas are both simple and profound and consist in thinking about quantum field theory as the analog of a thermodynamical description of a statistical system. To be more precise, let us consider an Ising spin system in a two-dimensional square lattice like the one depicted in Fig 15. In terms of the spin variables $s_i = \pm\frac{1}{2}$, where i labels the lattice site, the Hamiltonian of the system is given by

$$H = -J \sum_{\langle i,j \rangle} s_i \, s_j, \tag{8.35}$$

where $\langle i, j\rangle$ indicates that the sum extends over nearest neighbors and J is the coupling constant between neighboring spins (here we consider that there is no external magnetic field). The starting point to study

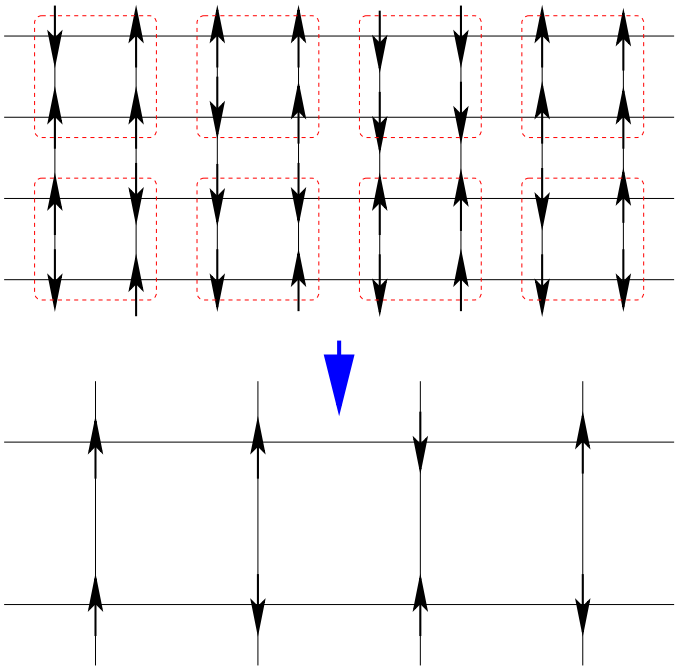


Fig. 16: Decimation of the spin lattice. Each block in the upper lattice is replaced by an effective spin computed according to the rule (8.39). Notice also that the size of the lattice spacing is doubled in the process.

the statistical mechanics of this system is the partition function defined as

$$\mathcal{Z} = \sum_{\{s_i\}} e^{-\beta H}, \tag{8.36}$$

where the sum is over all possible configurations of the spins and $\beta = \frac{1}{T}$ is the inverse temperature. For $J > 0$ the Ising model presents spontaneous magnetization below a critical temperature T_c, in any dimension higher than one. Away from this temperature correlations between spins decay exponentially at large distances

$$\langle s_i s_j \rangle \sim e^{-\frac{|x_{ij}|}{\xi}}, \tag{8.37}$$

with $|x_{ij}|$ the distance between the spins located in the i-th and j-th sites of the lattice. This expression serves as a definition of the correlation length ξ which sets the characteristic length scale at which spins can influence each other by their interaction through their nearest neighbors.

Suppose now that we are interested in a macroscopic description of this spin system. We can capture the relevant physics by integrating out somehow the physics at short scales. A way in which this can be done was proposed by Leo Kadanoff [41] and consists in dividing our spin system in spin-blocks like the ones shown in Fig. 16. Now we can construct another spin system where each spin-block of the original lattice is replaced by an effective spin calculated according to some rule from the spins contained in each block B_a

$$\{s_i : i \in B_a\} \quad \longrightarrow \quad {s_a}^{(1)}. \tag{8.38}$$

For example, we can define the effective spin associated with the block B_a by taking the majority rule with an additional prescription in case of a draw

$${s_a}^{(1)} = \frac{1}{2}\mathrm{sgn}\left(\sum_{i \in B_a} s_i\right), \tag{8.39}$$

where we have used the sign function, $\mathrm{sign}(x) \equiv \frac{x}{|x|}$, with the additional definition $\mathrm{sgn}(0) = 1$. This procedure is called decimation and leads to a new spin system with a doubled lattice space.

The idea now is to rewrite the partition function (8.36) only in terms of the new effective spins ${s_a}^{(1)}$. Then we start by splitting the sum over spin configurations into two nested sums, one over the spin-blocks and a second one over the spins within each block

$$\mathcal{Z} = \sum_{\{\vec{s}\}} e^{-\beta H[s_i]} = \sum_{\{\vec{s}^{\,(1)}\}} \sum_{\{\vec{s} \in B_a\}} \delta \left[{s_a}^{(1)} - \mathrm{sign}\left(\sum_{i \in B_a} s_i \right) \right] e^{-\beta H[s_i]}. \tag{8.40}$$

The interesting point now is that the sum over spins inside each block can be written as the exponential of a new effective Hamiltonian depending only on the effective spins, $H^{(1)}[{s_a}^{(1)}]$

$$\sum_{\{s \in B_a\}} \delta \left[{s_a}^{(1)} - \mathrm{sign}\left(\sum_{i \in B_a} s_i \right) \right] e^{-\beta H[s_i]} = e^{-\beta H^{(1)}[{s_a}^{(1)}]}. \tag{8.41}$$

The new Hamiltonian is of course more complicated

$$H^{(1)} = -J^{(1)} \sum_{\langle i,j \rangle} s_i^{(1)} s_j^{(1)} + \dots \tag{8.42}$$

where the dots stand for other interaction terms between the effective spin block. These new terms appear because in the process of integrating out short distance physics we induce interactions between the new effective degrees of freedom. For example the interaction between the spin-block variables $s_i^{(1)}$ will in general not be restricted to nearest neighbors in the new lattice. The important point is that we have managed to rewrite the partition function solely in terms of this new (renormalized) spin variables $s^{\,(1)}$ interacting through a new Hamiltonian $H^{(1)}$

$$\mathcal{Z} = \sum_{\{s^{\,(1)}\}} e^{-\beta H^{(1)}[{s_a}^{(1)}]}. \tag{8.43}$$

Let us now think about the space of all possible Hamiltonians for our statistical system including all kinds of possible couplings between the individual spins compatible with the symmetries of the system. If we denote by $\mathcal{R}$ the decimation operation, our previous analysis shows that $\mathcal{R}$ defines a map in this space of Hamiltonians

$$\mathcal{R} : H \to H^{(1)}. \tag{8.44}$$

At the same time the operation $\mathcal{R}$ replaces a lattice with spacing a by another one with double spacing $2a$. As a consequence the correlation length in the new lattice measured in units of the lattice spacing is divided by two, $\mathcal{R} : \xi \to \frac{\xi}{2}$.

Now we can iterate the operation $\mathcal{R}$ an indefinite number of times. Eventually we might reach a Hamiltonian $H_\star$ that is not further modified by the operation $\mathcal{R}$

$$H \xrightarrow{\mathcal{R}} H^{(1)} \xrightarrow{\mathcal{R}} H^{(2)} \xrightarrow{\mathcal{R}} \dots \xrightarrow{\mathcal{R}} H_\star. \tag{8.45}$$

The fixed-point Hamiltonian $H_\star$ is *scale invariant* because it does not change as $\mathcal{R}$ is performed. Notice that because of this invariance the correlation length of the system at the fixed point does not change under $\mathcal{R}$. This fact is compatible with the transformation $\xi \to \frac{\xi}{2}$ only if $\xi = 0$ or $\xi = \infty$. Here we will focus on the case of nontrivial fixed points with infinite correlation length.

The space of Hamiltonians can be parametrized by specifying the values of the coupling constants associated with all possible interaction terms between individual spins of the lattice. If we denote by

$\mathcal{O}_a[s_i]$ these (possibly infinite) interaction terms, the most general Hamiltonian for the spin system under study can be written as

$$H[s_i] = \sum_{a=1}^{\infty} \lambda_a \mathcal{O}_a[s_i], \tag{8.46}$$

where $\lambda_a \in \mathbb{R}$ are the coupling constants for the corresponding operators. These constants can be thought of as coordinates in the space of all Hamiltonians. Therefore the operation $\mathcal{R}$ defines a transformation in the set of coupling constants

$$\mathcal{R} : \lambda_a \longrightarrow \lambda_a^{(1)}. \tag{8.47}$$

For example, in our case we started with a Hamiltonian in which only one of the coupling constants is different from zero (say $\lambda_1 = -J$). As a result of the decimation $\lambda_1 \equiv -J \to -J^{(1)}$ while some of the originally vanishing coupling constants will take a nonzero value. Of course, for the fixed point Hamiltonian the coupling constants do not change under the scale transformation $\mathcal{R}$.

Physically the transformation $\mathcal{R}$ integrates out short distance physics. The consequence for physics at long distances is that we have to replace our Hamiltonian by a new one with different values for the coupling constants. That is, our ignorance of the details of the physics going on at short distances results in a *renormalization* of the coupling constants of the Hamiltonian that describes the long range physical processes. It is important to stress that although $\mathcal{R}$ is sometimes called a renormalization group transformation, in fact this is a misnomer. Transformations between Hamiltonians defined by $\mathcal{R}$ do not form a group: since these transformations proceed by integrating out degrees of freedom at short scales they cannot be inverted.

In statistical mechanics fixed points under renormalization group transformations with $\xi = \infty$ are associated with phase transitions. From our previous discussion we can conclude that the space of Hamiltonians is divided into regions corresponding to the basins of attraction of the different fixed points. We can ask ourselves now about the stability of those fixed points. Suppose we have a statistical system described by a fixed-point Hamiltonian $H_\star$ and we perturb it by changing the coupling constant associated with an interaction term $\mathcal{O}$. This is equivalent to replacing $H_\star$ by the perturbed Hamiltonian

$$H = H_\star + \delta\lambda\, \mathcal{O}, \tag{8.48}$$

where $\delta\lambda$ is the perturbation of the coupling constant corresponding to $\mathcal{O}$ (we can also consider perturbations in more than one coupling constant). At the same time thinking of the λ_a's as coordinates in the space of all Hamiltonians, this corresponds to moving slightly away from the position of the fixed point.

The question to decide now is in which direction the renormalization group flow will take the perturbed system. Working at first order in $\delta\lambda$ there are three possibilities:

- The renormalization group flow takes the system back to the fixed point. In this case the corresponding interaction $\mathcal{O}$ is called *irrelevant*.
- $\mathcal{R}$ takes the system away from the fixed point. If this is what happens the interaction is called *relevant*.
- It is possible that the perturbation actually does not take the system away from the fixed point at first order in $\delta\lambda$. In this case the interaction is said to be *marginal* and it is necessary to go to higher orders in $\delta\lambda$ in order to decide whether the system moves towards or away from the fixed point, or whether we have a family of fixed points.

Therefore we can picture the action of the renormalization group transformation as a flow in the space of coupling constants. In Fig. 17 we have depicted an example of such a flow in the case of a system with two coupling constants λ_1 and λ_2. In this example we find two fixed points, one at the

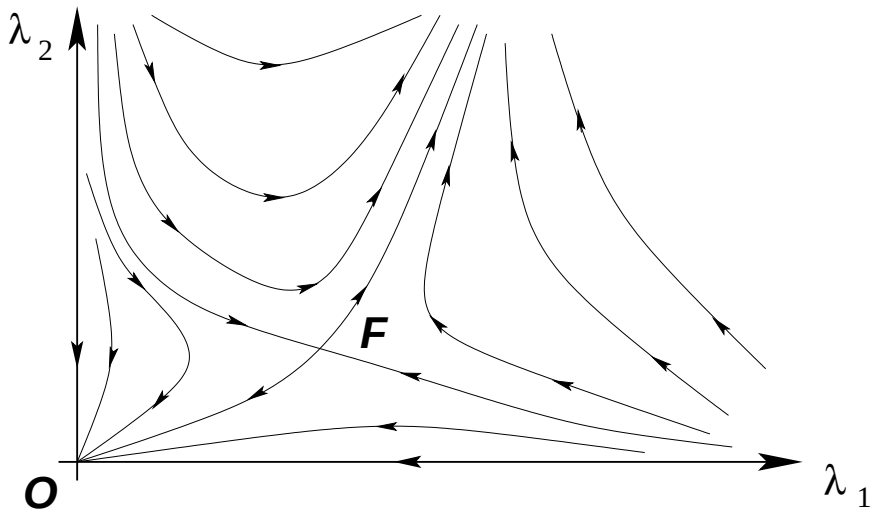


Fig. 17: Example of a renormalization group flow

origin O and another at F for a finite value of the couplings. The arrows indicate the direction in which the renormalization group flow acts. The free theory at $\lambda_1 = \lambda_2 = 0$ is a stable fixed point since any perturbation $\delta\lambda_1, \delta\lambda_2 > 0$ makes the theory flow back to the free theory at long distances. On the other hand, the fixed point F is stable with respect to certain types of perturbation (along the line with incoming arrows) whereas for any other perturbations the system flows either to the free theory at the origin or to a theory with infinite values for the couplings.

Quantum field theory. Let us see now how these ideas of the renormalization group apply to Field Theory. Let us begin with a quantum field theory defined by the Lagrangian

$$\mathcal{L}[\phi_a] = \mathcal{L}_0[\phi_a] + \sum_i g_i \mathcal{O}_i[\phi_a], \tag{8.49}$$

where $\mathcal{L}_0[\phi_a]$ is the kinetic part of the Lagrangian and g_i are the coupling constants associated with the operators $\mathcal{O}_i[\phi_a]$. In order to make sense of the quantum theory we introduce a cutoff in momenta Λ. In principle we include all operators $\mathcal{O}_i$ compatible with the symmetries of the theory.

In Section 8.2 we saw how in the case of QED and QCD, the value of the coupling constant changed with the scale from its value at the scale Λ. We can understand now this behavior along the lines of the analysis presented above for the Ising model. If we would like to compute the effective dynamics of the theory at an energy scale $\mu < \Lambda$ we only have to integrate out all physical models with energies between the cutoff Λ and the scale of interest μ. This is analogous to what we did in the Ising model by replacing the original spins by the spin block. In the case of field theory the effective action $S[\phi_a, \mu]$ at scale μ can be written in the language of functional integration as

$$e^{iS[\phi'_a,\mu]} = \int_{\mu<p<\Lambda} \prod_a \mathcal{D}\phi_a \, e^{iS[\phi_a,\Lambda]}. \tag{8.50}$$

Here $S[\phi_a, \Lambda]$ is the action at the cutoff scale

$$S[\phi_a, \Lambda] = \int d^4x \left\{ \mathcal{L}_0[\phi_a] + \sum_i g_i(\Lambda)\mathcal{O}_i[\phi_a] \right\} \tag{8.51}$$

and the functional integral in Eq. (8.50) is carried out only over the field modes with momenta in the range $\mu < p < \Lambda$. The action resulting from integrating out the physics at the intermediate scales between Λ and μ depends not on the original field variable ϕ_a but on some renormalized field ϕ'_a. At

the same time the couplings $g_i(\mu)$ differ from their values at the cutoff scale $g_i(\Lambda)$. This is analogous to what we learned in the Ising model: by integrating out short distance physics we ended up with a new Hamiltonian depending on renormalized effective spin variables and with renormalized values for the coupling constants. Therefore the resulting effective action at scale μ can be written as

$$S[\phi'_a, \mu] = \int d^4x \left\{ \mathcal{L}_0[\phi'_a] + \sum_i g_i(\mu) \mathcal{O}_i[\phi'_a] \right\}. \tag{8.52}$$

This Wilsonian interpretation of renormalization sheds light on what in Section 8.1 might have looked just a smart way to get rid of the infinities. The running of the coupling constant with the energy scale can be understood now as a way of incorporating into an effective action at scale μ the effects of field excitations at higher energies $E > \mu$.

As in statistical mechanics there are also quantum field theories that are fixed points of the renormalization group flow, i.e., whose coupling constants do not change with the scale. We have encountered them already in Section 8.2 when studying the properties of the beta function. The most trivial example of such theories are massless free quantum field theories, but there are also examples of four-dimensional interacting quantum field theories which are scale invariant. Again we can ask the question of what happens when a scale-invariant theory is perturbed with some operator. In general the perturbed theory is not scale invariant anymore but we may wonder whether the perturbed theory flows at low energies towards or away from the theory at the fixed point.

In quantum field theory this can be decided by looking at the canonical dimension $d[\mathcal{O}]$ of the operator $\mathcal{O}[\phi_a]$ used to perturb the theory at the fixed point. In four dimensions the three possibilities are defined by:

- $d[\mathcal{O}] > 4$: irrelevant perturbation. The running of the coupling constants takes the theory back to the fixed point.
- $d[\mathcal{O}] < 4$: relevant perturbation. At low energies the theory flows away from the scale-invariant theory.
- $d[\mathcal{O}] = 4$: marginal deformation. The direction of the flow cannot be decided only on dimensional grounds.

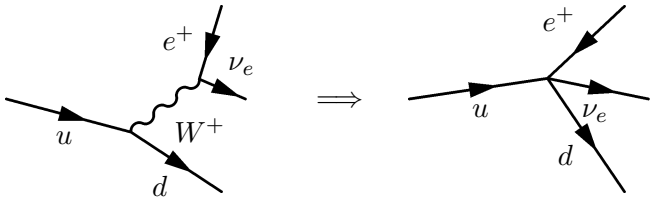
As an example, let us consider first a massless fermion theory perturbed by a four-fermion interaction term

$$\mathcal{L} = i\overline{\psi}\partial\!\!\!/\psi - \frac{1}{M^2}(\overline{\psi}\psi)^2. \tag{8.53}$$

This is indeed a perturbation by an irrelevant operator, since in four dimensions $[\psi] = \frac{3}{2}$. Interactions generated by the extra term are suppressed at low energies since typically their effects are weighted by the dimensionless factor $\frac{E^2}{M^2}$, where E is the energy scale of the process. This means that as we try to capture the relevant physics at lower and lower energies the effect of the perturbation is weaker and weaker rendering in the infrared limit $E \to 0$ again a free theory. Hence, the irrelevant perturbation in (8.53) makes the theory flow back to the fixed point.

On the other hand, relevant operators dominate the physics at low energies. This is the case, for example, of a mass term. As we lower the energy the mass becomes more important and once the energy goes below the mass of the field its dynamics is completely dominated by the mass term. This is, for example, how Fermi's theory of weak interactions emerges from the Standard Model at energies below

the mass of the $W^{\pm}$ boson



At energies below $M_W = 80.4$ GeV the dynamics of the W^+ boson is dominated by its mass term and therefore becomes nonpropagating, giving rise to the effective four-fermion Fermi theory.

To summarize our discussion so far, we found that while relevant operators dominate the dynamics in the infrared, taking the theory away from the fixed point, irrelevant perturbations become suppressed in the same limit. Finally we consider the effect of marginal operators. As an example we take the interaction term in massless QED, $\mathcal{O} = \overline{\psi}\gamma^\mu\psi\, A_\mu$. Taking into account that in $d = 4$ the dimension of the electromagnetic potential is $[A_\mu] = 1$, the operator $\mathcal{O}$ is a marginal perturbation. In order to decide whether the fixed-point theory

$$\mathcal{L}_0 = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\overline{\psi}\not{D}\,\psi \tag{8.54}$$

is restored at low energies or not, we need to study the perturbed theory in more detail. This we have done in Section 8.1 where we learned that the effective coupling in QED decreases at low energies. Then we conclude that the perturbed theory flows towards the fixed point in the infrared.

As an example of a marginal operator with the opposite behavior we can write the Lagrangian for a SU(N_c) gauge theory, $\mathcal{L} = -\frac{1}{4}F^a_{\mu\nu}F^{a\,\mu\nu}$, as

$$\begin{aligned}\mathcal{L} &= -\frac{1}{4}\left(\partial_\mu A^a_\nu - \partial_\nu A^a_\mu\right)\left(\partial^\mu A^{a\,\nu} - \partial^\nu A^{a\,\mu}\right) - 4g f^{abc} A^a_\mu A^b_\nu\, \partial^\mu A^{c\,\nu} \\ &+ g^2 f^{abc} f^{ade} A^b_\mu A^c_\nu A^{d\,\mu} A^{e\,\nu} \equiv \mathcal{L}_0 + \mathcal{O}_g,\end{aligned} \tag{8.55}$$

i.e., a marginal perturbation of the free theory described by $\mathcal{L}_0$, which is obviously a fixed point under renormalization group transformations. Unlike the case of QED we know that the full theory is asymptotically free, so the coupling constant grows at low energies. This implies that the operator $\mathcal{O}_g$ becomes more and more important in the infrared and therefore the theory flows away the fixed point in this limit.

It is very important to notice here that in the Wilsonian view the cutoff is not necessarily regarded as just some artifact to remove infinities but actually has a physical origin. For example, in the case of Fermi's theory of β-decay there is a natural cutoff $\Lambda = M_W$ at which the theory has to be replaced by the Standard Model. In the case of the Standard Model itself the cutoff can be taken at Planck scale $\Lambda \simeq 10^{19}$ GeV or the Grand Unification scale $\Lambda \simeq 10^{16}$ GeV, where new degrees of freedom are expected to become relevant. The cutoff serves the purpose of cloaking the range of energies at which new physics has to be taken into account.

Provided that in the Wilsonian approach the quantum theory is always defined with a physical cutoff, there is no fundamental difference between renormalizable and nonrenormalizable theories. Actually, a renormalizable field theory, like the Standard Model, can generate nonrenormalizable operators at low energies such as the effective four-fermion interaction of Fermi's theory. They are not sources of any trouble if we are interested in the physics at scales much below the cutoff, $E \ll \Lambda$, since their contribution to the amplitudes will be suppressed by powers of E/Λ.

9 Special topics

9.1 Creation of particles by classical fields

Particle creation by a classical source. In a free quantum field theory the total number of particles contained in a given state of the field is a conserved quantity. For example, in the case of the quantum

scalar field studied in Section 3 the number operator commutes with the Hamiltonian

$$\widehat{n} \equiv \int \frac{d^3k}{(2\pi)^3}\alpha^\dagger(\vec{k})\alpha(\vec{k}), \qquad [\widehat{H},\widehat{n}] = 0. \tag{9.1}$$

This means that any states with a well-defined number of particle excitations will preserve this number at all times. The situation, however, changes as soon as interactions are introduced, since in this case particles can be created and/or destroyed as a result of the dynamics.

Another case in which the number of particles might change is if the quantum theory is coupled to a classical source. The archetypical example of such a situation is the Schwinger effect, in which a classical strong electric field produces the creation of electron–positron pairs out of the vacuum. However, before plunging into this more involved situation we can illustrate the relevant physics involved in the creation of particles by classical sources with the help of the simplest example: a free scalar field theory coupled to a classical external source $J(x)$. The action for such a theory can be written as

$$S = \int d^4x \left[\frac{1}{2}\partial_\mu\phi(x)\partial^\mu\phi(x) - \frac{m^2}{2}\phi(x)^2 + J(x)\phi(x)\right], \tag{9.2}$$

where $J(x)$ is a real function of the coordinates. Its identification with a classical source is obvious once we calculate the equations of motion

$$\left(\nabla^2 + m^2\right)\phi(x) = J(x). \tag{9.3}$$

Our plan is to quantize this theory but, unlike the case analysed in Section 3, now the presence of the source $J(x)$ makes the situation a bit more involved. The general solution to the equations of motion can be written in terms of the retarded Green function for the Klein–Gordon equation as

$$\phi(x) = \phi_0(x) + i\int d^4x'\, G_R(x-x')J(x'), \tag{9.4}$$

where $\phi_0(x)$ is a general solution to the homogeneous equation and

$$G_R(t,\vec{x}) = \int \frac{d^4k}{(2\pi)^4}\frac{i}{k^2-m^2}e^{-ik\cdot x} = i\,\theta(t)\int\frac{d^3k}{(2\pi)^3}\frac{1}{2\omega_k}\left(e^{-i\omega_k t+\vec{k}\cdot\vec{x}} - e^{i\omega_k t - i\vec{p}\cdot\vec{x}}\right), \tag{9.5}$$

with $\theta(x)$ the Heaviside step function. The integration contour to evaluate the integral over p^0 surrounds the poles at $p^0 = \pm\omega_k$ from above. Since $G_R(t,\vec{x}) = 0$ for $t<0$, the function $\phi_0(x)$ corresponds to the solution of the field equation at $t\to-\infty$, before the interaction with the external source[21].

To make the argument simpler we assume that $J(x)$ is switched on at $t=0$, and only lasts for a time τ, that is

$$J(t,\vec{x}) = 0 \qquad \text{if } t<0 \text{ or } t>\tau. \tag{9.6}$$

We are interested in a solution of (9.3) for times after the external source has been switched off, $t>\tau$. In this case the expression (9.5) can be written in terms of the Fourier modes $\widetilde{J}(\omega,\vec{k})$ of the source as

$$\phi(t,\vec{x}) = \phi_0(x) + i\int\frac{d^3k}{(2\pi)^3}\frac{1}{2\omega_k}\left[\widetilde{J}(\omega_k,\vec{k})e^{-i\omega_k t+i\vec{k}\cdot\vec{x}} - \widetilde{J}(\omega_k,\vec{k})^* e^{i\omega_k t - i\vec{k}\cdot\vec{x}}\right]. \tag{9.7}$$

On the other hand, the general solution $\phi_0(x)$ has already been computed in Eq. (3.53). Combining this result with Eq. (9.7) we find the following expression for the late time general solution to the Klein–Gordon equation in the presence of the source

$$\phi(t,x) \;=\; \int\frac{d^3k}{(2\pi)^3}\frac{1}{\sqrt{2\omega_k}}\left\{\left[\alpha(\vec{k}) + \frac{i}{\sqrt{2\omega_k}}\widetilde{J}(\omega_k,\vec{k})\right]e^{-i\omega_k t+i\vec{k}\cdot\vec{x}}\right.$$

[21] We could have taken instead the advanced propagator $G_A(x)$, in which case $\phi_0(x)$ would correspond to the solution to the equation at large times, after the interaction with $J(x)$.

$$+ \quad \left[\alpha^*(\vec{k}) - \frac{i}{\sqrt{2\omega_k}}\widetilde{J}(\omega_k,\vec{k})^*\right] e^{i\omega_k t - i\vec{k}\cdot\vec{x}}\Bigg\}. \tag{9.8}$$

We should not forget that this is a solution valid for times $t > \tau$, i.e., once the external source has been disconnected. On the other hand, for $t < 0$ we find from Eqs. (9.4) and (9.5) that the general solution is given by Eq. (3.53).

Now we can proceed to quantize the theory. The conjugate momentum $\pi(x) = \partial_0\phi(x)$ can be computed from Eqs. (3.53) and (9.8). Imposing the canonical equal time commutation relations (3.50) we find that $\alpha(\vec{k})$, $\alpha^\dagger(\vec{k})$ satisfy the creation–annihilation algebra (3.27). From our previous calculation we find that for $t > \tau$ the expansion of the operator $\phi(x)$ in terms of the creation–annihilation operators $\alpha(\vec{k})$, $\alpha^\dagger(\vec{k})$ can be obtained from the one for $t < 0$ by the replacement

$$\begin{aligned}
\alpha(\vec{k}) \quad &\longrightarrow \quad \beta(\vec{k}) \equiv \alpha(\vec{k}) + \frac{i}{\sqrt{2\omega_k}}\widetilde{J}(\omega_k,\vec{k}), \\
\alpha^\dagger(\vec{k}) \quad &\longrightarrow \quad \beta^\dagger(\vec{k}) \equiv \alpha^\dagger(\vec{k}) - \frac{i}{\sqrt{2\omega_k}}\widetilde{J}(\omega_k,\vec{k})^*.
\end{aligned} \tag{9.9}$$

Actually, since $\widetilde{J}(\omega_k,\vec{k})$ is a c-number, the operators $\beta(\vec{k})$, $\beta^\dagger(\vec{k})$ satisfy the same algebra as $\alpha(\vec{k})$, $\alpha^\dagger(\vec{k})$ and therefore can be interpreted as well as a set of creation–annihilation operators. This means that we can define two vacuum states, $|0_-\rangle$, $|0_+\rangle$ associated with both sets of operators

$$\left.\begin{aligned}
\alpha(\vec{k})|0_-\rangle = 0 \\
\beta(\vec{k})|0_+\rangle = 0
\end{aligned}\right\} \qquad \forall\ \vec{k}. \tag{9.10}$$

For an observer at $t < 0$, $\alpha(\vec{k})$ and $\alpha(\vec{k})$ are the natural set of creation–annihilation operators in terms of which to expand the field operator $\phi(x)$. After the usual zero-point energy subtraction the Hamiltonian is given by

$$\widehat{H}^{(-)} = \int d^3k\, \omega_k\, \alpha^\dagger(\vec{k})\alpha(\vec{k}) \tag{9.11}$$

and the ground state of the spectrum for this observer is the vacuum $|0_-\rangle$. At the same time, a second observer at $t > \tau$ will also see a free scalar quantum field (the source has been switched off at $t = \tau$) and consequently will expand ϕ in terms of the second set of creation–annihilation operators $\beta(\vec{k})$, $\beta^\dagger(\vec{k})$. In terms of these operators the Hamiltonian is written as

$$\widehat{H}^{(+)} = \int d^3k\, \omega_k\, \beta^\dagger(\vec{k})\beta(\vec{k}). \tag{9.12}$$

Then for this late-time observer the ground state of the Hamiltonian is the second vacuum state $|0_+\rangle$.

In our analysis we have been working in the Heisenberg picture, where states are time-independent and the time dependence comes in the operators. Therefore the states of the theory are globally defined. Suppose now that the system is in the 'in' ground state $|0_-\rangle$. An observer at $t < 0$ will find that there are no particles

$$\widehat{n}^{(-)}|0_-\rangle = 0. \tag{9.13}$$

However the late-time observer will find that the state $|0_-\rangle$ contains an average number of particles given by

$$\langle 0_-|\widehat{n}^{(+)}|0_-\rangle = \int \frac{d^3k}{(2\pi)^3}\frac{1}{2\omega_k}\left|\widetilde{J}(\omega_k,\vec{k})\right|^2. \tag{9.14}$$

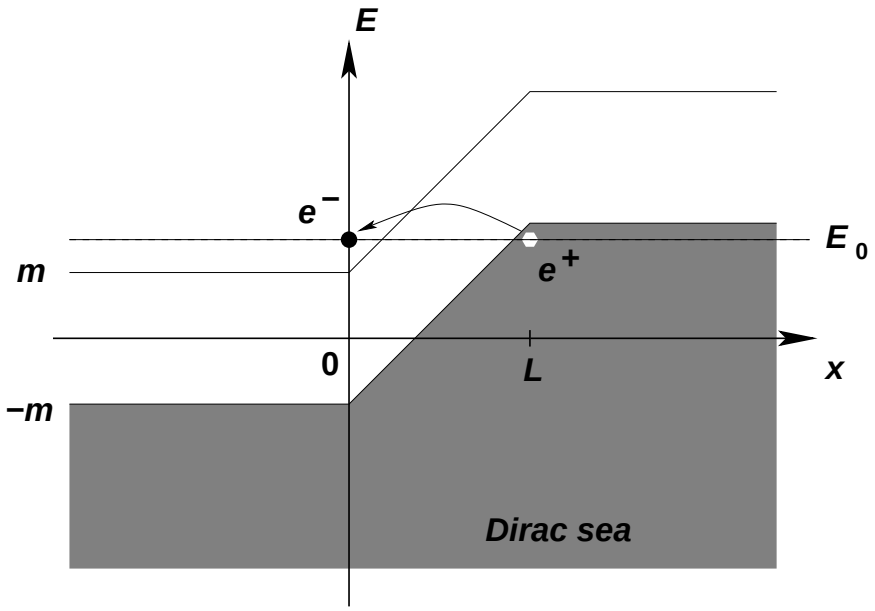


Fig. 18: Pair creation by an electric field in the Dirac sea picture

Moreover, $|0_-\rangle$ is no longer the ground state for the 'out' observer. On the contrary, this state has a vacuum expectation value for $\widehat{H}^{(+)}$

$$\langle 0_-|\widehat{H}^{(+)}|0_-\rangle = \frac{1}{2}\int \frac{d^3k}{(2\pi)^3}\left|\tilde{J}(\omega_k,\vec{k})\right|^2 . \tag{9.15}$$

The key to understand what is going on here lies in the fact that the external source breaks the invariance of the theory under space-time translations. In the particular case we have studied here where $J(x)$ has support over a finite time interval $0 < t < \tau$, this implies that the vacuum is not invariant under time translations, so observers at different times will make different choices of vacuum that will not necessarily agree with each other. This is clear in our example. An observer in $t < \tau$ will choose the vacuum to be the lowest energy state of her Hamiltonian, $|0_-\rangle$. On the other hand, the second observer at late times $t > \tau$ will naturally choose $|0_+\rangle$ as the vacuum. However, for this second observer, the state $|0_-\rangle$ is not the vacuum of his Hamiltonian, but actually an excited state that is a superposition of states with a well-defined number of particles. In this sense it can be said that the external source has the effect of creating particles out of the 'in' vacuum. Besides, this breaking of time translation invariance produces a violation in the energy conservation as we see from Eq. (9.15). Particles are actually created from the energy pumped into the system by the external source.

The Schwinger effect. A classical example of creation of particles by an external field was pointed out by Schwinger [42] and consists of the creation of electron–positron pairs by a strong electric field. In order to illustrate this effect we are going to follow a heuristic argument based on the Dirac sea picture and the WKB approximation.

In the absence of an electric field the vacuum state of a spin-$\frac{1}{2}$ field is constructed by filling all the negative energy states as depicted in Fig. 2. Let us now connect a constant electric field $\vec{\mathcal{E}} = -\mathcal{E}\vec{u}_x$ in the range $0 < x < L$ created by an electrostatic potential

$$V(\vec{r}) = \begin{cases} 0 & x < 0 \\ \mathcal{E}(x - x_0) & 0 < x < L \\ \mathcal{E}L & x > L \end{cases} . \tag{9.16}$$

After the field has been switched on, the Dirac sea looks like that in Fig. 18. In particular we find that if $\mathcal{E}L > 2m$ there are negative energy states at $x > L$ with the same energy as the positive energy states in

the region $x < 0$. Therefore it is possible for an electron filling a negative energy state with energy close to $-2m$ to tunnel through the forbidden region into a positive energy state. The interpretation of such a process is the production of an electron–positron pair out of the electric field.

We can compute the rate at which such pairs are produced by using the WKB approximation. Focusing for simplicity on an electron on top of the Fermi surface near $x = L$ with energy E_0, the transmission coefficient in this approximation is given by[22]

$$\begin{aligned} T_{\rm WKB} &= \exp\left[-2\int_{\frac{1}{e\mathcal{E}}\left(E_0-\sqrt{m^2+\vec{p}_T^{\,2}}\right)}^{\frac{1}{e\mathcal{E}}\left(E_0+\sqrt{m^2+\vec{p}_T^{\,2}}\right)} dx\,\sqrt{m^2-\left[E_0-e\mathcal{E}(x-x_0)\right]^2+\vec{p}_T^{\,2}}\right] \\ &= \exp\left[-\frac{\pi}{e\mathcal{E}}\left(\vec{p}_T^{\,2}+m^2\right)\right], \end{aligned} \tag{9.17}$$

where $p_T^2 \equiv p_y^2 + p_z^2$. This gives the transition probability per unit time and per unit cross section $dydz$ for an electron in the Dirac sea with transverse momentum $\vec{p}_T$ and energy E_0. To get the total probability per unit time and per unit volume we have to integrate over all possible values of $\vec{p}_T$ and E_0. Actually, in the case of the energy, because of the relation between E_0 and the coordinate x at which the particle penetrates into the barrier we can write $\frac{dE_0}{2\pi} = \frac{e\mathcal{E}}{2\pi}dx$ and the total probability per unit time and per unit volume for the creation of a pair is given by

$$W = 2\left(\frac{e\mathcal{E}}{2\pi}\right)\int \frac{d^2p_T}{(2\pi)^2} e^{-\frac{\pi}{e\mathcal{E}}\left(\vec{p}_T^{\,2}+m^2\right)} = \frac{e^2\mathcal{E}^2}{4\pi^3} e^{-\frac{\pi\, m^2}{e\mathcal{E}}}, \tag{9.18}$$

where the factor of 2 accounts for the two polarizations of the electron.

Then production of electron–positron pairs is exponentially suppressed and is only sizeable for strong electric fields. To estimate its order of magnitude it is useful to restore the powers of c and $\hbar$ in (9.18)

$$W = \frac{e^2\mathcal{E}^2}{4\pi^3 c\hbar^2} e^{-\frac{\pi\, m^2 c^3}{\hbar e\mathcal{E}}} \tag{9.19}$$

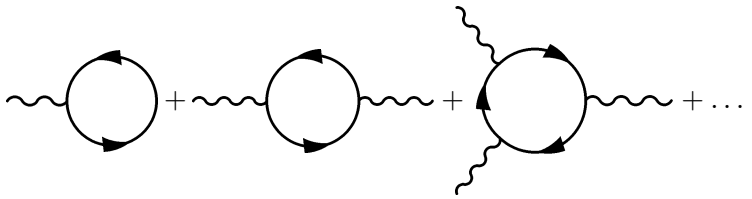
The exponential suppression of the pair production disappears when the electric field reaches the critical value $\mathcal{E}_{\rm crit}$ at which the exponent is of order one

$$\mathcal{E}_{\rm crit} = \frac{m^2c^3}{\hbar e} \simeq 1.3\times 10^{16}\,{\rm V\,cm^{-1}}. \tag{9.20}$$

This is indeed a very strong field which is extremely difficult to produce. A similar effect, however, takes place also in a time-varying electric field [43] and there is the hope that pair production could be observed in the presence of the alternating electric field produced by a laser.

The heuristic derivation that we followed here can be made more precise in QED. There the decay of the vacuum into electron–positron pairs can be computed from the imaginary part of the effective action $\Gamma[A_\mu]$ in the presence of a classical gauge potential A_μ

$$\begin{aligned} i\Gamma[A_\mu] &\equiv \\ &= \log\,\det\left[1 - ie\not{A}\frac{1}{i\not{\partial}-m}\right]. \end{aligned} \tag{9.21}$$



[22] Notice that the electron satisfies the relativistic dispersion relation $E = \sqrt{\vec{p}^{\,2}+m^2}+V$ and therefore $-p_x^2 = m^2 - (E-V)^2 + \vec{p}_T^{\,2}$. The integration limits are set by those values of x at which $p_x = 0$.

This determinant can be computed using the standard heat kernel techniques. The probability of pair production is proportional to the imaginary part of $i\Gamma[A_\mu]$ and gives

$$W = \frac{e^2\mathcal{E}^2}{4\pi^3}\sum_{n=1}^{\infty}\frac{1}{n^2}e^{-n\frac{\pi m^2}{e\mathcal{E}}}. \tag{9.22}$$

Our simple argument based on tunneling in the Dirac sea gave only the leading term of Schwinger's result (9.22). The remaining terms can be also captured in the WKB approximation by taking into account the probability of production of several pairs, i.e., the tunneling of more than one electron through the barrier.

Here we have illustrated the creation of particles by semiclassical sources in quantum field theory using simple examples. Nevertheless, what we learned has important applications to the study of quantum fields in curved backgrounds. In quantum field theory in Minkowski space-time the vacuum state is invariant under the Poincaré group and this, together with the covariance of the theory under Lorentz transformations, implies that all inertial observers agree on the number of particles contained in a quantum state. The breaking of such invariance, as happened in the case of coupling to a time-varying source analysed above, implies that it is not possible anymore to define a state which would be recognized as the vacuum by all observers.

This is precisely the situation when fields are quantized on curved backgrounds. In particular, if the background is time-dependent (as happens in a cosmological setup or for a collapsing star) different observers will identify different vacuum states. As a consequence what one observer calls the vacuum will be full of particles for a different observer. This is precisely what is behind the phenomenon of Hawking radiation [44]. The emission of particles by a physical black hole formed from gravitational collapse of a star is the consequence of the fact that the vacuum state in the asymptotic past contains particles for an observer in the asymptotic future. As a consequence, a detector located far away from the black hole detects a stream of thermal radiation with temperature

$$T_{\text{Hawking}} = \frac{\hbar c^3}{8\pi G_N\, k\, M} \tag{9.23}$$

where M is the mass of the black hole, G_N is Newton's constant and k is Boltzmann's constant. There are several ways in which these results can be obtained. A more heuristic way is perhaps to think of this particle creation as resulting from quantum tunneling of particles across the potential barrier posed by gravity [45].

9.2 Supersymmetry

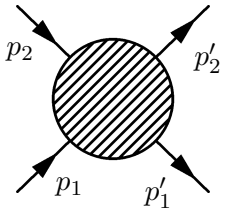
One of the things that we have learned in our journey around the landscape of quantum field theory is that our knowledge of the fundamental interactions in Nature is based on the idea of symmetry, and in particular gauge symmetry. The Lagrangian of the Standard Model can be written just including all possible renormalizable terms (i.e., with canonical dimension smaller o equal to 4) compatible with the gauge symmetry SU(3)×SU(2)×U(1) and Poincaré invariance. All attempts to go beyond start with the question of how to extend the symmetries of the Standard Model.

As explained in Section 5.1, in a quantum field theoretical description of the interaction of elementary particles the basic observable quantity to compute is the scattering or S-matrix giving the probability amplitude for the scattering of a number of incoming particles with a certain momentum into some final products

$$\mathcal{A}(\text{in} \longrightarrow \text{out}) = \langle \vec{p}_1{}', \ldots; \text{out}|\vec{p}_1, \ldots; \text{in}\rangle. \tag{9.24}$$

An explicit symmetry of the theory has to be necessarily a symmetry of the S-matrix. Hence it is fair to ask what is the largest symmetry of the S-matrix.

Let us ask this question in the simple case of the scattering of two particles with four-momenta p_1 and p_2 in the t-channel



We will make the usual assumptions regarding positivity of the energy and analyticity. Invariance of the theory under the Poincaré group implies that the amplitude can only depend on the scattering angle ϑ through

$$t = (p_1' - p_1)^2 = 2\left(m_1^2 - p_1 \cdot p_1'\right) = 2\left(m_1^2 - E_1 E_1' + |\vec{p}_1||\vec{p}_1{}'|\cos\vartheta\right). \tag{9.25}$$

If there would be any extra bosonic symmetry of the theory it would restrict the scattering angle to a set of discrete values. In this case the S-matrix cannot be analytic since it would vanish everywhere except for the discrete values selected by the extra symmetry.

Actually, the only way to extend the symmetry of the theory without renouncing the analyticity of the scattering amplitudes is to introduce 'fermionic' symmetries, i.e., symmetries whose generators are anticommuting objects [46]. This means that in addition to the generators of the Poincaré group[23] P^μ, $M^{\mu\nu}$ and the ones for the internal gauge symmetries G, we can introduce a number of fermionic generators Q_a^I, $\overline{Q}_{\dot{a}\,I}$ ($I = 1, \ldots, \mathcal{N}$), where $\overline{Q}_{\dot{a}\,I} = (Q_a^I)^\dagger$. The most general algebra that these generators satisfy is the $\mathcal{N}$-extended supersymmetry algebra [47]

$$\begin{aligned}
\{Q_a^I, \overline{Q}_{\dot{b}\,J}\} &= 2\sigma^\mu_{a\dot{b}} P_\mu \delta^I{}_J, \\
\{Q_a^I, Q_b^J\} &= 2\varepsilon_{ab}\mathcal{Z}^{IJ}, \qquad (9.26)\\
\{\overline{Q}_{\dot{a}}^I, \overline{Q}_{\dot{b}}^J\} &= -2\varepsilon_{\dot{a}\dot{b}}\overline{\mathcal{Z}}^{IJ}, \qquad (9.27)
\end{aligned}$$

where $\mathcal{Z}^{IJ} \in \mathbb{C}$ commute with any other generator and satisfy $\mathcal{Z}^{IJ} = -\mathcal{Z}^{JI}$. Besides we have the commutators that determine the Poincaré transformations of the fermionic generators Q_a^I, $Q_{\dot{a}\,J}$

$$\begin{aligned}
[Q_a^I, P^\mu] &= [\overline{Q}_{\dot{a}\,I}, P^\mu] = 0, \\
[Q_a^I, M^{\mu\nu}] &= \frac{1}{2}(\sigma^{\mu\nu})_a{}^b Q_b^I, \qquad (9.28)\\
[\overline{Q}_{a\,I}, M^{\mu\nu}] &= -\frac{1}{2}(\overline{\sigma}^{\mu\nu})_{\dot{a}}{}^{\dot{b}}\,\overline{Q}_{\dot{b}\,I},
\end{aligned}$$

where $\sigma^{0i} = -i\sigma^i$, $\sigma^{ij} = \varepsilon^{ijk}\sigma^k$ and $\overline{\sigma}^{\mu\nu} = (\sigma^{\mu\nu})^\dagger$. These identities simply mean that Q_a^I, $\overline{Q}_{\dot{a}\,J}$ transform respectively in the $(\frac{\mathbf{1}}{\mathbf{2}}, \mathbf{0})$ and $(\mathbf{0}, \frac{\mathbf{1}}{\mathbf{2}})$ representations of the Lorentz group.

We know that the presence of a global symmetry in a theory implies that the spectrum can be classified in multiplets with respect to that symmetry. In the case of supersymmetry start with the case case $\mathcal{N} = 1$ in which there is a single pair of supercharges Q_a, $\overline{Q}_{\dot{a}}$ satisfying the algebra

$$\{Q_a, \overline{Q}_{\dot{b}}\} = 2\sigma^\mu_{a\dot{b}} P_\mu, \qquad \{Q_a, Q_b\} = \{\overline{Q}_{\dot{a}}, \overline{Q}_{\dot{b}}\} = 0. \tag{9.29}$$

Notice that in the $\mathcal{N} = 1$ case there is no possibility of having central charges.

We study now the representations of the supersymmetry algebra (9.29), starting with the massless case. Given a state $|k\rangle$ satisfying $k^2 = 0$, we can always find a reference frame where the four-vector k^μ

[23] The generators $M^{\mu\nu}$ are related with the ones for boost and rotations introduced in Section 4.1 by $J^i \equiv M^{0i}$, $M^i = \frac{1}{2}\varepsilon^{ijk}M^{jk}$. In this section we also use the 'dotted spinor' notation, in which spinors in the $(\frac{\mathbf{1}}{\mathbf{2}}, \mathbf{0})$ and $(\mathbf{0}, \frac{\mathbf{1}}{\mathbf{2}})$ representations of the Lorentz group are indicated respectively by undotted $(a, b, \ldots)$ and dotted $(\dot{a}, \dot{b}, \ldots)$ indices.

takes the form $k^\mu = (E, 0, 0, E)$. Since the theory is Lorentz covariant we can obtain the representation of the supersymmetry algebra in this frame where the expressions are simpler. In particular, the right-hand side of the first anticommutator in Eq. (9.29) is given by

$$2\sigma^\mu_{a\dot{b}} P_\mu = 2(P^0 - \sigma^3 P^3) = \begin{pmatrix} 0 & 0 \\ 0 & 4E \end{pmatrix}. \tag{9.30}$$

Therefore the algebra of supercharges in the massless case reduces to

$$\begin{aligned} \{Q_1, Q_1^\dagger\} &= \{Q_1, Q_2^\dagger\} = 0, \\ \{Q_2, Q_2^\dagger\} &= 4E. \end{aligned} \tag{9.31}$$

The commutator $\{Q_1, Q_1^\dagger\} = 0$ implies that the action of Q_1 on any state gives a zero-norm state of the Hilbert space $\|Q_1|\Psi\rangle\| = 0$. If we want the theory to preserve unitarity we must eliminate these null states from the spectrum. This is equivalent to setting $Q_1 \equiv 0$. On the other hand, in terms of the second generator Q_2 we can define the operators

$$a = \frac{1}{2\sqrt{E}} Q_2, \qquad a^\dagger = \frac{1}{2\sqrt{E}} Q_2^\dagger, \tag{9.32}$$

which satisfy the algebra of a pair of fermionic creation–annihilation operators, $\{a, a^\dagger\} = 1$, $a^2 = (a^\dagger)^2 = 0$. Starting with a vacuum state $a|\lambda\rangle = 0$ with helicity λ we can build the massless multiplet

$$|\lambda\rangle, \qquad |\lambda + \tfrac{1}{2}\rangle \equiv a^\dagger|\lambda\rangle. \tag{9.33}$$

Here we consider two important cases:

- Scalar multiplet: we take the vacuum state to have zero helicity $|0^+\rangle$ so the multiplet consists of a scalar and a helicity-$\frac{1}{2}$ state

  $$|0^+\rangle, \qquad |\,\tfrac{1}{2}\rangle \equiv a^\dagger|0^+\rangle. \tag{9.34}$$

  However, this multiplet is not invariant under the CPT transformation which reverses the sign of the helicity of the states. In order to have a CPT-invariant theory we have to add to this multiplet its CPT-conjugate which can be obtained from a vacuum state with helicity $\lambda = -\frac{1}{2}$

  $$|0^-\rangle, \qquad |-\tfrac{1}{2}\rangle. \tag{9.35}$$

  Putting them together we can combine the two zero-helicity states with the two fermionic ones into the degrees of freedom of a complex scalar field and a Weyl (or Majorana) spinor.
- Vector multiplet: now we take the vacuum state to have helicity $\lambda = \frac{1}{2}$, so the multiplet contains also a massless state with helicity $\lambda = 1$

  $$|\,\tfrac{1}{2}\rangle, \qquad |1\rangle \equiv a^\dagger|\,\tfrac{1}{2}\rangle. \tag{9.36}$$

  As with the scalar multiplet we add the CPT conjugate obtained from a vacuum state with helicity $\lambda = -1$

  $$|-\tfrac{1}{2}\rangle, \qquad |-1\rangle, \tag{9.37}$$

  which together with (9.36) give the propagating states of a gauge field and a spin-$\frac{1}{2}$ gaugino.

In both cases we see the trademark of supersymmetric theories: the number of bosonic and fermionic states within a multiplet are the same.

In the case of extended supersymmetry we have to repeat the previous analysis for each supersymmetry charge. At the end, we have $\mathcal{N}$ sets of fermionic creation–annihilation operators $\{a^I, a^\dagger_J\} = \delta^I{}_J$, $(a_I)^2 = (a^\dagger_I)^2 = 0$. Let us work out the case of $\mathcal{N} = 8$ supersymmetry. Since for several reasons we do not want to have states with helicity larger than 2, we start with a vacuum state $|-2\rangle$ of helicity $\lambda = -2$. The rest of the states of the supermultiplet are obtained by applying the eight different creation operators $a^\dagger_I$ to the vacuum:

$$\begin{aligned}
\lambda = 2: &\quad a^\dagger_1 \dots a^\dagger_8 |-2\rangle && \binom{8}{8} = 1 \text{ state},\\
\lambda = \frac{3}{2}: &\quad a^\dagger_{I_1} \dots a^\dagger_{I_7} |-2\rangle && \binom{8}{7} = 8 \text{ states},\\
\lambda = 1: &\quad a^\dagger_{I_1} \dots a^\dagger_{I_6} |-2\rangle && \binom{8}{6} = 28 \text{ states},\\
\lambda = \frac{1}{2}: &\quad a^\dagger_{I_1} \dots a^\dagger_{I_5} |-2\rangle && \binom{8}{5} = 56 \text{ states},\\
\lambda = 0: &\quad a^\dagger_{I_1} \dots a^\dagger_{I_4} |-2\rangle && \binom{8}{4} = 70 \text{ states},\\
\lambda = -\frac{1}{2}: &\quad a^\dagger_{I_1} a^\dagger_{I_2} a^\dagger_{I_3} |-2\rangle && \binom{8}{3} = 56 \text{ states},\\
\lambda = -1: &\quad a^\dagger_{I_1} a^\dagger_{I_2} |-2\rangle && \binom{8}{2} = 28 \text{ states},\\
\lambda = -\frac{3}{2}: &\quad a^\dagger_{I_1} |-2\rangle && \binom{8}{1} = 8 \text{ states},\\
\lambda = -2: &\quad |-2\rangle && 1 \text{ state}.
\end{aligned} \tag{9.38}$$

Putting together the states with opposite helicity we find that the theory contains:

- 1 spin-2 field $g_{\mu\nu}$ (a graviton),
- 8 spin-$\frac{3}{2}$ gravitino fields ψ^I_μ,
- 28 gauge fields $A^{[IJ]}_\mu$,
- 56 spin-$\frac{1}{2}$ fermions $\psi^{[IJK]}$,
- 70 scalars $\phi^{[IJKL]}$,

where by $[IJ...]$ we have denoted that the indices are antisymmetrized. We see that, unlike the massless multiplets of $\mathcal{N} = 1$ supersymmetry studied above, this multiplet is CPT invariant by itself. As in the case of the massless $\mathcal{N} = 1$ multiplet, here we also find as many bosonic as fermionic states:

$$\begin{aligned}
\text{bosons:} &\quad 1 + 28 + 70 + 28 + 1 = 128 \quad \text{states},\\
\text{fermions:} &\quad 8 + 56 + 56 + 8 = 128 \quad \text{states}.
\end{aligned}$$

Now we study briefly the case of massive representations $|k\rangle$, $k^2 = M^2$. Things become simpler if we work in the rest frame where $P^0 = M$ and the spatial components of the momentum vanish. Then, the supersymmetry algebra becomes:

$$\{Q^I_\alpha, \overline{Q}_{\dot\beta\, J}\} = 2M\delta_{\alpha\dot\beta}\delta^I{}_J. \tag{9.39}$$

We proceed now in a similar way to the massless case by defining the operators

$$a^I_\alpha \equiv \frac{1}{\sqrt{2M}} Q^I_\alpha, \qquad a^\dagger_{\dot\alpha\, I} \equiv \frac{1}{\sqrt{2M}} \overline{Q}_{\dot\alpha\, I}. \tag{9.40}$$

The multiplets are found by choosing a vacuum state with a definite spin. For example, for $\mathcal{N} = 1$ and taking a spin-0 vacuum $|0\rangle$ we find three states in the multiplet transforming irreducibly with respect to the Lorentz group:

$$|0\rangle, \qquad a^\dagger_{\dot\alpha}|0\rangle, \qquad \varepsilon^{\dot\alpha\dot\beta} a^\dagger_{\dot\alpha} a^\dagger_{\dot\beta}|0\rangle, \tag{9.41}$$

which, once transformed back from the rest frame, correspond to the physical states of two spin-0 bosons and one spin-$\frac{1}{2}$ fermion. For $\mathcal{N}$-extended supersymmetry the corresponding multiplets can be worked out in a similar way.

The equality between bosonic and fermionic degrees of freedom is at the root of many of the interesting properties of supersymmetric theories. For example, in Section 4 we computed the divergent vacuum energy contributions for each real bosonic or fermionic propagating degree of freedom as[24]

$$E_{\rm vac} = \pm \frac{1}{2} \delta(\vec{0}) \int d^3p\, \omega_p, \tag{9.42}$$

where the sign $\pm$ corresponds respectively to bosons and fermions. Hence, for a supersymmetric theory the vacuum energy contribution exactly cancels between bosons and fermions. This boson–fermion degeneracy is also responsible for supersymmetric quantum field theories being less divergent than non-supersymmetric ones.

Acknowledgements

It is a great pleasure to thank the organizers of the CERN-CLAF 2005 and 2009 Schools, and in particular Teresa Dova, Marta Losada and Enrico Nardi, for the opportunity to present this material, and for the wonderful atmosphere they created during the Schools. The work of M.A.V.-M. has been partially supported by Spanish Science Ministry Grants FPA2002-02037, FPA2005-04823, and BFM2003-02121.

Appendix: A crash course in group theory

In this Appendix we summarize some basic facts about group theory. Given a group G a representation of G is a correspondence between the elements of G and the set of linear operators acting on a vector space V, such that for each element of the group $g \in G$ there is a linear operator $D(g)$

$$D(g) : V \longrightarrow V \tag{A.43}$$

satisfying the group operations

$$D(g_1)D(g_2) = D(g_1 g_2), \qquad D(g_1^{-1}) = D(g_1)^{-1}, \qquad g_1, g_2 \in \mathcal{G}. \tag{A.44}$$

The representation $D(g)$ is irreducible if and only if the only operators $A : V \to V$ commuting with all the elements of the representation $D(g)$ are the ones proportional to the identity

$$[D(g), A] = 0, \ \ \forall g \qquad \Longleftrightarrow \qquad A = \lambda \mathbf{1}, \quad \lambda \in \mathbb{C}\,. \tag{A.45}$$

More intuitively, we can say that a representation is irreducible if there is no proper subspace $U \subset V$ (i.e., $U \neq V$ and $U \neq \emptyset$) such that $D(g)U \subset U$ for every element $g \in G$.

[24] For a boson, this can be read off Eq. (3.56). In the case of fermions, the result of Eq. (4.44) gives the vacuum energy contribution of the four real propagating degrees of freedom of a Dirac spinor.

Here we are specially interested in Lie groups whose elements are labelled by a number of continuous parameters. In mathematical terms this means that a Lie group is a manifold $\mathcal{M}$ together with an operation $\mathcal{M} \times \mathcal{M} \longrightarrow \mathcal{M}$ that we will call multiplication that satisfies the associativity property $g_1 \cdot (g_2 \cdot g_3) = (g_1 \cdot g_2) \cdot g_3$ together with the existence of unity $g\mathbf{1} = \mathbf{1}g = g$, for every $g \in \mathcal{M}$ and inverse $gg^{-1} = g^{-1}g = \mathbf{1}$.

The simplest example of a Lie group is SO(2), the group of rotations in the plane. Each element $R(\theta)$ is labelled by the rotation angle θ, with the multiplication acting as $R(\theta_1)R(\theta_2) = R(\theta_1 + \theta_2)$. Because the angle θ is defined only modulo 2π, the manifold of SO(2) is a circumference S^1.

One of the interesting properties of Lie groups is that in a neighborhood of the identity element they can be expressed in terms of a set of generators T^a ($a = 1, \ldots, \dim G$) as

$$D(g) = \exp[-i\alpha_a T^a] \equiv \sum_{n=0}^{\infty} \frac{(-i)^n}{n!} \alpha_{a_1} \ldots \alpha_{a_n} T^{a_1} \ldots T^{a_n}, \tag{A.46}$$

where $\alpha_a \in \mathbb{C}$ are a set of coordinates of $\mathcal{M}$ in a neighborhood of $\mathbf{1}$. Because of the general Baker–Campbell–Haussdorf formula, the multiplication of two group elements is encoded in the value of the commutator of two generators, that in general has the form

$$[T^a, T^b] = if^{abc}T^c, \tag{A.47}$$

where $f^{abc} \in \mathbb{C}$ are called the structure constants. The set of generators with the commutator operation form the Lie algebra associated with the Lie group. Hence, given a representation of the Lie algebra of generators we can construct a representation of the group by exponentiation (at least locally near the identity).

We illustrate this concept with some particular examples. For SU(2) each group element is labelled by three real number α_i, $i = 1, 2, 3$. We have two basic representations: one is the fundamental representation (or spin $\frac{1}{2}$) defined by

$$D_{\frac{1}{2}}(\alpha_i) = e^{-\frac{i}{2}\alpha_i \sigma^i}, \tag{A.48}$$

with σ^i the Pauli matrices. The second one is the adjoint (or spin 1) representation which can be written as

$$D_1(\alpha_i) = e^{-i\alpha_i J^i}, \tag{A.49}$$

where

$$J^1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \qquad J^2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad J^3 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{A.50}$$

Actually, J^i ($i = 1, 2, 3$) generate rotations around the x, y and z axis respectively. Representations of spin $j \in \mathbb{N} + \frac{1}{2}$ can also be constructed with dimension

$$\dim D_j(g) = 2j + 1. \tag{A.51}$$

As a second example we consider SU(3). This group has two basic three-dimensional representations denoted by $\mathbf{3}$ and $\overline{\mathbf{3}}$ which in QCD are associated with the transformation of quarks and antiquarks under the color gauge symmetry SU(3). The elements of these representations can be written as

$$D_{\mathbf{3}}(\alpha^a) = e^{\frac{i}{2}\alpha^a \lambda_a}, \qquad D_{\overline{\mathbf{3}}}(\alpha^a) = e^{-\frac{i}{2}\alpha^a \lambda_a^T} \qquad (a = 1, \ldots, 8), \tag{A.52}$$

where λ_a are the eight hermitian Gell-Mann matrices

$$\begin{aligned}
\lambda_1 &= \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \lambda_2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\
\lambda_4 &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \lambda_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, \qquad \lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \\
\lambda_7 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \qquad \lambda_8 = \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{3}} & 0 \\ 0 & 0 & -\frac{2}{\sqrt{3}} \end{pmatrix}.
\end{aligned} \tag{A.53}$$

Hence the generators of the representations $\mathbf{3}$ and $\overline{\mathbf{3}}$ are given by

$$T^a(\mathbf{3}) = \frac{1}{2}\lambda_a, \qquad T^a(\overline{\mathbf{3}}) = -\frac{1}{2}\lambda_a^T. \tag{A.54}$$

Irreducible representations can be classified in three groups: real, complex and pseudoreal.

– Real representations: a representation is said to be real if there is a *symmetric matrix* S which acts as intertwiner between the generators and their complex conjugates

$$\overline{T}^a = -ST^aS^{-1}, \qquad S^T = S. \tag{A.55}$$

This is for example the case of the adjoint representation of SU(2) generated by the matrices (A.50)

– Pseudoreal representations: are the ones for which an *antisymmetric matrix* S exists with the property

$$\overline{T}^a = -ST^aS^{-1}, \qquad S^T = -S. \tag{A.56}$$

As an example we can mention the spin-$\frac{1}{2}$ representation of SU(2) generated by $\frac{1}{2}\sigma^i$.

– Complex representations: finally, a representation is complex if the generators and their complex conjugate are not related by a similarity transformation. This is for instance the case of the two three-dimensional representations $\mathbf{3}$ and $\overline{\mathbf{3}}$ of SU(3).

There are a number of invariants that can be constructed associated with an irreducible representation R of a Lie group G and that can be used to label such a representation. If T_R^a are the generators in a certain representation R of the Lie algebra, it is easy to see that the matrix $\sum_{a=1}^{\dim G} T_R^a T_R^a$ commutes with every generator T_R^a. Therefore, because of Schur's lemma, it has to be proportional to the identity[25]. This defines the Casimir invariant $C_2(R)$ as

$$\sum_{a=1}^{\dim G} T_R^a T_R^a = C_2(R)\mathbf{1}. \tag{A.57}$$

A second invariant $T_2(R)$ associated with a representation R can also be defined by the identity

$$\mathrm{Tr}\, T_R^a T_R^b = T_2(R)\delta^{ab}. \tag{A.58}$$

[25]Schur's lemma states that a representation of a group is irreducible if and only if all matrices commuting with every element of the representation are proportional to the identity.

Actually, taking the trace in Eq. (A.57) and combining the result with (A.58) we find that both invariants are related by the identity

$$C_2(R)\dim R = T_2(R)\dim G, \tag{A.59}$$

with $\dim R$ the dimension of the representation R.

These two invariants appear frequently in quantum field theory calculations with non-Abelian gauge fields. For example $T_2(R)$ comes about as the coefficient of the one-loop calculation of the beta-function for a Yang–Mills theory with gauge group G. In the case of SU(N), for the fundamental representation, we find the values

$$C_2(\text{fund}) = \frac{N^2-1}{2N}, \qquad T_2(\text{fund}) = \frac{1}{2}, \tag{A.60}$$

whereas for the adjoint representation the results are

$$C_2(\text{adj}) = N, \qquad T_2(\text{adj}) = N. \tag{A.61}$$

A third invariant $A(R)$ is specially important in the calculation of anomalies. As discussed in Section 7, the chiral anomaly in gauge theories is proportional to the group-theoretical factor $\text{Tr}\,\left[T_R^a\{T_R^b, T_R^c\}\right]$. This leads us to define $A(R)$ as

$$\text{Tr}\,\left[T_R^a\{T_R^b, T_R^c\}\right] = A(R), d^{abc} \tag{A.62}$$

where d^{abc} is symmetric in its three indices and does not depend on the representation. Therefore, the cancellation of anomalies in a gauge theory with fermions transformed in the representation R of the gauge group is guaranteed if the corresponding invariant $A(R)$ vanishes.

It is not difficult to prove that $A(R) = 0$ if the representation R is either real or pseudoreal. Indeed, if this is the case, then there is a matrix S (symmetric or antisymmetric) that intertwines the generators T_R^a and their complex conjugates $\overline{T}_R^a = -ST_R^aS^{-1}$. Then, using the hermiticity of the generators we can write

$$\text{Tr}\,\left[T_R^a\{T_R^b, T_R^c\}\right] = \text{Tr}\,\left[T_R^a\{T_R^b, T_R^c\}\right]^T = \text{Tr}\,\left[\overline{T}_R^a\{\overline{T}_R^b, \overline{T}_R^c\}\right]. \tag{A.63}$$

Now, using (A.55) or (A.56) we have

$$\text{Tr}\,\left[\overline{T}_R^a\{\overline{T}_R^b, \overline{T}_R^c\}\right] = -\text{Tr}\,\left[ST_R^aS^{-1}\{ST_R^bS^{-1}, ST_R^cS^{-1}\}\right] = -\text{Tr}\,\left[T_R^a\{T_R^b, T_R^c\}\right], \tag{A.64}$$

which proves that $\text{Tr}\,\left[T_R^a\{T_R^b, T_R^c\}\right]$ and therefore $A(R) = 0$ whenever the representation is real or pseudoreal. Since the gauge anomaly in four dimensions is proportional to $A(R)$ this means that anomalies appear only when the fermions transform in a complex representation of the gauge group.

References

[1] J. D. Bjorken and S. D. Drell, *Relativistic Quantum Fields* (McGraw-Hill, New York, 1965).

[2] C. Itzykson and J.-B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980).

[3] P. Ramond, *Field Theory: A Modern Primer*, 2nd ed. (Addison-Wesley, Redwood City, CA, 1989).

[4] M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory* (Addison-Wesley, Reading, MA, 1995).

[5] S. Weinberg, *The Quantum Theory of Fields*, 3 vols. (Cambridge University Press, 1995).

[6] P. Deligne et al. (editors), *Quantum Fields and Strings: A Course for Mathematicians*, 2 vols. (American Mathematical Society, Providence, RI, 1999–2000).

[7] A. Zee, *Quantum Field Theory in a Nutshell* (Princeton University Press, 2003).

[8] B. S. DeWitt, *The Global Approach to Quantum Field Theory*, 2 vols. (Clarendon Press, Oxford, 2003).

[9] V. P. Nair, *Quantum Field Theory: A Modern Perspective* (Springer, 2005).

[10] T. Banks, *Modern Quantum Field Theory* (Cambridge University Press, 2008).

[11] O. Klein, *Die Reflexion von Elektronen an einem Potentialsprung nach der Relativischen Dynamik von Dirac*, Z. Phys. **53** (1929) 157.

[12] B. R. Holstein, *Klein's paradox*, Am. J. Phys. **66** (1998) 507.

[13] N. Dombey and A. Calogeracos, *Seventy years of the Klein paradox*, Phys. Rep. **315** (1999) 41.
N. Dombey and A. Calogeracos, *History and physics of the Klein paradox*, Contemp. Phys. **40** (1999) 313 (`quant-ph/9905076`).

[14] F. Sauter, *Zum Kleinschen Paradoxon*, Z. Phys. **73** (1932) 547.

[15] H. B. G. Casimir, *On the attraction between two perfectly conducting plates*, Proc. Kon. Ned. Akad. Wet. **60** (1948) 793.

[16] G. Plunien, B. Müller and W. Greiner, *The Casimir effect*, Phys. Rep. **134** (1986) 87.
K. A. Milton, *The Casimir Effect: Physical Manifestation of Zero-Point Energy*, (`hep-th/9901011`).
K. A. Milton, *The Casimir effect: recent controversies and progress*, J. Phys. **A37** (2004) R209 (`hep-th/0406024`).
S. K. Lamoreaux, *The Casimir force: background, experiments, and applications*, Rep. Prog. Phys. **68** (2005) 201.

[17] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1972).

[18] M. J. Sparnaay, *Measurement of attractive forces between flat plates*, Physica **24** (1958) 751.

[19] Y. Aharonov and D. Bohm, *Significance of the electromagnetic potentials in the quantum theory*, Phys. Rev. **115** (1955) 485.

[20] P. A. M. Dirac, *Quantised singularities in the electromagneticfield*, Proc. R. Soc. **133** (1931) 60.

[21] P. A. M. Dirac, *Lectures on Quantum Mechanics* (Dover, 2001).

[22] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton University Press, 1992).

[23] R. Jackiw, *Quantum meaning of classical field theory*, Rev. Mod. Phys. **49** (1977) 681.
R. Jackiw, *Introduction to the Yang–Mills quantum theory*, Rev. Mod. Phys. **52** (1980) 661.

[24] P. Ramond, *Journeys Beyond the Standard Model* (Perseus Books, Cambridge, MA, 1999).
R. N. Mohapatra, *Unification and Supersymmetry: The Frontiers of Quark–Lepton Physics*, 3rd ed. (Springer, Berlin, 2003).

[25] S. Dodelson, *Modern Cosmology* (Academic Press, New York, 2003).

[26] L. Álvarez-Gaumé, An introduction to anomalies, in: *Fundamental Problems of Gauge Field Theory*, Eds. G. Velo and A. S. Wightman (Plenum Press, New York, 1986).

[27] R. Jackiw, Topological investigations of quantized gauge theories, in: *Current Algebra and Anomalies*, Eds. S. B. Treiman, R. Jackiw, B. Zumino, and E. Witten (World Scientific, Singapore, 1986).

[28] S. Adler, *Axial-vector vertex in spinor electrodynamics*, Phys. Rev. **177** (1969) 2426.
J. S. Bell and R. Jackiw, *A PCAC puzzle:* $\pi^0 \to 2\gamma$ *in the sigma model*, Nuovo Cimento **A60** (1969) 47.

[29] F. J. Ynduráin, *The Theory of Quark and Gluon Interactions*, 3rd ed. (Springer, Berlin, 1999).

[30] G. 't Hooft, *How the instantons solve the U(1) problem*, Phys. Rep. **142** (1986) 357.

[31] D. G. Sutherland, *Current algebra and some nonstrong mesonic decays*, Nucl. Phys. **B2** (1967) 433.

M. J. G. Veltman, *Theoretical aspects of high-energy neutrino interactions*, Proc. R. Soc. **A301** (1967) 107.

[32] J. Steinberger, *On the use of substraction fields and the lifetimes of some types of meson decay*, Phys. Rev. **76** (1949) 1180.

[33] S. L. Adler and W. A. Bardeen, *Absence of higher order corrections in the anomalous axial vector divergence equation*, Phys. Rev. **182** (1969) 1517.

[34] E. Witten, *An SU(2) anomaly*, Phys. Lett. **B117** (1982) 324.

[35] S. Eidelman et al., *Review of particle physics*, Phys. Lett. **B592** (2004) 1 (`http://pdg.lbl.gov`).

[36] D. J. Gross and F. Wilczek, *Ultraviolet behavior of non-Abelian gauge theories*, Phys. Rev. Lett. **30** (1973) 1343.

[37] H. D. Politzer, *Reliable perturbative results for strong interations?*, Phys. Rev. Lett. **30** (1973) 1346.

[38] G. 't Hooft, remarks at the *Colloquium on Renormalization of Yang–Mills Fields and Applications to Particle Physics*, Marseille, 1972 (CNRS, Marseille, 1972).

[39] I. B. Khriplovich, *Green's functions in theories with a non-Abelian gauge group*, Yad. Fiz. **10** (1969) 409 [Sov. J. Nucl. Phys. **10** (1970) 235].
M. V. Terentiev and V. S. Vanyashin, *The vacuum polarization of a charged vector field*, Zh. Eskp. Teor. Fiz. **48** (1965) 565 [Sov. Phys. JETP **21** (1965) 375].

[40] K. G. Wilson, *Renormalization group and critical phenomena 1. Renormalization group and the Kadanoff scaling picture*, Phys. Rev. **B4** (1971) 3174.
K. G. Wilson, *Renormalization group and critical phenomena 2. Phase space cell analysis of critical behavior*, Phys. Rev. **B4** (1971) 3184
K. G. Wilson, *The renormalization group and critical phenomena*, Rev. Mod. Phys. **55** (1983) 583.

[41] L. P. Kadanoff, *Scaling laws for Ising models near T_c*, Physics **2** (1966) 263.

[42] J. Schwinger, *On gauge invariance and vacuum polarization*, Phys. Rev. **82** (1951) 664.

[43] E. Brezin and C. Itzykson, *Pair production in vacuum by an alternating field*, Phys. Rev. **D2** (1970) 1191.

[44] S. W. Hawking, *Particle creation by black holes*, Commun. Math. Phys. **43** (1975) 199.

[45] M. K. Parikh and F. Wilczek, *Hawking radiation as tunneling*, Phys. Rev. Lett. **85** (2000) 5042 (`hep-th/9907001`)

[46] Yu. A. Golfand and E. P. Likhtman, *Extension of the algebra of Poincaré group generators and violations of P-invariance*, JETP Lett. **13** (1971) 323.
D. V. Volkov and V. P. Akulov, *Is the neutrino a Goldstone particle*, Phys. Lett. **B46** (1973) 109.
J. Wess and B. Zumino, *A Lagrangian model invariant under supergauge transformations*, Phys. Lett. **B49** (1974) 52.

[47] R. Haag, J. Łopuszański and M. Sohnius, *All possible generators of supersymmetries of the S-matrix*, Nucl. Phys. **B88** (1975) 257.

Quantum ChromoDynamics

M. H. Seymour
University of Manchester, UK, and CERN, Geneva, Switzerland

Abstract
These lectures on QCD stress the theoretical elements that underlie a wide range of phenomenological studies, particularly gauge invariance, renormalization, factorization and infrared safety. The three parts cover the basics of QCD, QCD at tree level, and higher order corrections.

1 Basics of QCD

1.1 Introduction

QCD is the theory of the strong nuclear force, one of the four fundamental forces of nature. It describes the interactions of quarks, via their colour quantum numbers. It is an unbroken gauge theory. The gauge bosons are gluons. It has a similar structure to QED, but with one important difference: the gauge group is non-Abelian, SU(3), and hence the gluons are self-interacting. This results in a negative β-function and hence asymptotic freedom at high energies and strong interactions at low energies.

These strong interactions are confining: only colour-singlet states can propagate over macroscopic distances. The only stable colour singlets are quark–antiquark pairs, mesons, and three-quark states, baryons. In high energy reactions, like deep inelastic scattering, the quark and gluon constituents of hadrons act as quasi-free particles, partons. Such reactions can be factorized into the convolution of non-perturbative functions that describe the distribution of partons in the hadron, which cannot be calculated from first principles (at present) but are universal (process-independent), with process-dependent functions, which can be calculated as perturbative expansions in the coupling constant $\alpha_{\rm S}$.

Beyond leading order in $\alpha_{\rm S}$, the parton distribution functions and coefficient functions become intermixed. They can still be factorized, but the parton distribution functions become energy-dependent. Although the input distributions at some fixed energy scale still cannot be calculated, the energy dependence is given by perturbative evolution equations.

In sufficiently inclusive cross sections, called infrared safe, the non-perturbative distributions cancel and distributions can directly be calculated in perturbation theory. Non-perturbative corrections are then suppressed by powers of the high energy scale. The most important examples are jet cross sections, where jets of hadrons have a direct connection to the perturbatively-calculable quarks and gluons.

This course will attempt to give a brief overview of the subject. The approach will be pretty phenomenological, with most results stated rather than derived. I will however attempt to sketch in most cases roughly how they would be derived. One thing I will not have time to go into in much detail will be heavy quarks: in most cases we will treat the d, u, s, c and b quarks as massless and neglect the top quark, an approximation that I will motivate in Section 1.9.

It is hard to give a better introduction to the subject than the book ‘QCD and Collider Physics’, by Keith Ellis, James Stirling and Bryan Webber [1]. So I will follow the ESW approach and notation pretty closely. In most cases they will be able to give you a few more details and references to much more detailed treatments if you want to go further. For a much more detailed treatment of the formulation of QCD and its renormalization in particular Peskin and Schroeder [2] is also unbeatable.

As there are many parallels with QED I will have to assume prior knowledge of the basics of QED and that you can calculate a few simple cross sections. However we start by recapitulating a few features.

1.2 Basics of QED

QED is a gauge theory with gauge group U(1). It can be derived using the gauge principle. The classical Lagrangian density for n types of non-interacting fermion is

$$\mathcal{L}_{\rm ferm} = \sum_i^n \bar{f}_i(i\not{\partial} - m_i)f_i, \tag{1.1}$$

where f_i is a spinor-valued wave function describing plane waves of momentum p_i, $\bar{f}_i$ its Dirac conjugate $f_i^\dagger\gamma^0$, $\not{a}$ is shorthand for $\gamma^\mu a_\mu$ and γ^μ are Dirac spinor matrices with anticommutation relation

$$\{\gamma^\mu,\gamma^\nu\} = 2g^{\mu\nu}. \tag{1.2}$$

The Lagrangian density (1.1) is invariant under global changes of gauge,

$$f_i \to f_i' = \exp(ie_i\theta)f_i, \tag{1.3}$$

where e_i is an arbitrary flavour-dependent parameter, which will turn out to be proportional to electric charge. We can derive QED by asking how we would need to modify (1.1) to make it also invariant under local changes of gauge,

$$f_i(x) \to f_i'(x) = \exp(ie_i\theta(x))f_i(x). \tag{1.4}$$

This can be done by introducing a new vector-valued field A_μ, which transforms under the same change of gauge like

$$A_\mu(x) \to A_\mu'(x) = A_\mu(x) + \frac{i}{e}\Big(\partial_\mu \exp(i\theta(x))\Big)\exp(-i\theta(x)), \tag{1.5}$$

and by replacing the derivative ∂_μ by the covariant derivative,

$$D_\mu = \partial_\mu + ie\hat{Q}\,A_\mu, \tag{1.6}$$

where $\hat{Q}$ is the charge operator, defined by

$$\hat{Q}\,f_i = e_i f_i. \tag{1.7}$$

Since A_μ is a new field that we have introduced, we must make it physical by adding a kinetic term,

$$\mathcal{L}_{\rm kin} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu}, \tag{1.8}$$

where the field strength tensor $F_{\mu\nu}$ is defined by

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{1.9}$$

The classical QED Lagrangian density is therefore given by

$$\mathcal{L}_{\rm classical} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \sum_i^n \bar{f}_i(i\not{D} - m_i)f_i. \tag{1.10}$$

This is now invariant under local changes of gauge.

Perturbative calculations are made according to the Feynman rules. These can be read off from the action, defined by

$$S = i\int d^4x\,\mathcal{L}. \tag{1.11}$$

There is however one complication. The photon propagator $\Delta_{\gamma,\mu\nu}(p)$ is derived from the inverse of the bilinear term in A_μ:

$$\Delta_{\gamma,\mu\nu}(p) \times i\Big[p^2 g^{\nu\sigma} - p^\nu p^\sigma\Big] = \delta_\mu^\sigma. \tag{1.12}$$

This does not have an inverse. However, we can exploit the gauge invariance of the theory to rewrite it in a physically equivalent form that is invertible. Since the Lagrangian density is gauge invariant, we can choose some convenient gauge to work in and the final answer should be independent of which we chose. For example, in the covariant gauge, we have the condition

$$\partial^\mu A_\mu = 0 \tag{1.13}$$

at every space-time point. We can therefore add an extra term to the Lagrangian density

$$\mathcal{L}_{\rm gauge-fixing} = -\frac{1}{2\lambda}\left(\partial^\mu A_\mu\right)^2, \tag{1.14}$$

where λ is an arbitrary parameter, and provided we work in a covariant gauge we cannot have changed the physics, since we have only added zero. (This is essentially just the method of undetermined Lagrange multipliers for minimizing an action subject to a constraint: the constraint is (1.13) and the multiplier is $1/2\lambda$.) The final results must clearly be independent of λ, although it will appear at intermediate steps of calculations. Common choices are $\lambda = 1$ (Feynman gauge) and $\lambda \to 0$ (Landau gauge). For arbitrary λ, we must now solve

$$\Delta_{\gamma,\mu\nu}(p) \times i\Big[p^2 g^{\nu\sigma} - (1-\frac{1}{\lambda})p^\nu p^\sigma\Big] = \delta^\sigma_\mu, \tag{1.15}$$

which yields

$$\Delta_{\gamma,\mu\nu} = \frac{i}{p^2}\left(-g_{\mu\nu} + (1-\lambda)\frac{p_\mu p_\nu}{p^2}\right). \tag{1.16}$$

Clearly the Feynman gauge offers significant calculational advantages, so we use it for most of the rest of this course.

Another popular class of gauges are the axial (or physical) gauges, defined in terms of an arbitrary vector n, by

$$\mathcal{L}_{\rm gauge-fixing} = -\frac{1}{2\lambda}\left(n^\mu A_\mu\right)^2. \tag{1.17}$$

These have the result that an on-shell photon has two polarization states, which, in the $(n+p)$ rest-frame, are purely transverse to its direction. The penalty is that the propagator becomes more complicated,

$$\Delta_{\gamma,\mu\nu} = \frac{i}{p^2}\left(-g_{\mu\nu} + \frac{n_\mu p_\nu + p_\mu n_\nu}{n\cdot p} - \frac{(n^2+\lambda p^2)p_\mu p_\nu}{(n\cdot p)^2}\right). \tag{1.18}$$

Obviously some simplification is obtained by setting $n^2 = 0$ and $\lambda \to 0$ (the 'lightcone' gauge), but practical calculations are still considerably more complicated than in covariant gauges. In particular, if making a numerical calculation, it is difficult to guarantee that the spurious singularities $n\cdot p \to 0$ cancel as they should.

We therefore have the Feynman rules (in Feynman gauge):

$$\Delta_i = \frac{i}{\not p - m_i} = i\frac{\not p + m_i}{p^2 - m_i^2}, \tag{1.19}$$

$$\Delta_{\gamma,\mu\nu} = i\frac{-g_{\mu\nu}}{p^2}, \tag{1.20}$$

$$\Gamma^\mu_{\gamma f_i \bar f_i} = -i\, e_i e \gamma^\mu. \tag{1.21}$$

To calculate the cross section for a given process, we must write down all possible diagrams, use the Feynman rules to give us the amplitude $i\mathcal{M}$, use Dirac algebra and trace theorems to calculate $\sum|\mathcal{M}|^2$, where the sum is over all unobserved quantum numbers for example spin, divide by the overcounting of incoming states, and integrate over phase space:

$$\sigma = \frac{1}{S}\,\frac{1}{2s}\int d\Gamma\ \sum|\mathcal{M}|^2. \tag{1.22}$$

An element of n-body phase space is given by

$$d\Gamma = \prod_{i=1}^{n}\left(\frac{d^4p_i}{(2\pi)^4}\,(2\pi)\delta(p_i^2-m_i^2)\right)(2\pi)^4\delta^4(p_{tot}-\textstyle\sum_i^n p_i) \tag{1.23}$$

$$= \prod_{i=1}^{n}\left(\frac{d^3p_i}{(2\pi)^3 2E_i}\right)(2\pi)^4\delta^4(p_{tot}-\textstyle\sum_i^n p_i). \tag{1.24}$$

For example, the cross section for $e^+e^-\to\mu^+\mu^-$ is calculated as follows. The amplitude is

$$i\mathcal{M} = \bar v(p_{e^+})(ie)\gamma^\mu u(p_{e^-})\; i\frac{-g_{\mu\nu}}{(p_{e^+}+p_{e^-})^2}\;\bar u(p_{\mu^-})(ie)\gamma^\nu v(p_{\mu^+}) \tag{1.25}$$

$$= \frac{-ie^2}{(p_{e^+}+p_{e^-})^2}\bar v(p_{e^+})\gamma^\mu u(p_{e^-})\;\bar u(p_{\mu^-})\gamma_\mu v(p_{\mu^+}) \tag{1.26}$$

and hence

$$\sum|\mathcal{M}|^2 = \frac{(4\pi\alpha)^2}{s^2}\;\mathrm{Tr}\left\{\not p_{e^+}\gamma^\mu\not p_{e^-}\gamma^\nu\right\}\;\mathrm{Tr}\left\{\not p_{\mu^-}\gamma_\mu\not p_{\mu^+}\gamma_\nu\right\}, \tag{1.27}$$

where $\alpha=e^2/4\pi$ and $s=(p_{e^+}+p_{e^-})^2$, or

$$\sum|\mathcal{M}|^2 = \frac{16(4\pi\alpha)^2}{s^2}\;\left(p^\mu_{e^+}p^\nu_{e^-}+p^\mu_{e^-}p^\nu_{e^+}-p_{e^+}\cdot p_{e^-}g^{\mu\nu}\right)\;\left(p_{\mu^-,\mu}p_{\mu^+,\nu}+p_{\mu^+,\mu}p_{\mu^-,\nu}-p_{\mu^+}\cdot p_{\mu^-}g_{\mu\nu}\right) \tag{1.28}$$

$$= 8(4\pi\alpha)^2\frac{t^2+u^2}{s^2}, \tag{1.29}$$

where $t=(p_{e^-}-p_{\mu^-})^2$ and $u=(p_{e^-}-p_{\mu^+})^2=-s-t$. The cross section is therefore

$$\sigma = \frac{1}{4}\,\frac{1}{2s}\int_{-s}^{0}\frac{dt}{8\pi s}\;8(4\pi\alpha)^2\frac{t^2+u^2}{s^2} \tag{1.30}$$

$$= \frac{4\pi\alpha^2}{3s}. \tag{1.31}$$

1.3 SU(3) and colour

QCD can be derived in exactly the same way as QED: we start from the Lagrangian density for a set of non-interacting quarks and modify it in just such a way that it is invariant under changes of gauge. The only difference is that instead of the gauge transformation being a simple phase (U(1) group), we consider a non-Abelian group SU(N_c). This has several important consequences. Fermion charges will come in N_c different types, called colours, they will be quantized (in contrast to the electric charges e_i, which could take any value) and, most importantly, the gauge bosons will be self-interacting.

It has been well-known since the early days of QCD that there are three colours, for example from baryon wave functions, the total e^+e^- cross section (which is proportional to N_c) and π^0 decay rate (which is proportional to N_c^2). However, in most calculations it is useful to keep the number of colours N_c arbitrary until the very last step when it is set equal to three. The N_c-dependent coefficients are a useful diagnostic tool in understanding the physical origins of different terms, comparing different calculations and tracking down errors.

We start by restating briefly some features of SU(N), the group of $N\times N$ unitary matrices ($U^\dagger U=1$) with determinant +1. Let U be an element of SU(N) that is infinitesimally close to the identity and write it as

$$U = 1+iG, \tag{1.32}$$

where G has infinitesimal entries. It must be hermitian ($G^\dagger = G$) and traceless. One can choose a basis set of N^2-1 matrices, t^A, $A = 1, \ldots, N^2-1$, such that any G can be written as

$$G = \sum_A^{N^2-1} \epsilon^A t^A, \tag{1.33}$$

where ϵ_A are infinitesimal numbers. Note that I will always denote colour indices that run from 1 to N by a and from 1 to N^2-1 by A. The t^A are called the generators of the group and define its fundamental representation. You can show that $[t^A, t^B]$ is antihermitian and traceless and hence can be written as a linear combination of other t^Cs,

$$[t^A, t^B] \equiv i\, f^{ABC} t^C, \tag{1.34}$$

where f^{ABC} are a set of real constants, called the structure constants of the group. It is straightforward to see that f^{ABC} is antisymmetric in A, B, and with a little more work, one can prove that it is antisymmetric in all its indices. Equation (1.34) defines the Lie algebra of the group.

We can also define a set of $(N^2-1) \times (N^2-1)$ matrices that obey the same algebra:

$$\left(T^A\right)_{BC} \equiv -if^{ABC}, \tag{1.35}$$

$$[T^A, T^B] = i\, f^{ABC} T^C. \tag{1.36}$$

These define the group's adjoint representation.

Although we started with elements infinitesimally close to the identity matrix, we can calculate an arbitrary element U by stringing together an infinite number of infinitesimal elements,

$$U = \lim_{N\to\infty} (1 + i\theta^A t^A/N)^N = \exp(i\theta^A t^A) \equiv \exp(it\cdot\theta). \tag{1.37}$$

Since U is unitary and t^A hermitian, we have

$$U^{-1} = \exp(-it\cdot\theta). \tag{1.38}$$

There are several identities we will require time and time again:

$$\begin{aligned}
\mathrm{Tr}(t^A t^B) &= \tfrac{1}{2}\delta^{AB} \equiv T_R \delta^{AB} && (1.39)\\
\sum_A t^A_{ab} t^A_{bc} &= \frac{N^2-1}{2N}\delta_{ac} \equiv C_F \delta_{ac} && (1.40)\\
\mathrm{Tr}(T^C T^D) &= \sum_{A,B} f^{ABC} f^{ABD} = N\delta^{CD} \equiv C_A \delta^{CD}, && (1.41)
\end{aligned}$$

where the constants C_F and C_A are the Casimir operators of the fundamental and adjoint representations of the group respectively. Although we know the numerical values of these constants:

$$\begin{aligned}
T_R &= \frac{1}{2}, && (1.42)\\
C_F &= \frac{4}{3}, && (1.43)\\
C_A &= 3, && (1.44)
\end{aligned}$$

it is good practice, as I said, to leave them unexpanded in all algebraic results.

In fact for practical calculations one only requires these, and other similar, identities and never an explicit representation for t^A or f^{ABC}.

1.4 The QCD Lagrangian

The classical Lagrangian density for n non-interacting quarks with masses m_i is

$$\mathcal{L}_{\text{quarks}} = \sum_i^n \bar{q}_i^a (i\partial\!\!\!/ - m_i)_{ab} q_i^b, \tag{1.45}$$

where the factor $(i\partial\!\!\!/ - m_i)_{ab}$ is proportional to the identity matrix in colour space. This is invariant under global SU(N_c) transformations,

$$q_a \to q'_a = \exp(it\cdot\theta)_{ab} q_b. \tag{1.46}$$

To make it invariant under local transformations,

$$q_a(x) \to q'_a(x) = \exp(it\cdot\theta(x))_{ab} q_b(x), \tag{1.47}$$

we have to introduce the covariant derivative,

$$D_{\mu,ab} = \partial_\mu 1_{ab} + ig_s (t\cdot A_\mu)_{ab}, \tag{1.48}$$

where A_μ^A are coloured vector fields that transform in just the right way that we have

$$D'_{\mu,ab} q'_b(x) = \exp(it\cdot\theta(x))_{ab} D_{\mu,bc} q_c(x), \tag{1.49}$$

giving

$$t\cdot A'_\mu = \exp(it\cdot\theta(x))\, t\cdot A_\mu \, \exp(-it\cdot\theta(x)) + \frac{i}{g_s}\Big(\partial_\mu \exp(it\cdot\theta(x))\Big) \exp(-it\cdot\theta(x)). \tag{1.50}$$

We again have to introduce a kinetic term for this new field,

$$\mathcal{L}_{\text{kin}} = -\frac{1}{4} F^A_{\mu\nu} F_A^{\mu\nu}, \tag{1.51}$$

where $F^A_{\mu\nu}$ is the non-Abelian field strength tensor. However, the definition we used in QED (1.9) does not result in an invariant Lagrangian density under transformation (1.50). One must add an extra term,

$$F^A_{\mu\nu} = \partial_\mu A^A_\nu - \partial_\nu A^A_\mu - g_s f^{ABC} A^B_\mu A^C_\nu, \tag{1.52}$$

and only then is (1.51) invariant under gauge transformations.

This extra term has profound consequences for the theory: it means that gluons are self-interacting, through three- and four-point vertices. This will turn out to give rise to asymptotic freedom at high energies and strong interactions at low energies, among the most fundamental properties of QCD. We therefore see that these are absolute requirements of the SU(N_c) gauge symmetry.

Before reading off the Feynman rules we again have to fix the gauge. This proceeds in exactly the same way as in QED, leading to, in covariant gauges,

$$\mathcal{L}_{\text{gauge-fixing}} = -\frac{1}{2\lambda}\left(\partial^\mu A^A_\mu\right)^2. \tag{1.53}$$

Finally, it turns out that in a non-Abelian gauge theory, it is necessary to add one extra term to the Lagrangian density, related to the need for ghost particles. These are beyond the scope of this course, but basically they arise because when a non-Abelian gauge theory is renormalized it is possible for unphysical degrees of freedom to propagate freely. These are cancelled off by introducing into the theory an unphysical set of fields, the ghosts, which are scalars but have Fermi statistics. For practical purposes it is enough to know that there exist Feynman rules for ghosts and that in every diagram with a closed loop of internal gluons containing only triple-gluon vertices, we must add a diagram with the gluons in the loop replaced by ghosts. It is worth noting that in physical gauges, as the name suggests, ghost contributions always vanish and they can be ignored.

The final Lagrangian is therefore

$$\mathcal{L}_{\text{QCD}} = -\frac{1}{4} F^A_{\mu\nu} F_A^{\mu\nu} + \sum_i^n \bar{q}_i^a (i\not{D} - m_i)_{ab} q_i^b - \frac{1}{2\lambda}\left(\partial^\mu A^A_\mu\right)^2 + \mathcal{L}_{\text{ghost}}. \tag{1.54}$$

1.5 Feynman rules

Just as in QED it is straightforward to read off the Feynman rules from the action. We obtain in Feynman gauge (only the gluon propagator is gauge dependent)

$$\Delta_i^{ab} = \delta^{ab}\frac{i}{\not{p}-m_i} = \delta^{ab} i\frac{\not{p}+m_i}{p^2-m_i^2}, \tag{1.55}$$

$$\Delta_{g,\mu\nu}^{AB} = \delta^{AB} i\frac{-g_{\mu\nu}}{p^2}, \tag{1.56}$$

$$\Gamma_{gq\bar{q}}^{\mu} = -i\,g_s\,t^A\,\gamma^\mu, \tag{1.57}$$

$$\Gamma_{ggg} = -g_s f^{ABC}\left[(p-q)^\lambda g^{\mu\nu} + (q-r)^\mu g^{\nu\lambda} + (r-p)^\nu g^{\lambda\mu}\right]. \tag{1.58}$$

Note that, apart from the triple-gluon vertex, the only difference relative to QED is in the colour structure: propagators are diagonal in colour and the vertex for a gluon of colour A to scatter a quark of colour b to a quark of colour c contains $(t^A)_{cb}$. Note also that unlike QED the quark–gluon vertex is flavour-independent (it is straightforward to check that, unlike in QED, we cannot introduce a flavour-dependence into the gauge transformation, Eq. (1.47) and retain gauge invariance). In the triple-gluon vertex, the three gluons have momenta p, q, r, Lorentz indices μ, ν, λ and colour indices A, B, C respectively. The momenta are all ingoing: $p + q + r = 0$.

The Feynman rules for ghosts and for the four-gluon vertex can be found in ESW [1] (p. 10). They will not be needed for this course.

Note also that in analogy with QED the strong charge g_s is usually substituted by $\alpha_{\rm S}$,

$$\alpha_{\rm S} \equiv \frac{g_s^2}{4\pi}. \tag{1.59}$$

1.6 $e^+e^- \to q\bar{q}$

One of the most fundamental quantities in QCD is the total e^+e^- annihilation cross section to hadrons. We will see in a later lecture that to leading order in $\alpha_{\rm S}$ this is equal to the total $e^+e^- \to q\bar{q}$ cross section. The calculation is very similar to that for $e^+e^- \to \mu^+\mu^-$, the only difference being in the colour structure. The photon is colour blind, so the Feynman rule for a photon to couple to a quark contains a trivial colour matrix, δ^{ab}. Summing over colours and dividing by the number of incoming colour states (1 in this case since electrons are not coloured), we therefore obtain

$$\sigma(e^+e^- \to q\bar{q}) = \sigma(e^+e^- \to \mu^+\mu^-) \times e_q^2 \times \sum_{a,b}\delta^{ab}\delta^{ba}. \tag{1.60}$$

We obtain

$$\sum_{a,b}\delta^{ab}\delta^{ba} = \sum_a \delta^{aa} = N_c, \tag{1.61}$$

and hence

$$R_{\rm had} \equiv \frac{\sigma(e^+e^- \to \text{hadrons})}{\sigma(e^+e^- \to \mu^+\mu^-)} = \sum_q e_q^2 N_c. \tag{1.62}$$

1.7 $e^+e^- \to q\bar{q}g$

This process will be important for the higher order corrections to $\sigma(e^+e^- \to \text{hadrons})$ and particularly for the study of three-jet final states in e^+e^- annihilation, among the most important test-beds for QCD.

There are two Feynman diagrams, shown in Fig. 1.1. We label the momenta and colours $e^-(p_-) +$

Fig. 1.1: Feynman diagrams for the process $e^+e^- \to q\bar{q}g$

$e^+(p_+) \to q_a(p_1) + \bar{q}_b(p_2) + g_A(p_3)$. For the matrix element we obtain

$$\begin{aligned} i\mathcal{M} &= \bar{v}(p_+)(ie)\gamma^\mu u(p_-)\, i\frac{-g_{\mu\nu}}{s}\,\varepsilon_A^{*\lambda} \\ &\quad \bar{u}_a(p_1)\left\{(-ig_s)t^A_{ab}\gamma^\lambda\frac{\not{p}_1+\not{p}_3}{(p_1+p_3)^2}(-iee_q)\gamma^\nu + (-iee_q)\gamma^\nu\frac{-\not{p}_2-\not{p}_3}{(p_2+p_3)^2}(-ig_s)t^A_{ab}\gamma^\lambda\right\}v_b(p_2). \end{aligned} \tag{1.63}$$

We will evaluate the cross section from this matrix element later. Here we are interested in the colour algebra. Using the fact that the spin sum of a massless vector particle is proportional to the colour identity matrix,

$$\sum_{\text{spin}} \varepsilon_A^{*\mu}\varepsilon_B^{\nu} = -g^{\mu\nu}\delta_{AB}, \tag{1.64}$$

we obtain

$$\sum|\mathcal{M}|^2 \propto \sum_{a,b,A} t^A_{ab}\left(t^A_{ab}\right)^* = \sum_{a,b,A} t^A_{ab}t^A_{ba} = \sum_A \mathrm{Tr}(t^At^A) = C_F\mathrm{Tr}(1) = C_FN_c, \tag{1.65}$$

where the first step uses the fact that t^A are hermitian, the second is simply a trivial rewrite, switching to matrix notation, the third uses Eq. (1.40) and the fourth uses the fact that the matrix being traced is the identity matrix of the fundamental representation, i.e. the $N_c \times N_c$ identity matrix. Note that since the colour factor of the lowest order process is N_c, we can associate C_F with the emission of the additional gluon. Since the emission probability of a gluon from a quark is proportional to C_F, and we will later see that that from a gluon is proportional to C_A, C_F and C_A are sometimes referred to as the squares of the colour charges of the quark and gluon respectively.

Performing the trace Dirac algebra on the matrix element, we finally obtain

$$\sum|\mathcal{M}|^2 = \frac{16C_FN_ce^4e_q^2g_s^2}{s\,p_1{\cdot}p_3\,p_2{\cdot}p_3}\left((p_1{\cdot}p_+)^2 + (p_2{\cdot}p_+)^2 + (p_1{\cdot}p_-)^2 + (p_2{\cdot}p_-)^2\right). \tag{1.66}$$

(Note the misprint in ESW [1] — their result is a factor of 4 too large.)

1.8 The coupling constant α_{S} and renormalization

As we mentioned above, in practical calculations, α_{S} is usually used rather than g_s. Besides the quark masses, which we will neglect in most of this course, g_s is the only parameter in the QCD Lagrangian and therefore assumes a central role in our study of QCD. However, it is not *a priori* clear that parameters in the Lagrangian are physically observable quantities — any physical observable can be calculated as a function of them (at least in perturbation theory) and their values can be extracted from measured values of physical observables, but they are not necessarily themselves physical. It is worthwhile therefore to consider whether we can reformulate our theory in such a way that one physical observable can be written as a function of another. This reformulation is known as renormalization.

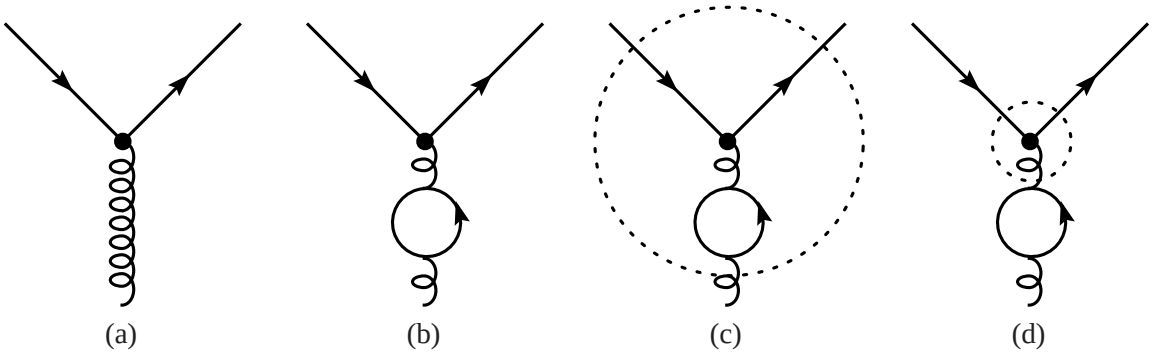


Fig. 1.2: When a quark–gluon vertex (a) is corrected by a loop (b), one must decide whether to describe it as a correction to the vertex (c), or to the rest of the diagram (d)

In this section I give a very handwaving description of renormalization, which I believe conveys the important physical point. Of course for practical calculations one needs a much more precise definition of the renormalization prescription, which I describe at the end.

We redefine g_s to be the strength of the quark–gluon coupling, as in Fig. 1.2a. At first sight, this seems like a trivial statement and at the lowest order of perturbation theory it is — the two definitions are identical. However, when we calculate higher orders of perturbation theory, we encounter loop corrections like the one in Fig. 1.2b, which correct the vertex. To avoid double-counting, we must uniquely decide whether these corrections are part of the vertex, as in Fig. 1.2c, or the rest of the diagram, as in Fig. 1.2d. One way to decide is to introduce a *renormalization scale* μ_R and say that physics at high scales (therefore short distances) above μ_R is part of the vertex and physics at lower scales (longer distances) below μ_R is part of the rest of the diagram. Of course, this is simply a book-keeping device, which does not change the physics, it simply ensures that each physical contribution to the process is counted once and only once. Since μ_R is a completely arbitrary book-keeping scale, introduced by hand, its value should not affect the physical prediction — changing it simply moves contributions between what we call the vertex and what we call the rest of the diagram. Since the amount of physics that we include in the vertex depends on μ_R, and we defined g_s to be the strength of the vertex, it is clear that g_s must now be a function of μ_R.

It is worth mentioning that, although I defined g_s as the strength of the quark–gluon vertex, I could equally well have defined it as the strength of the triple-gluon vertex. It is one of the remarkable features of gauge theories that, as a direct result of the gauge symmetry, I would get exactly the same result for the renormalized coupling $g_s(\mu_R)$. That is, the equality of the strengths of the quark–gluon and triple-gluon vertices is true even after renormalization.

When it is clear that I am talking about the renormalization scale, I will henceforth drop the subscript $_R$.

1.8.1 Renormalization group equation

As I said, varying μ moves physical contributions (loop corrections) around within a calculation, but it should not change the result of the physical calculation. We can use this fact to derive an equation for how g_s varies as a function of μ. This is one of a set of equations that together describe how the whole theory varies with renormalization scale (and scheme), which formally form a group.

We study this by considering a dimensionless physical observable R that is a function of only one physical scale Q^2 (think of $R_{\rm had}$ at $\sqrt{s} = Q$ for example). Assume that this observable is not sensitive to quark masses (we will return to this point shortly). After renormalization, R can only be a function of Q^2, μ^2 and $\alpha_{\rm S}(\mu^2)$. By dimensional analysis, the only way the dimensionless function R can depend on

the dimensionful variables Q^2 and μ^2 is through their ratio. We can therefore write

$$R = R(Q^2/\mu^2, \alpha_{\rm S}(\mu^2)). \tag{1.67}$$

We can use the fact that R, as a physical quantity, must be independent of the value of μ, and the chain rule for partial derivatives, to write

$$\mu^2\frac{d}{d\mu^2}R(Q^2/\mu^2,\alpha_{\rm S}) = 0 \quad = \quad \left[\mu^2\frac{\partial}{\partial\mu^2} + \mu^2\frac{\partial\alpha_{\rm S}}{\partial\mu^2}\frac{\partial}{\partial\alpha_{\rm S}}\right]R \tag{1.68}$$

$$\equiv \quad \left[\mu^2\frac{\partial}{\partial\mu^2} + \beta(\alpha_{\rm S})\frac{\partial}{\partial\alpha_{\rm S}}\right]R\,, \tag{1.69}$$

i.e., $\beta(\alpha_{\rm S}) \equiv \mu^2\frac{\partial\alpha_{\rm S}}{\partial\mu^2}$. There are several points to note about this.

- A physical solution is provided by $R(1, \alpha_{\rm S}(Q))$, i.e., by setting the renormalization scale equal to the physical scale in the problem.
- Q-dependence of the physical quantity R comes about only because of the renormalization of the theory and would not be present in the classical theory. Thus measuring the Q-dependence of R directly probes the quantum structure of the theory.
- By rearranging Eq. (1.69), one can derive the μ^2 dependence of $\alpha_{\rm S}$ from a calculation of R,

$$\beta(\alpha_{\rm S}) = -\frac{\mu^2\frac{\partial R}{\partial\mu^2}}{\frac{\partial R}{\partial\alpha_{\rm S}}}. \tag{1.70}$$

- If $\alpha_{\rm S}$ is small, R is perturbatively calculable and hence $\beta(\alpha_{\rm S})$ is too.

The β function of QCD is now known to four-loop accuracy,

$$\beta(\alpha_{\rm S}) = -\alpha_{\rm S}^2(\beta_0 + \beta_1\alpha_{\rm S} + \beta_2\alpha_{\rm S}^2 + \beta_3\alpha_{\rm S}^2 + \ldots). \tag{1.71}$$

Although the higher orders are essential for quantitative calculation, they are not for qualitative understanding: almost all QCD phenomenology can be understood using the one loop result,

$$\beta_0 = \frac{11C_A - 4T_RN_f}{12\pi}, \tag{1.72}$$

where N_f is the number of quark flavours that can appear in loops, to be discussed further shortly.

Note that β_0 is positive and hence that the β function is negative, at least when $\alpha_{\rm S}$ is small. This results in asymptotic freedom: the fact that the interactions become weak at high energies (short distances) and infrared slavery: the fact that they become strong at low energy.

If we neglect the higher orders, we can solve the renormalization group equation exactly, to obtain $\alpha_{\rm S}$ at some scale Q as a function of its value at the renormalization scale μ,

$$\alpha_s(Q^2) = \frac{\alpha_s(\mu^2)}{1 + \alpha_s(\mu^2)\beta_0\ln\frac{Q^2}{\mu^2}}. \tag{1.73}$$

1.8.2 Choosing μ^2

Although physical quantities do not depend on μ, a calculation truncated at a finite order of perturbation theory does. We must therefore choose some value for μ. To illustrate this, suppose that our dimensionless physical quantity R has a perturbative expansion that starts at $\mathcal{O}(\alpha_{\rm S})$,

$$R = R_1\alpha_{\rm S} + \ldots, \tag{1.74}$$

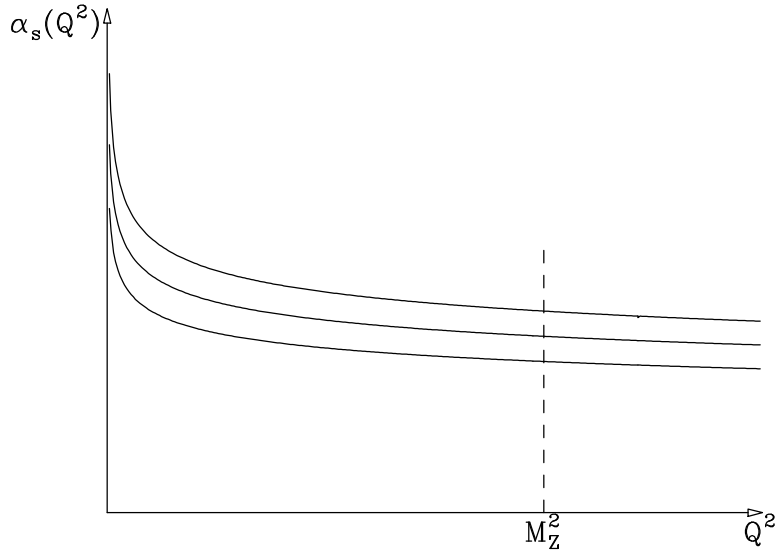


Fig. 1.3: A measurement of $\alpha_{\rm S}$ at any scale Q fixes which curve our universe lies on, but to compare measurements at different scales we have to agree to label the curves in a standard way, for example using $\alpha_{\rm S}(M_z)$

then if we truncate at leading order,

$$R \approx R_1 \alpha_{\rm S}, \tag{1.75}$$

our truncated expression for $R(1, \alpha_{\rm S}(Q))$ can be expanded as a power series in $\alpha_{\rm S}(\mu^2)$

$$\begin{aligned} R(1, \alpha_{\rm S}(Q^2)) &\approx R_1 \alpha_{\rm S}(Q^2) & (1.76)\\ &= R_1 \alpha_{\rm S}(\mu^2) \left[1 - \beta_0 \alpha_{\rm S}(\mu^2) \ln \frac{Q^2}{\mu^2} + \beta_0^2 \alpha_{\rm S}^2(\mu^2) \ln^2 \frac{Q^2}{\mu^2} + \ldots \right]. & (1.77) \end{aligned}$$

The leading order result in renormalized perturbation theory is the first term of this series, i.e., $R_1 \alpha_{\rm S}(\mu^2)$. It is therefore clear that although μ is completely arbitrary, choosing it far from Q guarantees a large truncation error (note that the converse is not true). One should therefore choose μ^2 'close' to Q^2, but how close is close?

The conventional approach is to set $\mu = Q$ and to use the μ variation in a reasonable range, e.g., $Q/2$ to $2Q$ as an estimate of the truncation uncertainty. It should be clear from the foregoing discussion that this is an extremely arbitrary procedure. However, the folklore is that in almost all cases where higher order corrections have been calculated, they have fallen within the band given by this procedure.

1.8.3 Measuring $\alpha_{\rm S}$

The β function tells us how $\alpha_{\rm S}$ varies with scale, but it does not tell us the value of $\alpha_{\rm S}$ at any particular scale: we need an experimental measurement to do that. Effectively $\beta(\alpha_{\rm S})$ defines a family of curves, as illustrated in Fig. 1.3, and one measurement at any scale is sufficient to tell us which curve our universe lies on. However, in order to compare and combine measurements of $\alpha_{\rm S}$ at different scales, we have to agree on some convenient labeling of the curves. The measurement at any given scale can then be converted into a measurement of the label. Historically, this was often done using the 'QCD scale', Λ_{QCD}, described in the next section, but more recently it has been realized that the value of $\alpha_{\rm S}$ at some fixed scale at which it is relatively small is a lot more convenient. Since some of the best measurements come from Z^0 decays, it has become universal to use $\alpha_{\rm S}(M_z)$ as the label. We will discuss the measurements of $\alpha_{\rm S}$ further in Section 3.1.4.

1.8.4 The 'QCD Scale', Λ

As I just mentioned, this is another way to label the running coupling, which is to construct a renormalization group invariant scale from $\alpha_{\rm S}(\mu)$. Although the Lagrangian of massless QCD has no scale, the renormalization process introduces a dimensionful parameter,

$$\Lambda^2 = \mu^2 \exp - \int^{\alpha_{\rm S}(\mu^2)} \frac{dx}{\beta(x)} \approx \mu^2 {\rm e}^{-1/\beta_0 \alpha_{\rm S}(\mu^2)}, \tag{1.78}$$

where the approximation uses only the one-loop term in the β function[1.1]. This process by which a scaleless theory gets a physically observable scale by the introduction of the unphysical renormalization scale is known as dimensional transmutation.

At leading order, Λ has a simple interpretation, it is the scale at which the coupling becomes infinite. However, this interpretation is not self-consistent, since it relies on a truncation of the perturbation series in a region in which the coupling is large, ultimately divergent. More generally, Λ can be viewed as a renormalization group invariant parameterization of the scale at which the theory becomes non-perturbative. All non-perturbative quantities, for example the hadron masses, would be expected to be of order Λ.

However, Λ is not a very practically useful label for the value of $\alpha_{\rm S}$. This is because its precise value, for a given measured value of $\alpha_{\rm S}$, depends strongly on the theoretical input used in the calculation, for example which order of perturbation theory we truncate β at, which renormalization scheme we use, the number of flavours we assume, the way we match the running coupling at the flavour thresholds, etc.

In principle any labeling suffers from these problems, but by using the value of $\alpha_{\rm S}$ in a region where it is small, and where the scale is not too different from that at which the measurements are made, the impact on $\alpha_{\rm S}(M_z)$ is small, whereas Λ is related to the region where $\alpha_{\rm S}$ is large, far away from where the measurements are made, and these effects are large.

1.8.5 Renormalization in practice

To give a simple physical picture of renormalization, I have described it in terms of a cutoff on the scale of the physical effects that are included in different components of a Feynman diagram calculation. However, in practice, this definition is extremely unattractive, because it breaks Lorentz and gauge invariance, two of the fundamental symmetries of our theory. If calculating in this scheme, these symmetries will get violated by a truncation at any finite order of perturbation theory and only restored in an all orders calculation. There are other simple schemes that work well in certain cases, for example the Pauli–Villars regularization, but the only known scheme consistent with all the symmetries of QCD, and hence guaranteed to work at any order of perturbation theory, is *dimensional regularization*. In this section I give a very brief description of how this works in practice. The difference between μ and μ_R will be (slightly) relevant here, so I temporarily reinstate the subscript.

The basic observation is that the loop corrections that we have been discussing are divergent in four or more space-time dimensions, but are finite in less than four dimensions. We therefore choose to calculate our Feynman diagrams in $d < 4$ dimensions (we always work in Minkowski space, with one time dimension and $d-1$ space dimensions). With a little thought, we can analytically continue

[1.1] Note that the definition in the first equality of Eq. (1.78), while formally renormalization group invariant, is not practically useful, since the lower limit of the integration is not defined (corresponding to the fact that any definition of Λ that differs by a multiplicative constant is equally renormalization group invariant). For perturbative calculations, various definitions, equivalent to Eq. (1.78) to the order to which they are defined, can be used. For non-perturbative calculations, for example in lattice QCD, the precise definition is more critical. A commonly-used convention (see for example [3]) is $\Lambda^2 = \mu^2 \exp\left\{-\frac{1}{\beta_0\alpha_{\rm S}(\mu^2)} - \frac{\beta_1}{\beta_0^2}\log\alpha_{\rm S}(\mu^2) - \int_0^{\alpha_{\rm S}(\mu^2)} dx \left(\frac{1}{\beta(x)} + \frac{1-\frac{\beta_1}{\beta_0}x}{\beta_0 x^2}\right)\right\}$. In contrast to the definition given in [1], for example, this can be seen to depend only on the β function at $\alpha_{\rm S}(\mu^2)$ and at smaller values, so is well-defined perturbatively and, as can be easily checked, is exactly renormalization scheme invariant.

the number of dimensions to be a complex number such that at the end of the calculation, after the renormalization prescription has been followed, we can let it smoothly tend back to 4 and obtain finite results. We therefore define $d = 4 - 2\epsilon$ and consider the $\epsilon \to 0_+$ limit.

By counting the dimensionality of terms in the Lagrangian, we discover that the coupling constant becomes dimensionful in $d \neq 4$ dimensions. This is not very convenient, so we define a dimensionless parameter $\alpha_{\rm S}$, by introducing a completely arbitrary scale μ,

$$\alpha_{\rm S}^{(d)} = \alpha_{\rm S}\, \mu^{2\epsilon}, \tag{1.79}$$

where $\alpha_{\rm S}^{(d)}$ is the dimensionful d-dimensional coupling. μ is called the regularization scale. It is often set equal to the renormalization scale μ_R, but I consider this confusing since we have not yet renormalized the theory, so, for now, I keep them distinct and only set them equal again at the end of this section.

When calculating loop corrections, we then find terms that have $1/\epsilon$ singularities for small ϵ. These have the right form to be absorbed by a redefinition (i.e. a renormalization) of the coupling. Since we also want the renormalized coupling to be dimensionless, we have to introduce a dimensionful scale at which the renormalization is performed, μ_R. To make this concrete, at one-loop order, the prescription is straightforward: after calculating all the one-loop diagrams, rewrite all occurrences of $\alpha_{\rm S}$ in terms of the renormalized coupling,

$$\alpha_{\rm S}(\mu_R) = \alpha_{\rm S} + \beta_0\, F(\epsilon)\, \left(\frac{\mu^2}{\mu_R^2}\right)^{\epsilon} \frac{1}{\epsilon}\, \alpha_{\rm S}^2\,. \tag{1.80}$$

Provided $F(0) = 1$, once this substitution has been made, the amplitude is finite. That is, the ϵ poles that this expression produces exactly cancel those from the one-loop calculation. Moreover, the arbitrary scale μ cancels from the amplitude at this point. One is left with a finite amplitude that depends only on μ_R and $\alpha_{\rm S}(\mu_R)$, in exactly the same way as discussed earlier.

The arbitrary function $F(\epsilon) = 1 + \mathcal{O}(\epsilon)$ defines the renormalization scheme. More precisely, it defines what finite parts of the loop amplitude are subtracted into the renormalized coupling, in addition to the divergent part. The MS, or minimal subtraction, scheme, is defined by subtracting nothing else,

$$F_{\rm MS}(\epsilon) = 1. \tag{1.81}$$

The most commonly used scheme is the $\overline{\rm MS}$, or modified minimal subtraction, scheme, in which one identifies some additional overall factors coming from the analytical continuation of the angular integrations in the one-loop calculation. Since they are universal it is convenient to subtract them into the coupling,

$$F_{\overline{\rm MS}}(\epsilon) = \frac{(4\pi)^{\epsilon}}{\Gamma(1-\epsilon)} = 1 + (\ln 4\pi - \gamma_E)\epsilon, \tag{1.82}$$

where Γ is the Euler gamma function and γ_E the Euler gamma constant, $\gamma_E \approx 0.577216$. Note that the two expressions on the right-hand side of Eq. (1.82) differ at order ϵ^2. Different practitioners use either of the two definitions, resulting in a finite difference at two loops that is straightforward to keep track of.

1.9 Quark masses and decoupling

The quark masses m_q are also parameters of the Lagrangian and face the same issues: for a physical calculation we should redefine them in a physical way. For the electron mass, we have a simple definition: we can isolate a single electron and 'weigh' it in the laboratory. That is, we can define its mass through the classical limit. We cannot use the same procedure for quarks, because confinement means that we can never take a single quark off to our laboratory to weigh it individually. We must therefore define some other renormalization procedure.

It is possible to proceed in close analogy with the coupling strength. We renormalize our theory at the same scale μ. We encounter gluon loop corrections to the quark propagator and absorb the part of them at scales above μ into the definition of the mass. We therefore obtain a 'running' (i.e. scale-dependent) mass. Just like for the coupling, we can obtain a renormalization group equation with perturbatively-calculable coefficients,

$$\frac{\mu^2}{m}\frac{dm}{d\mu^2} = -\frac{1}{\pi}\alpha_{\rm S}(\mu^2) + \dots . \tag{1.83}$$

At leading order it can be solved exactly, to give

$$m(\mu^2) = M\left[\alpha_{\rm S}(\mu^2)\right]^{\frac{1}{\pi\beta_0}}, \tag{1.84}$$

where M is a renormalization group invariant constant (c.f. Λ_{QCD}). Note that increasing μ^2 decreases m^2. Thus quarks appear to get lighter as they are probed at scales further and further above their masses.

An alternative scheme, which is often used in electroweak physics, and in the physics of heavy mesons, is the *pole mass*. Here one defines m_q to be the pole of the propagator $i(\not{p}+m_q)/(p^2-m_q^2)$ to all orders. This is very useful for $Q\sim m_q$, but it turns out that it is similar to a running mass scheme with μ of order m_q and hence generates large logarithms and a large truncation error for $Q\gg m_q$.

If our dimensionless observable R is finite for massless quarks then the quark mass effects must vanish smoothly as the mass goes to zero. Therefore the mass effects must be suppressed by $(m_q/Q)^n$, with $n\geq 1$. If there are quarks with mass much greater than Q, they can only affect our observable through loop corrections. A dimensional argument shows that such corrections must be suppressed by $(Q/m_q)^n$, with $n\geq 2$.

These observations form the basis of the decoupling theorem, in which quarks heavier than our physical scale can be ignored, and quarks lighter than it can be treated as massless. Thus, for most QCD calculations, we work with N_f flavours of massless quark (recall the N_f that appeared in β_0). Care must be taken when Q is close to a quark mass, or we study a range of processes at scales that span a quark mass, but in fact for most of the phenomenology considered in this course we can simply take N_f to be fixed, $N_f=5$.

1.10 Summary

We have seen that QCD is a gauge theory. The fact that the gauge symmetry is non-Abelian predicts that the gluon is self-interacting. This leads to the fact that the theory becomes strongly interacting at low energies, and hence non-perturbative, and weakly interacting at high energies so that perturbation theory can be used.

The main tools that we will use to study QCD are the *factorization* of non-perturbative effects and the *renormalization* and *decoupling* of high-energy physics. These allow us to use perturbation theory and, in particular, the Feynman rules, to study the phenomenology of QCD.

2 QCD phenomenology at tree level

Leading order perturbation theory, together with the one-loop renormalization group equation is enough to understand a wide variety of QCD phenomenology. In this section, we briefly review the phenomenology of QCD before introducing the complications of loop corrections to it in the following section. Most of the salient ideas are introduced in the context of e^+e^- annihilation and deep inelastic scattering, but apply equally well to hadron collisions and photoproduction, which we discuss more briefly at the end.

2.1 The cross section for $e^+e^-\to$ hadrons

One of the most striking features of e^+e^- annihilation events is the fact that many of them produce many hadrons. In trying to calculate the cross section for this process, however, we are immediately faced with

a problem: the Lagrangian does not contain any information about hadrons, so there are no Feynman rules involving them. Even if there were, calculating all the diagrams for events involving thirty or forty particles would be prohibitively complicated, let alone integrating them over the corresponding phase space to produce a total cross section. Fortunately a simple application of the Feynman rules of QED, together with some simple symmetry arguments, allows us to make a surprisingly strong statement about the cross section for e^+e^- annihilation to hadrons.

We postulate that the matrix element for the sum of all diagrams in which a virtual photon with Lorentz index ν and momentum q produces a particular set of n hadrons with momenta $\{p_1 \dots p_n\}$ is known and parameterize it by a function $T_\nu(n, q, \{p_1 \dots p_n\})$. Using this function, it is straightforward to write down the matrix element for the full process,

$$\mathcal{M} = \{\bar{v}(q_2) e \gamma_\mu u(q_1)\} \; \frac{-g^{\mu\nu}}{q^2} \; T_\nu(n, q, \{p_1 \dots p_n\}) \tag{2.1}$$

and hence the phase-space integral for its total cross section. The total cross section to produce any number of any type of hadrons is then simply given by the sum of this integral over hadron type and multiplicity (both generically represented by $\sum_n$),

$$\sigma = \frac{1}{2s} \frac{1}{4} \frac{e^2}{s^2} \mathrm{Tr}(\not{q}_2 \gamma^\mu \not{q}_1 \gamma^\nu) \tag{2.2}$$

$$\times \sum_n \int dPS_n \; T_\mu(n, q, \{p_1 \dots p_n\}) \; T^*_\nu(n, q, \{p_1 \dots p_n\}). \tag{2.3}$$

We then define a new two-index tensor, $H_{\mu\nu}$, to represent this sum of integrals,

$$H_{\mu\nu}(q) \equiv \sum_n \int dPS_n \; T_\mu \; T^*_\nu, \tag{2.4}$$

which after the integration and summation can only be a function of q[2.1]. Now, there are only two possible Lorentz covariant two-index tensor functions of one four-vector, $g_{\mu\nu}$ and $q_\mu q_\nu$. We therefore parameterize $H_{\mu\nu}$ as a linear combination of these, with coefficients that are functions of the only available Lorentz scalar, q^2,

$$H_{\mu\nu} = A(q^2) g_{\mu\nu} + B(q^2) q_\mu q_\nu. \tag{2.5}$$

Finally, since the theory is gauge invariant (in practice boiling down to invariance under the change $\varepsilon^\mu \to \varepsilon^\mu + q^\mu$ for the polarization vector of a photon of momentum q), $H_{\mu\nu}$ must be perpendicular to both q^μ and q^ν,

$$q^\mu H_{\mu\nu} = q^\nu H_{\mu\nu} = 0, \tag{2.6}$$

giving a relation between the two functions,

$$A = -q^2 B. \tag{2.7}$$

The final step is to realize that $B(s)$ has to be dimensionless. Since it is a function of only one dimensionful parameter, it must therefore be constant. We therefore have the fundamental prediction that (for energies well above all hadron masses) the cross section to produce any number of hadrons is proportional to that to produce a muon–antimuon pair,

$$R(e^+e^-) \equiv \frac{\sigma(e^+e^- \to \text{hadrons})}{\sigma(e^+e^- \to \mu^+\mu^-)} = \text{constant}, \tag{2.8}$$

[2.1] Can you spot the flaw in this argument? It assumes that all information about the hadron momenta is washed out by the integration, which is only true if they are massless. In general since p_h^2 is fixed at m_h^2 during the integration, H also depends in a complicated way on the masses of all possible hadrons. In fact we will shortly justify, on the basis of a space-time picture, neglecting these, in the limit that q^2 is much greater than all m_h^2. It also ignores any other masses in the problem, like the Z mass, which we remedy later on.

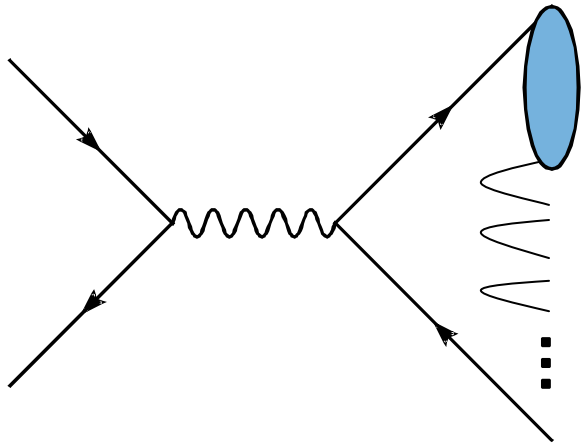


Fig. 2.1: Space-time sketch of the production of a hadron in e^+e^- annihilation

without knowing anything about the interactions of hadrons!

In order to go further than this and try to predict this constant, or learn something from its measurement, we need a specific model of the production of hadrons. This is provided by the *quark parton model*. Of course this can be more rigorously derived, but I find it more useful to illustrate the physics with a space-time argument, see Fig. 2.1. Since the photon is highly virtual, it is produced and decays to quarks in a small space-time volume, $t \sim 1/\sqrt{s}$. On the other hand, the wavefunction of a hadron with mass $\sim m_{\rm had}$ has spatial extent $\sim 1/m_{\rm had}$ and hence the confinement of a quark pair into the hadron takes $t \sim 1/m_{\rm had}$. Thus there is no time for the confinement to affect the annihilation cross section and we expect

$$\sigma(e^+e^- \to \text{hadrons}) \approx \sigma(e^+e^- \to \text{quarks}), \tag{2.9}$$

and the Feynman rules do tell us how to calculate that.

In fact, we can go further than that and use an argument from quantum mechanics to postulate the form of the corrections to this approximation. Over a region of size $\sim 1/\sqrt{s}$, the amount by which the wave function of a hadron with spatial extent $\sim 1/m_{\rm had}$, could vary is $\sim m_{\rm had}/\sqrt{s}$ and the corrections should be at least this to some positive power,

$$\sigma(e^+e^- \to \text{hadrons}) = \sigma(e^+e^- \to \text{quarks}) \times \left(1 + \mathcal{O}\left(\frac{m_{\rm had}}{\sqrt{s}}\right)^n\right). \tag{2.10}$$

On the basis of the space-time picture, we can only justify that the corrections to the quark parton model are suppressed by some (positive) power of the ratio of scales. In practice, n is believed to be 6 for e^+e^- annihilation, making these corrections so small as to be almost impossible to measure. For most cross sections however, n is 2, and for jet cross sections, 1.

We calculated the cross section for $e^+e^- \to q\bar{q}$ in Section 1.6 and obtained

$$R_{e^+e^-} \equiv \frac{\sigma(\text{hadrons})}{\sigma(\text{muons})} = N_c \sum_q e_q^2, \tag{2.11}$$

where the sum over q is over all quark flavours that are kinematically allowed, i.e. for which $\sqrt{s} > 2m_q$. If we ignore effects close to threshold, such as the formation of bound states, we can expect a plot of $R_{e^+e^-}$ against $\sqrt{s}$ to present a series of steps at twice the quark masses and be flat in between. In principle one can read off the quark masses and charges from this plot.

Looking at the data in Fig. 2.2, we see that the general trend is as expected, but there are clearly corrections that are not accounted for by the quark parton model. One of these is the effect of higher

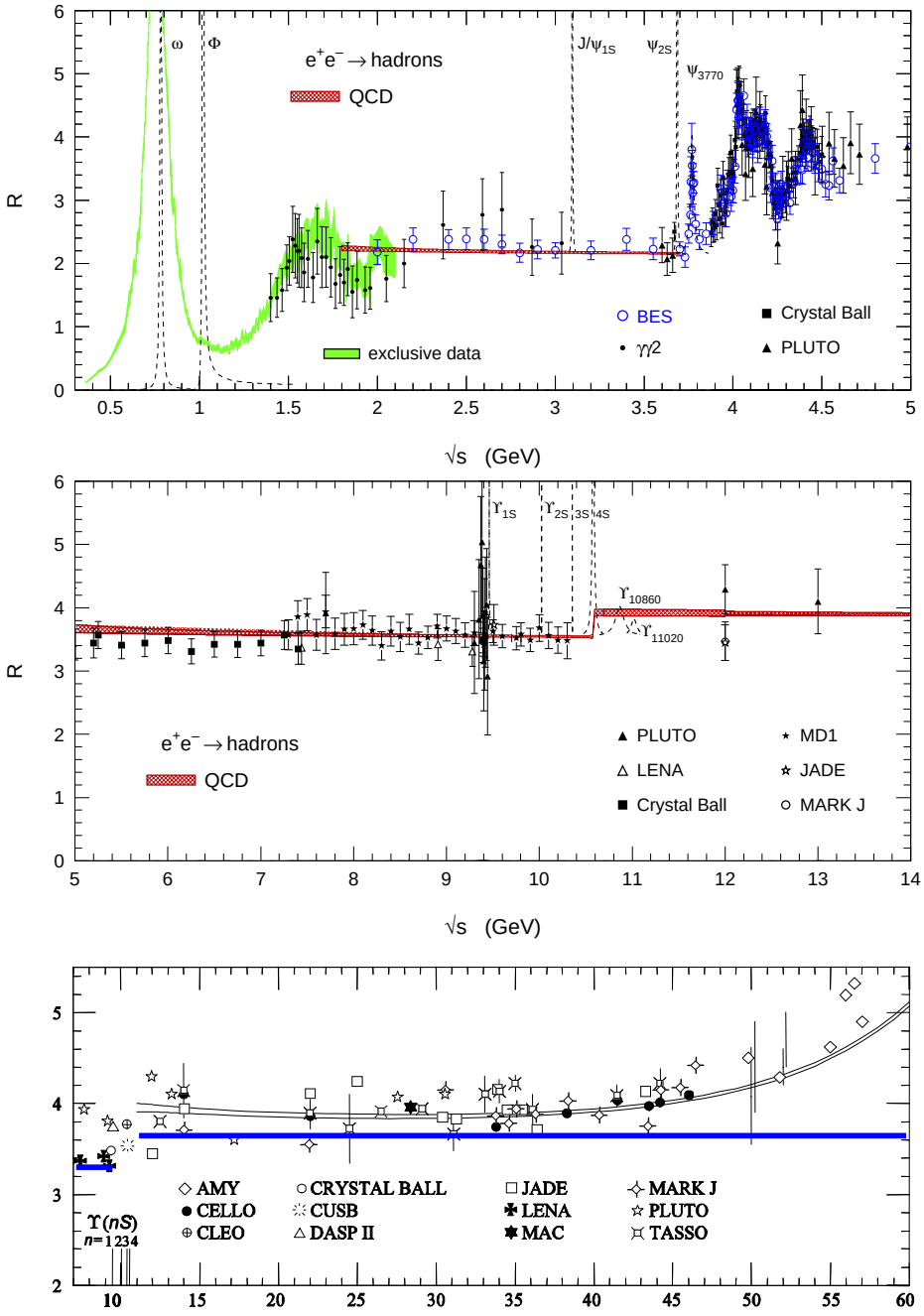


Fig. 2.2: Data on $R_{e^+e^-}$ as a function of centre-of-mass energy. Upper two panels taken from [4], lower from ESW [1]. The bands (red above, white below) show the QCD prediction, while the horizontal lines in the lower panel show the quark parton model expectations.

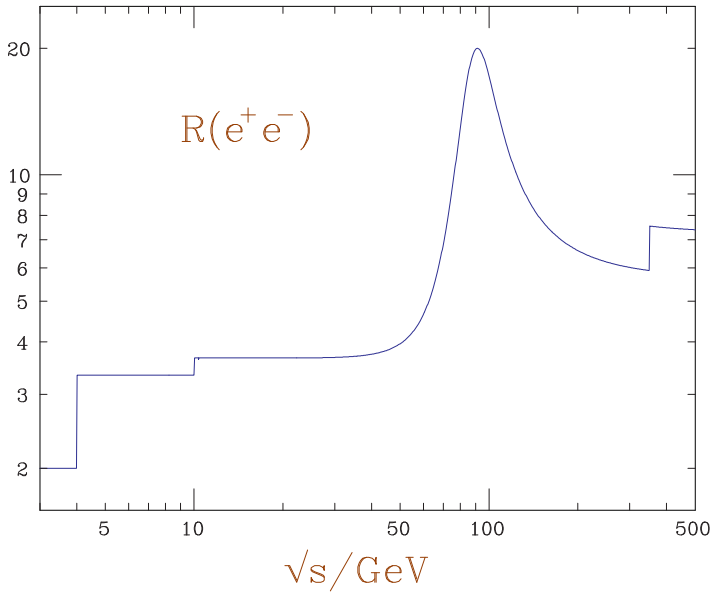


Fig. 2.3: Calculation of R_{had} as a function of centre-of-mass energy

order QCD corrections, which we include in the next lecture. Another is the effect of the Z^0 boson, which is clearly seen at the high energy end of Fig. 2.2, which we include shortly.

Before including the Z^0 contribution, it is worth remarking on a historical ambiguity that affects this figure. Although people wrote

$$R \equiv \frac{\sigma(e^+e^- \to \text{hadrons})}{\sigma(e^+e^- \to \mu^+\mu^-)}$$

they often didn't actually use that formula to show their experimental results, but rather

$$R \equiv \frac{\sigma(e^+e^- \to \text{hadrons})}{\frac{4\pi\alpha^2}{3s}},$$

using the leading order QED result for the denominator. Clearly many of the experimental and theoretical systematic errors would be smaller if the former was used, although of course the statistical errors would be larger, by around a factor of 2. More recent measurements, for example from LEP, have used the more honest notation in which the numerator and denominator are calculated or measured in the same way. This is sometimes called R_{had} to differentiate it from R.

In Fig. 2.3 I show the calculation of R_{had} in the quark parton model, including the Z^0 contribution. It is clear that γ–Z interference is important, even far from the Z peak. However, exactly on the peak the interference is zero (you might like to think about a simple explanation for why) and R_{had} is given to a good approximation by the Z contribution alone,

$$R_{had} = N_c \frac{\sum_q v_q^2 + a_q^2}{v_\mu^2 + a_\mu^2} = 20.095, \tag{2.12}$$

where v_i and a_i are the vector and axial couplings of the Z^0 to fermion type i. I note for future reference that the value including the photon contribution is 19.984. This number compares well with the LEP average measured value of 20.767 ± 0.025. However, the difference is still large on the scale of the experimental uncertainty, again indicating a clear need for the QCD corrections.

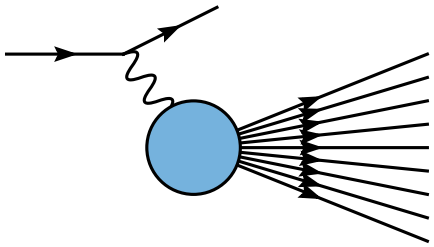


Fig. 2.4: Decay of τ lepton to hadrons

2.1.1 τ decays

We conclude this section by mentioning the closely-related process of τ decay to hadrons, depicted in Fig. 2.4. One can apply exactly the same arguments to the blob in this diagram as to annihilation of e^+e^- to hadrons. The only differences are that we have a virtual W boson producing hadrons instead of a virtual photon, and that we have an integral over all virtualities of the W between the τ mass and zero, rather than a single virtuality fixed by the beam energies. Nevertheless exactly the same arguments follow through and one obtains

$$R_\tau \equiv \frac{\mathcal{B}(\tau \to \text{hadrons})}{\mathcal{B}(\tau \to \mu)} = N_c \sum_{i,j} |V_{ij}|^2 \approx N_c, \tag{2.13}$$

where the sum is over the flavours of quark and antiquark that can appear in the W decay and V is the CKM matrix. Since a τ^- can decay to $\bar{u}d$ or $\bar{u}s$, to a good approximation this sum is $\cos^2\theta_C + \sin^2\theta_C$ and the final result follows.

We will see later that this process provides an excellent measurement of $\alpha_{\rm S}$.

2.2 Deep inelastic scattering

Historically, the quark model developed as a way of rationalizing the vast array of strongly-interacting particles that had been found by the 1960s. However, it was not clear whether quarks were really physical constituents of hadrons, or merely a convenient mathematical language to describe the hadrons' wave functions. The decisive evidence came from deep inelastic scattering experiments at SLAC. Today, deep inelastic scattering experiments give us by far the best information about the internal structure of the proton.

2.2.1 Quarks as partons in hadronic scattering

The classic probe of nuclear structure is electron–nucleus scattering. Assuming the scattering takes place by exchanging a single photon, measuring the kinematics of the scattered electron uniquely constrains that of the photon. The scattered electron has two non-trivial kinematic variables, its energy and scattering angle. These can more conveniently be converted into the photon virtuality ($Q^2 \equiv -q\cdot q$) and energy in the nucleus rest frame ν. Q^2 controls the resolving power of the photon, $Q^2 \sim 1/\lambda^2$. For fixed small $Q^2 \ll 1/R^2$, where R is the nuclear radius, the photon is absorbed elastically by the nucleus, giving a narrow peak in the ν distribution at $\nu = Q^2/2M_N$. For increased $Q^2 \sim 1/R^2$ one begins to resolve nuclear resonances as additional peaks at higher ν. Finally, for large $Q^2 \gg 1/R^2$, one resolves the proton constituents of the nucleus, with the photon being absorbed elastically by individual protons. These show up as a peak at $\nu = Q^2/2M_p$, broadened by the internal motion of the proton within the nucleus.

The scattering of electrons off hadrons, protons for example, is exactly analogous: at low Q^2 one sees only elastic proton scattering, but as Q^2 is increased, the photon can be elastically absorbed by the

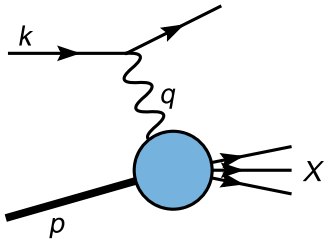


Fig. 2.5: Deep inelastic scattering

(charged) quark constituents of the proton. (Eventually at very large Q^2 and ν something new happens relative to the nuclear case, but we will not discuss that until the next lecture.)

We are interested in the region of Deep ($Q^2 \gg M_p^2$) Inelastic ($W^2 \gg M_p^2$, where W is the invariant mass of the photon–proton system) Scattering, DIS. We are therefore justified in neglecting the proton mass throughout, provided we do not work in the proton rest-frame, which is not well defined in that case. This is most conveniently done by working in terms of Lorentz-invariant variables.

2.2.2 Lorentz-invariant variables

It is convenient to describe this in terms of Lorentz-invariant variables. We label the momenta as shown in Fig. 2.5. For an electron of momentum k to scatter to one of momentum k' by exchanging a photon of momentum q with a proton of momentum p we again have, for fixed centre-of-mass energy s, only two independent kinematic variables,

$$s = (k+p)^2, \tag{2.14}$$
$$Q^2 = -q^2, \tag{2.15}$$
$$x = \frac{Q^2}{2p\cdot q}, \tag{2.16}$$

in terms of which we can calculate two other commonly-used variables

$$W^2 = (p+q)^2 = Q^2\frac{1-x}{x}, \tag{2.17}$$
$$y = \frac{p\cdot q}{p\cdot k} = \frac{Q^2}{xs}. \tag{2.18}$$

The kinematic limits are

$$Q^2 < s, \tag{2.19}$$
$$x > \frac{Q^2}{s}. \tag{2.20}$$

The coverage of the (x, Q^2) plane by the HERA, and earlier fixed target, DIS experiments is shown in Fig. 2.6

2.2.3 Structure functions

Since we do not yet know anything about the internal structure of protons, we cannot calculate the matrix element for the interaction of a photon with the proton to produce some arbitrary state X. However, like in the case of e^+e^- to hadrons we can get a surprisingly long way just by considering the properties that that matrix element must satisfy.

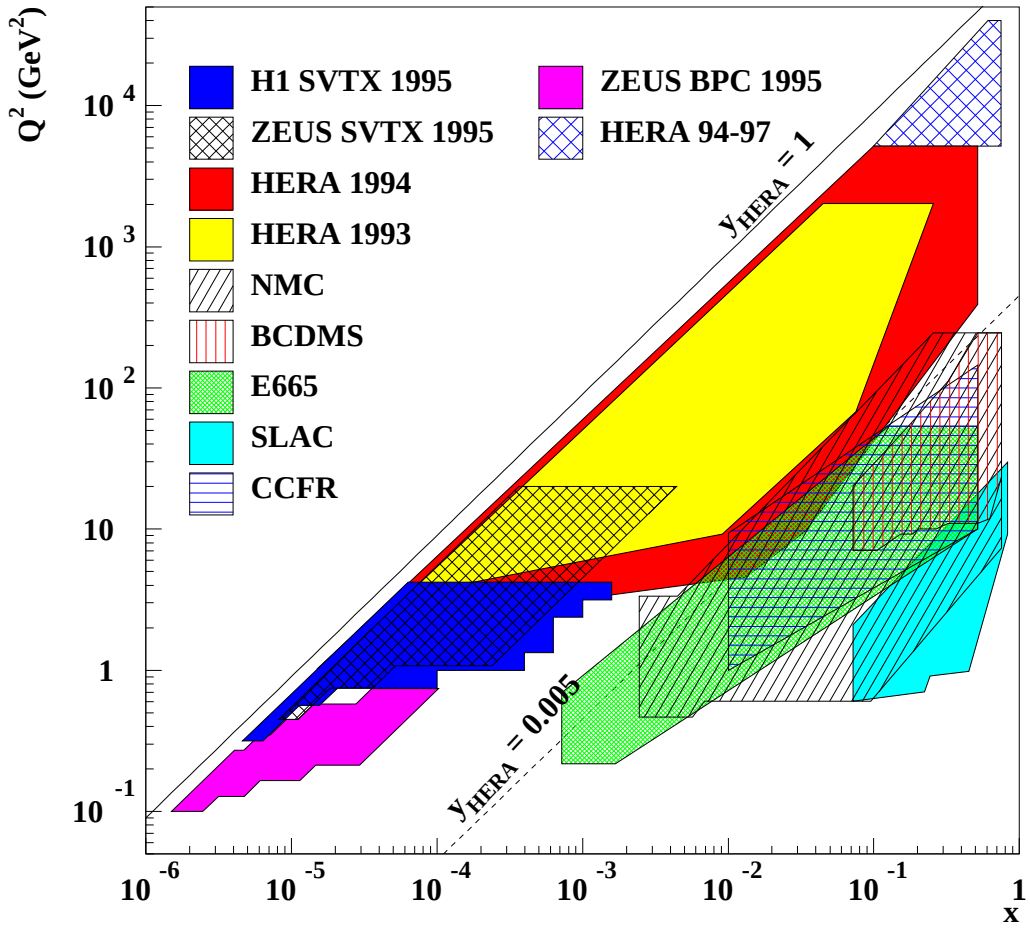


Fig. 2.6: The x–Q^2 plane, showing the coverage of measurements by various experiments

We parameterize the matrix element for a proton of momentum p to absorb a photon of momentum q and Lorentz index μ to produce an arbitrary set of hadrons X with fixed momenta $\{p_X\}$ as

$$e\,T_\mu(p,q;\{p_X\}). \tag{2.21}$$

We therefore have the matrix element squared for the whole process

$$\frac{1}{4}|\mathcal{M}|^2 = \frac{1}{4}\,\frac{e^4}{Q^4}\,\mathrm{Tr}\left\{\not{k}\gamma^\mu\not{k}'\gamma^\nu\right\}\,T_\mu(p,q;\{p_X\})\,T^*_\nu(p,q;\{p_X\}). \tag{2.22}$$

For convenience we define the Lorentz tensor

$$L^{\mu\nu} = \mathrm{Tr}\left\{\not{k}\gamma^\mu\not{k}'\gamma^\nu\right\}. \tag{2.23}$$

If the state X consists of n hadrons, then the $n+1$-body phase space for the whole process can be factorized into a part describing the electron kinematics times the n-body phase space for X,

$$dPS = \frac{Q^2}{16\pi^2 s x^2}\,dQ^2\,dx\,dPS_X. \tag{2.24}$$

This is as far as we can go for a specific state X, but we can get further by integrating over the phase space of X and summing over all possible states X. We define

$$\sum_X \int dPS_X\,\frac{1}{4}|\mathcal{M}|^2 \equiv \frac{e^4}{Q^4}\,L^{\mu\nu}H_{\mu\nu}, \tag{2.25}$$

or

$$\sum_X \int dPS_X \, T_\mu(p,q;\{p_X\}) \, T^*_\nu(p,q;\{p_X\}) = H_{\mu\nu}. \tag{2.26}$$

Since we have summed and integrated out all dependence on X, $H_{\mu\nu}$ can only depend on the vectors p and q. Since the electromagnetic and strong interactions conserve parity, it must be symmetric in μ and ν. There are only four possible symmetric two-index tensors that can be constructed from two vectors, so we can parameterize the hadronic tensor as a linear combination of them:

$$H_{\mu\nu} = -H_1 g_{\mu\nu} + H_2 \frac{p_\mu p_\nu}{Q^2} + H_4 \frac{q_\mu q_\nu}{Q^2} + H_5 \frac{p_\mu q_\nu + q_\mu p_\nu}{Q^2}, \tag{2.27}$$

where the Hs are scalar functions of the only two Lorentz scalars available $q{\cdot}q = -Q^2$ and $p{\cdot}q = Q^2/2x$, i.e., of x and Q^2 only (not s). (Note that we neglect $p{\cdot}p = M_p^2$ since we work in the limit $|q{\cdot}q|, p{\cdot}q \gg p{\cdot}p$.)

If we include Z^0 exchange (or charged current scattering) we can construct one further tensor, which is antisymmetric in μ and ν, $H_3 \, \epsilon_{\mu\nu\lambda\sigma} p^\lambda q^\sigma$, where $\epsilon_{\mu\nu\lambda\sigma}$ is the totally antisymmetric Lorentz tensor.

Contracting with $L^{\mu\nu}$ we find that H_4 and H_5 cannot contribute to physical cross sections (think about a simple explanation why not) and we have

$$L^{\mu\nu} H_{\mu\nu} = 4k{\cdot}k' \, H_1 + 4\frac{p{\cdot}k \; p{\cdot}k'}{Q^2} \, H_2. \tag{2.28}$$

Redefining (just a matter of convention) $H_1 = 4\pi F_1$ and $H_2 = 8\pi x F_2$, we obtain the final result for the scattering cross section

$$\frac{d^2\sigma}{dx \, dQ^2} = \frac{4\pi\alpha^2}{xQ^4} \left[y^2 x F_1(x,Q^2) + (1-y) F_2(x,Q^2) \right]. \tag{2.29}$$

Without knowing anything about the interactions of hadrons, we have been able to derive the s dependence of the scattering cross section for fixed x and Q^2 (which enters through the y dependence: recall $y = Q^2/xs$).

The Fs are called the structure functions of the proton. It is common to see other linear combinations of the structure functions,

$$F_T(x,Q^2) = 2xF_1(x,Q^2), \tag{2.30}$$

$$F_L(x,Q^2) = F_2(x,Q^2) - 2xF_1(x,Q^2), \tag{2.31}$$

which correspond to scattering of transverse and longitudinally polarized photons respectively. We therefore have

$$\frac{d^2\sigma}{dx \, dQ^2} = \frac{2\pi\alpha^2}{xQ^4} \left[(1+(1-y)^2) F_T(x,Q^2) + 2(1-y) F_L(x,Q^2) \right]. \tag{2.32}$$

In fact the most common form you will see this in nowadays is

$$\frac{d^2\sigma}{dx \, dQ^2} = \frac{2\pi\alpha^2}{xQ^4} \left[(1+(1-y)^2) F_2(x,Q^2) - y^2 F_L(x,Q^2) \right]. \tag{2.33}$$

For the majority of current data, y^2 is small and F_L can be neglected: only close to the kinematic limit, or for very precise data, need it be considered.

We have isolated all the non-trivial x and Q^2 dependence into the two functions $F_2(x,Q^2)$ and $F_L(x,Q^2)$, but we still have no idea how those functions behave. If we make the assumption that the

Fig. 2.7: In the Breit frame, the proton of diameter $2R$ is contracted to a pancake of thickness $4RxM_p/Q$ (a) so that a photon of high virtuality Q interacts incoherently with a single parton within it (b)

interaction of the photon with the innards of the proton does not involve any dimensionful scale, then we immediately get the result that the dimensionless Fs cannot depend on the dimensionful Q^2 and we get

$$\frac{d^2\sigma}{dx\,dQ^2} = \frac{2\pi\alpha^2}{xQ^4}\left[(1+(1-y)^2)F_2(x) - y^2F_L(x)\right], \tag{2.34}$$

known as Bjorken scaling. Experimentally this is true to a pretty good approximation, but given that the proton is supposed to consist of quarks, bound together with a distance scale $\sim 1/M_p$, how can the interaction possibly be M_p-independent? The answer to this lies in the parton model.

2.2.4 Parton distribution functions and Bjorken scaling

Although it is of course Lorentz-invariant, the parton model is most easily formulated in a frame in which the proton is fast moving. Most convenient is the so-called Breit frame, in which the photon has zero energy and collides head-on with the proton. In this frame, the proton energy is $Q/2x$. Assuming that in its own restframe it is a sphere of radius R, in the Breit frame it is massively Lorentz contracted to a flat pancake, still with transverse diameter $2R$, but with length $4RxM_p/Q \ll 2R$, as illustrated in Fig. 2.7a. The transverse size of the photon is $\sim 1/Q \ll 2R$. The photon therefore interacts with a tiny fraction of a thin disk, so provided that the quarks are sufficiently dilute the photon is not able to resolve the quarks' interactions and they act as if they were free. That is, the photon effectively collides with a single free quark, as illustrated in Fig. 2.7b.

Since they act as if they do not interact, their interactions do not introduce a dimensionful scale, and so the structure functions will obey Bjorken scaling.

More precisely, we suppose that the proton consists of a bundle of comoving partons, which carry a range of the proton's momentum. We posit probability distribution functions (more often called parton distribution functions, pdfs), such that partons of type q carry a fraction of the proton's momentum between η and $\eta+d\eta$ a fraction $f_q(\eta)d\eta$ of the time. Provided that these partons are pointlike $r^2 \ll 1/Q^2$ and dilute $f_q(\eta) \ll Q^2R^2$, the photons will scatter incoherently off individual partons. The cross section can then be factorized as the convolution of the pdfs with the cross section for parton scattering,

$$\frac{d^2\sigma(e+p(p))}{dx\,dQ^2} = \sum_q \int_0^1 d\eta\, f_q(\eta)\frac{d^2\sigma(e+q(\eta p))}{dx\,dQ^2}. \tag{2.35}$$

We will calculate the partonic cross section shortly, but first let me point out a couple of features it must have.

Firstly if we assume that the scattering is elastic, then the outgoing parton must be on mass-shell. Since we are then considering a two-to-two collision, which has only one nontrivial kinematic variable, the double-differential cross section in x and Q^2 must be proportional to a δ function fixing one of the

variables. Specifically, if we assume that the partons are massless, then we obtain the relation

$$(q+\eta p)^2 = 2\eta\, p\cdot q - Q^2 = 0, \tag{2.36}$$

or

$$\eta = x. \tag{2.37}$$

Secondly if we assume that the struck partons are the quarks of the quark model, they must be fermions. Simply from helicity conservation, we can then show that $F_L = 0$. This is known as the Callan–Gross relation and was one of the first proofs that the quarks of the quark model really were the partons of the parton model. (If the partons were instead scalars we would have $F_T = 0$ and hence completely different y-dependence of the cross section.)

2.2.5 Scattering cross sections

To calculate the parton model prediction for the structure functions, we need the matrix elements for $eq \to eq$. These can be obtained by crossing symmetry from those for $e^+e^- \to q\bar{q}$. That is,

$$\sum |\mathcal{M}|^2 = 8(4\pi\alpha)^2\, e_q^2 N_c\, \frac{(p_e\cdot p_q)^2 + (p_e\cdot p_q')^2}{(p_e\cdot p_e')^2}. \tag{2.38}$$

Converting to the kinematic variables we defined earlier, we have

$$\sum |\mathcal{M}|^2 = 8(4\pi\alpha)^2\, e_q^2 N_c\, \frac{1+(1-y)^2}{y^2}. \tag{2.39}$$

Using (2.24), we have

$$dPS = \frac{Q^2}{16\pi^2 s x^2} dQ^2\, dx\, dPS_X. \tag{2.40}$$

Since X consists only of one massless parton, we have

$$\begin{aligned} dPS_X &= \frac{d^4p_X}{(2\pi)^3}\delta(p_X^2)\,(2\pi)^4\delta^4(\eta p + q - p_X) && (2.41)\\ &= (2\pi)\delta((\eta p+q)^2) && (2.42)\\ &= \frac{2\pi x}{Q^2}\delta(\eta - x). && (2.43)\end{aligned}$$

The full cross section is therefore

$$\begin{aligned} \frac{d\sigma}{dx\, dQ^2} &= \frac{1}{4N_c}\,\frac{1}{2\hat{s}}\,\frac{Q^2}{16\pi^2 s x^2}\,\frac{2\pi x}{Q^2}\,\delta(x-\eta)\,\sum|\mathcal{M}|^2 && (2.44)\\ &= \frac{1}{4N_c}\,\frac{y^2}{16\pi Q^4}\,\delta(x-\eta)\,\sum|\mathcal{M}|^2, && (2.45)\end{aligned}$$

where the factor of $1/N_c$ is the average over incoming colours. We therefore have

$$\frac{d\sigma(e+q)}{dx\, dQ^2} = \frac{2\pi\alpha^2}{Q^4}\,\delta(x-\eta)\; e_q^2\,\left(1+(1-y)^2\right) \tag{2.46}$$

and hence

$$\frac{d\sigma(e+p)}{dx\, dQ^2} = \frac{2\pi\alpha^2}{xQ^4}\left(1+(1-y)^2\right)\sum_q e_q^2\, x f_q(x). \tag{2.47}$$

Comparing (2.47) with (2.33) we therefore have

$$F_2(x,Q^2) = \sum_q e_q^2\, x f_q(x), \tag{2.48}$$

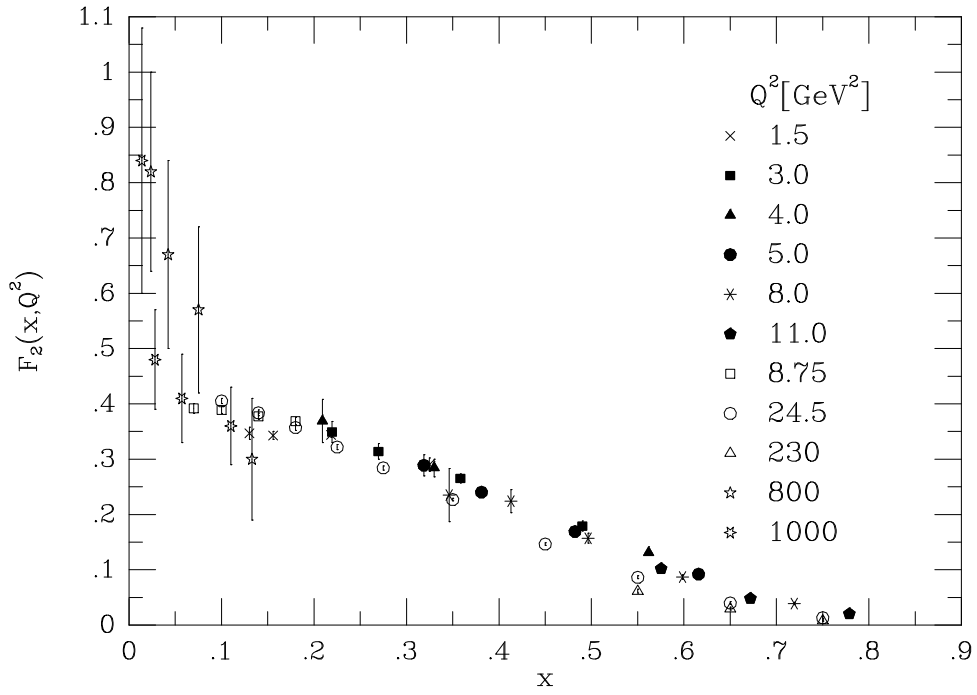


Fig. 2.8: The structure function F_2 as a function of x for various Q^2 values, exhibiting Bjorken scaling, taken from ESW [1]

$$F_L(x,Q^2) \;=\; 0. \tag{2.49}$$

Note that F_2 is Q^2-independent, showing Bjorken scaling.

Although we will see that QCD corrections do violate Bjorken scaling, it is satisfied pretty well by the data, as can be seen in Fig. 2.8.

2.2.6 Charged current neutrino DIS

We can consider charged current neutrino scattering in exactly the same way. Since the scattering takes place by the weak interaction, parity is violated, allowing one additional Lorentz structure,

$$L^{\stackrel{\nu}{\bar\nu}}_{\mu\nu} \;=\; L^e_{\mu\nu} \pm 2i\epsilon_{\mu\nu\rho\sigma}k^\rho k'^\sigma, \tag{2.50}$$

$$H^{\mu\nu} \;=\; -H_1 g^{\mu\nu} + H_2\frac{p^\mu p^\nu}{Q^2} - \frac{i}{Q^2}\epsilon^{\mu\nu\rho\sigma}p_\rho q_\sigma H_3, \tag{2.51}$$

$$\Rightarrow L^{\stackrel{\nu}{\bar\nu}}_{\mu\nu}H^{\mu\nu} \;=\; 2Q^2H_1 + Q^2\frac{1-y}{x^2y^2}H_2 \pm \frac{Q^2}{xy}H_3\,(1-y/2). \tag{2.52}$$

Thus, defining $H_3 = 8\pi x F_3$, we have a third structure function F_3:

$$\frac{d^2\sigma(^{\nu}_{\bar\nu}+p)}{dx\,dQ^2} = \frac{G_F^2}{4\pi x}\left(\frac{M_w^2}{Q^2+M_w^2}\right)^2\left[\left(1+(1-y)^2\right)F_2^{\stackrel{\nu}{\bar\nu}} - y^2F_L^{\stackrel{\nu}{\bar\nu}} \pm \left(1-(1-y)^2\right)xF_3^{\stackrel{\nu}{\bar\nu}}\right], \tag{2.53}$$

where G_F is the Fermi constant and M_w the W boson mass. In the parton model we have

$$F_2^{\stackrel{\nu}{\bar\nu}}(x,Q^2) \;=\; \sum_q 2xf_q(x) + \sum_{\bar q} 2xf_{\bar q}(x), \tag{2.54}$$

$$xF_3^{\stackrel{\nu}{\bar\nu}}(x,Q^2) \;=\; \sum_q 2xf_q(x) - \sum_{\bar q} 2xf_{\bar q}(x), \tag{2.55}$$

where the sums for neutrino scattering are over all partons that can absorb a W^+, i.e., d, s, $\bar{\rm u}$ and $\bar{\rm c}$ and for antineutrino over those that can absorb a W^-, i.e., u, c, $\bar{\rm d}$ and $\bar{\rm s}$.

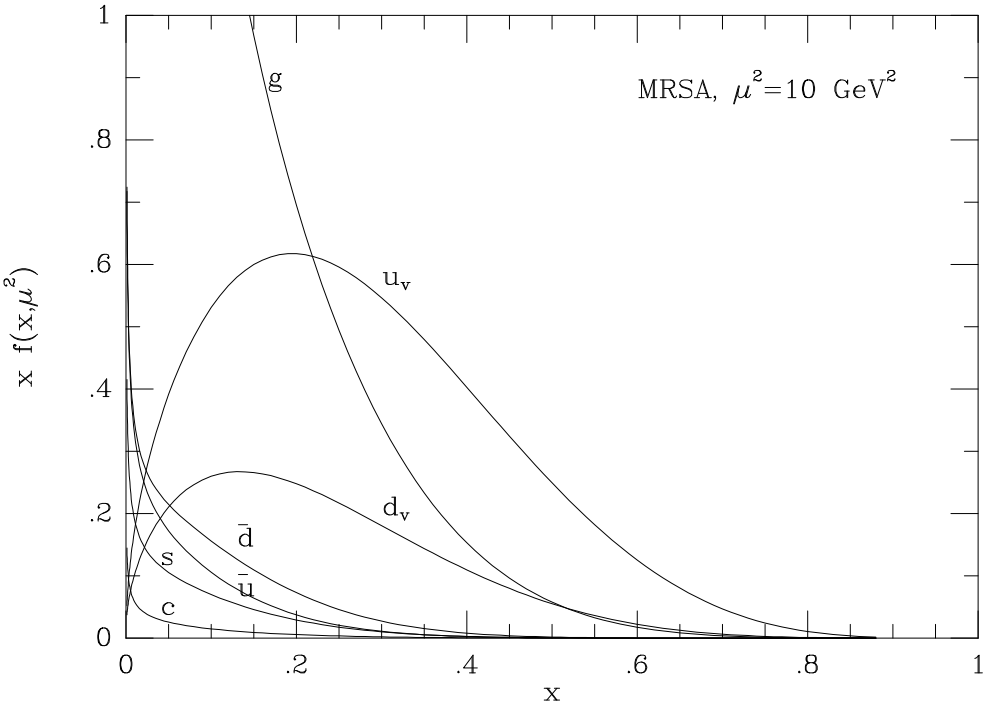


Fig. 2.9: Parton distribution function set A from the Martin-Roberts-Stirling group, taken from ESW [1]

2.2.7 Global fits

It is also possible to measure DIS on the neutron, or at least on deuterium from which the neutron structure functions can be derived. Using strong isospin symmetry, we have the relations

$$f_{u/n}(x) = f_{d/p}(x), \tag{2.56}$$
$$f_{\bar{u}/n}(x) = f_{\bar{d}/p}(x), \tag{2.57}$$
$$f_{d/n}(x) = f_{u/p}(x), \tag{2.58}$$
$$f_{s/n}(x) = f_{s/p}(x), \tag{2.59}$$

and so on. It is conventional to always refer to the proton case, dropping the "$/p$" subscript. We therefore have the slightly confusing result for F_2^{en} shown below, in which f_d is multiplied by $(2/3)^2$, and so on.

We therefore have

$$F_2^{ep} = \tfrac{1}{9}xf_d + \tfrac{4}{9}xf_u + \tfrac{1}{9}xf_{\bar{d}} + \tfrac{4}{9}xf_{\bar{u}} + \tfrac{1}{9}xf_s + \tfrac{1}{9}xf_{\bar{s}} + \tfrac{4}{9}xf_c + \tfrac{4}{9}xf_{\bar{c}}, \tag{2.60}$$
$$F_2^{en} = \tfrac{4}{9}xf_d + \tfrac{1}{9}xf_u + \tfrac{4}{9}xf_{\bar{d}} + \tfrac{1}{9}xf_{\bar{u}} + \tfrac{1}{9}xf_s + \tfrac{1}{9}xf_{\bar{s}} + \tfrac{4}{9}xf_c + \tfrac{4}{9}xf_{\bar{c}}, \tag{2.61}$$
$$F_2^{\nu p} = 2xf_d + 2xf_{\bar{u}} + 2xf_s + 2xf_{\bar{c}}, \tag{2.62}$$
$$xF_3^{\nu p} = 2xf_d - 2xf_{\bar{u}} + 2xf_s - 2xf_{\bar{c}}, \tag{2.63}$$
$$F_2^{\bar{\nu}p} = 2xf_u + 2xf_{\bar{d}} + 2xf_c + 2xf_{\bar{s}}, \tag{2.64}$$
$$xF_3^{\bar{\nu}p} = 2xf_u - 2xf_{\bar{d}} + 2xf_c - 2xf_{\bar{s}}. \tag{2.65}$$

If we make the assumption that $f_{\bar{s}} = f_s$ and $f_{\bar{c}} = f_c$, then we have six unknowns for six pieces of data so, given precise enough data, we could solve for all the pdfs exactly. In practice of course it is never so simple and one must make global fits to as wide a variety of data as possible.

One gets typical results like those shown in Fig. 2.9. Note that this uses the common notation of defining valence quark distributions,

$$f_{u_v} \equiv f_u - f_{\bar{u}}, \tag{2.66}$$

$$f_{d_v} \equiv f_d - f_{\bar{d}}. \tag{2.67}$$

Non-valence quarks are generically referred to as the sea.

2.2.8 Sum rules

Having results for the pdfs, one can form interesting integrals over them, for example,

$$\int_0^1 dx\, f_{u_v}(x) = 2, \tag{2.68}$$

$$\int_0^1 dx\, f_{d_v}(x) = 1. \tag{2.69}$$

Various such integrals can be constructed directly from the structure functions. It is worth checking that you can reproduce the physical interpretation of each.

2.2.8.1 The Gross–Llewellyn-Smith sum rule

$$\tfrac{1}{2}\int_0^1 dx \left(F_3^{\nu p} + F_3^{\bar{\nu} p}\right) = 3, \tag{2.70}$$

which counts the number of valence quarks in the proton. In QCD this provides a useful measurement of $\alpha_{\rm S}$, because the right-hand side is actually equal to $3\left(1 - \frac{\alpha_{\rm S}}{\pi} + \mathcal{O}(\alpha_{\rm S}^2)\right)$.

2.2.8.2 The Adler sum rule

$$\tfrac{1}{2}\int_0^1 \frac{dx}{x} \left(F_2^{\bar{\nu} p} - F_2^{\nu p}\right) = 1, \tag{2.71}$$

which counts the difference between the number of up and down valence quarks. This has the property that it is exact even in QCD, i.e., all higher order corrections vanish.

2.2.8.3 The Gottfried sum rule

$$\int_0^1 \frac{dx}{x} \left(F_2^{ep} - F_2^{en}\right) \approx 0.23, \tag{2.72}$$

where the result is experimental. This is sensitive to the difference between the number of up and down sea quarks: it would be 1/3 if they were equal.

2.2.8.4 The momentum sum rule

Finally, we have the particularly significant result

$$\tfrac{1}{2}\int_0^1 dx \left(F_2^{\nu p} + F_2^{\bar{\nu} p}\right) \approx 0.5, \tag{2.73}$$

where the result is again experimental. This tells us that only about half of the proton's momentum is carried by quarks and antiquarks.

2.3 Hadronic collisions

2.3.1 The Drell–Yan process

If the parton model is correct, the parton distribution functions should be universal. We should therefore be able to use the DIS measurements to make predictions for other hadronic scattering processes. The classic example is the so-called Drell–Yan process, of lepton pair production in hadron collisions,

$$h_1 + h_2 \to \mu^+ + \mu^- + X, \tag{2.74}$$

where the state X goes unmeasured. In the parton model this arises as the sum over all quark types of

$$q + \bar{q} \to \mu^+ + \mu^-. \tag{2.75}$$

The cross section can be written as the convolution of pdfs with a partonic cross section, exactly like in DIS:

$$\frac{d\sigma(h_1(p_1) + h_2(p_2) \to \mu^+\mu^-)}{dM^2} = \sum_q \int_0^1 d\eta_1 f_{q/h_1}(\eta_1) \int_0^1 d\eta_2 f_{\bar{q}/h_2}(\eta_2) \frac{d\sigma(q(\eta_1 p_1) + \bar{q}(\eta_2 p_2) \to \mu^+\mu^-)}{dM^2}, \tag{2.76}$$

where M is the mass of the $\mu^+\mu^-$ pair. Note that since the partonic cross section contains a $\delta(M^2 - \eta_1\eta_2 s)$ term, binning the data in M gives extra information about the pdfs. In fact, binning also in the rapidity of the lepton pair, defined by

$$y \equiv \frac{1}{2} \ln \frac{E_{\mu^+\mu^-} + p_{z,\mu^+\mu^-}}{E_{\mu^+\mu^-} - p_{z,\mu^+\mu^-}}, \tag{2.77}$$

both η values are fixed, providing a direct measurement of the parton distribution functions (the partonic cross section can easily be obtained by crossing the $e^+e^- \to q\bar{q}$ one we calculated in Section 1.6, divided by a factor of N_c^2 for the average over incoming colours):

$$\frac{d^2\sigma}{dM^2 dy} = \frac{4\pi\alpha^2}{3N_c M^2 s} \sum_q e_q^2 f_{q/h_1}(e^y M/\sqrt{s}) f_{\bar{q}/h_2}(e^{-y} M/\sqrt{s}). \tag{2.78}$$

Note that the case $h_1 = h_2 = p$ provides a particularly good measure of the sea quark distribution functions, which are hard to extract from DIS data.

2.3.2 Prompt photon and jet production

Although we have not yet mentioned gluons, we will see in the next lecture that there is also a non-zero pdf for the gluon, $f_g(\eta)$, as can also be inferred from the momentum sum rule mentioned earlier. As well as being important for higher order corrections to the processes given above, there are many processes in which they participate at tree level. The most important of these are prompt photon production,

$$h_1 + h_2 \to \gamma + X, \tag{2.79}$$

and jet production

$$\begin{aligned} h_1 + h_2 &\to q + q + X, && (2.80)\\ h_1 + h_2 &\to q + \bar{q} + X, && (2.81)\\ h_1 + h_2 &\to q + g + X, && (2.82)\\ h_1 + h_2 &\to g + g + X, \text{ etc.} && (2.83) \end{aligned}$$

The gluon pdf is used in exactly the same way as the quark ones, and hadronic cross sections can still be calculated as the sum of convolutions of pdfs with partonic cross sections. Prompt photon production receives contributions from two partonic processes,

$$\begin{aligned} q + \bar{q} &\to \gamma + g, && (2.84)\\ q + g &\to \gamma + q. && (2.85) \end{aligned}$$

In the case $h_1 = h_2 = p$, the latter dominates, providing a measure of the gluon pdf. However there is a slight complication, in that processes (2.84), (2.85) are proportional to $\alpha_{\rm S}$, which is less well-known than α, which controls the other processes we have studied. In fact this is always the case, that measurements of the gluon pdf actually measure $\alpha_{\rm S} \times f_g$ in general. The QCD corrections to this process turn out to be a lot larger than any of the others we have considered, further complicating this measurement.

2.4 Summary

We have considered the tree-level phenomenology of e^+e^- annihilation, deep inelastic scattering and, more briefly, hadron collisions. It is remarkable how much QCD phenomenology can be understood using tree level results. However, we have to worry that $\alpha_{\rm S}$ is not so small, so higher order corrections must be important. Equally importantly, it would be nice to see whether, and if so how, the parton model emerges from QCD.

We discuss both these issues in the next lecture.

3 Higher order corrections

3.1 e^+e^- annihilation at one loop

In this section, I go through the calculation of the NLO correction to the $e^+e^- \to$ hadrons cross section in some detail. I will briefly describe some of the more technical aspects of the calculation, for those interested, in Section 3.1.2, but those who are not can safely skip this section, since I recap the important results at the start of Section 3.1.3.

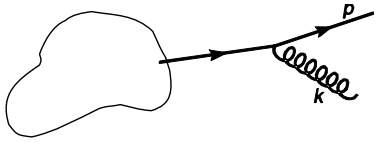
In discussing the $e^+e^- \to$ hadrons cross section at tree level, we assumed that quarks produce hadrons with probability 1. Therefore we calculated the $e^+e^- \to q\bar{q}$ cross section in Section 1.6. In discussing jet cross sections, we extended this to say that all partons produce hadrons with probability 1. Therefore we should calculate the total cross section to produce any number or type of partons. At leading order this makes no difference, since the only possible process is $e^+e^- \to q\bar{q}$, but at order $\alpha_{\rm S}$ we have to calculate and sum the cross sections for $q\bar{q}$ and $q\bar{q}g$ final states. We start with the latter.

Recall that the total $q\bar{q}g$ cross section is divergent,

$$\sigma = \sigma_0 \, C_F \frac{\alpha_s}{2\pi} \int dx_1 \, dx_2 \frac{x_1^2 + x_2^2}{(1-x_1)(1-x_2)}, \tag{3.1}$$

where the region of integration is the upper right triangle of the unit square, bordered by the lines $x_1 = 1$ and $x_2 = 1$, which are the singular regions. This divergence must be regularized in some way, before we can make progress.

First though we discuss the origin of the divergences. They arise from propagator factors that diverge,



$$\frac{1}{(p+k)^2} = \frac{1}{2p\cdot k} = \frac{1}{2E\omega(1-\cos\theta)} \approx \frac{1}{E\omega\theta^2}, \tag{3.2}$$

where E and ω are the quark and gluon energies and θ is the angle between them.

In the collinear limit, $\theta \to 0$, one in principle obtains $1/\theta^4$ in the matrix element squared, but in fact the numerators always contribute a factor of θ^2, so one obtains

$$|\mathcal{M}|^2 \sim \frac{1}{\theta^2}. \tag{3.3}$$

In the soft limit, $\omega \to 0$, one has in the interference between diagrams in which the gluon is attached to quark 1 and quark 2,

$$|\mathcal{M}|^2 \sim \frac{p_1\cdot p_2}{p_1\cdot k \, p_2\cdot k} \sim \frac{1}{\omega^2}. \tag{3.4}$$

In terms of ω and θ the phase space is given by

$$\frac{d^3k}{2\omega} = \tfrac{1}{2}\,\omega d\omega \, \sin\theta \, d\theta \, d\phi \sim \omega d\omega \, \theta d\theta. \tag{3.5}$$

We therefore have logarithmic singularities in both the soft and collinear limits. We generically refer to both of these as the infrared limit.

3.1.1 Regularization

As in the discussion of renormalization, the simplest way we could regularize this cross section is with a cutoff, for example on the transverse momentum of the gluon, which would prevent the integration entering both the soft and collinear regions. However, we will see that infrared singularities cancel between different contributions, in this case $q\bar{q}$ and $q\bar{q}g$, so we must use a regularization that can be consistently applied in all contributions. It is not clear that this is the case for a cutoff, since it must be applied in both real and virtual contributions, which have very different structures. Instead, to ensure consistent application across all processes, it is better to modify the theory in such a way that some dimensionless parameter ϵ regulates the divergences. Then the complete calculation can be performed in this modified theory and at the end of the calculation, when all the divergences have cancelled, the limit $\epsilon \to 0$ can be smoothly taken. Remarkably, dimensional regularization, which we used for ultraviolet singularities, also provides a consistent regulator for infrared singularities, as we shall discuss in detail shortly.

Another regularization scheme, which actually works well in QED, and for simple processes in QCD, is the gluon (or photon) mass regularization. We introduce a non-zero gluon mass $m_g^2 = \epsilon Q^2$. This prevents the propagators from reaching zero and diverging: for massless quarks the minimum value is m_g^2 and for a quark of mass m_q it is $2m_q m_g$. With this modification one can recalculate the differential cross section and integrate it to give a finite result,

$$\sigma_{q\bar{q}g} = \sigma_0\, C_F \frac{\alpha_s}{2\pi} \left(\log^2 \frac{1}{\epsilon} - 3 \log \frac{1}{\epsilon} + 7 - \frac{\pi^2}{3} + \mathcal{O}(\epsilon) \right). \tag{3.6}$$

However, since a non-zero gluon mass violates gauge invariance, this method is bound to fail in general. In particular, it is not suitable for any process in which any lowest order contributions have external gluons. As in the ultraviolet case, the only scheme that is known to be consistent with all the symmetries of QCD, and hence to work to arbitrary orders in arbitrary processes, is dimensional regularization.

The reason why I said that it is remarkable that dimensional regularization works in the infrared limit is the fact that the two limits have non-overlapping regions of applicability in the complex d plane. Ultraviolet-singular integrals are regularized by working in $d < 4$ dimensions, but infrared-singular integrals are only rendered finite by working in $d > 4$ dimensions. However, by carefully splitting contributions that are singular in both the infrared and ultraviolet one can consider the regularization schemes that are used in each as independent. In each region, one considers the appropriate dimensionality ($d = 4 - 2\epsilon$ with $\epsilon > 0$ in the ultraviolet and with $\epsilon < 0$ in the infrared) and then analytically continues to the whole complex ϵ plane. Since analytical continuation is unique, this gives a unique result for each, in the region of applicability of the other, and the two can be combined before the limit $\epsilon \to 0$ is taken. This subtlety leads to some surprising results, for example for the self-energy of a massless quark, discussed below.

As the calculation of cross sections in dimensional regularization is rather technical, it is rare to see it done in summer school lectures, but I think it brings out some interesting points, so I at least sketch how the calculation works in Section 3.1.2. As I said, those who disagree can safely skip ahead to Section 3.1.3.

3.1.2 Aside: Real and virtual corrections in dimensional regularization

It is straightforward to generalize the Feynman rules to d dimensions and fairly straightforward to generalize the Dirac algebra. The result is that d-dimensional matrix elements still have propagators $\sim 1/p^2$, but that the numerators become d dependent. (It is worth mentioning the closely-related dimensional reduction scheme, which is often used for supersymmetry calculations, since conventional dimensional regularization violates supersymmetry. In this scheme one works in d dimensions, but modifies the theory in such a way that fermions and massless vector bosons still have 2 spin states, instead of $d-2$ as in dimensional regularization. The result is that the matrix elements themselves are equal to the 4-

dimensional ones and it is only on performing the loop and phase space integrals that the d dimensionality gets introduced.)

3.1.2.1 Phase space integrals

We will have to integrate over d-dimensional phase space. We begin by considering integer values of d and then continue the results to real values. It is straightforward to write down the basic integration measure,

$$d^d k\, \delta_+(k^2) = \frac{d^{d-1}k}{2\omega} = \frac{1}{2}\omega^{d-3} d\omega\, d\Omega_{d-2}, \tag{3.7}$$

where ω is the energy of k and $d\Omega_{d-2}$ is an element of $d-2$-dimensional solid angle. The only difficulty concerns the evaluation of integrals over this solid angle. In four dimensions we have

$$k = \omega(1; \sin\phi\, \sin\theta, \cos\phi\, \sin\theta, \cos\theta) \qquad \text{(4 dimensions)}, \tag{3.8}$$

where θ and ϕ are the usual spherical polar coordinates with θ the polar angle and ϕ the azimuthal angle. In five dimensions we have

$$k = \omega(1; \sin\psi\, \sin\phi\, \sin\theta, \cos\psi\, \sin\phi\, \sin\theta, \cos\phi\, \sin\theta, \cos\theta) \qquad \text{(5 dimensions)}, \tag{3.9}$$

where ψ is an azimuthal angle in the additional dimension. Generalizing to d dimensions, we have $d-4$ additional azimuths and we write k generically as

$$k = \omega(1; \ldots, \cos\phi\, \sin\theta, \cos\theta) \qquad (d \text{ dimensions}), \tag{3.10}$$

where the ellipsis represents a $d-3$-vector of length $\sin\phi\, \sin\theta$ containing $d-4$ azimuths. Depending on the complexity of the calculation, more or less of these additional components have to be specified precisely. In fact in our case, since we only consider the relative orientations of three momenta that have zero total momentum, and therefore all lie in a plane, it is sufficient to specify

$$k = \omega(1; \ldots, \cos\theta) \qquad (d \text{ dimensions}), \tag{3.11}$$

where the ellipsis represents a $d-2$-vector of length $\sin\theta$ containing $d-3$ azimuths.

We can see how to integrate over the additional azimuths by again considering integer d and then generalizing,

$$\int d\Omega_1 = \int d\phi = 2\pi, \tag{3.12}$$

$$\int d\Omega_2 = \int d\phi\, \sin\theta\, d\theta = 4\pi, \tag{3.13}$$

$$\int d\Omega_3 = \int d\psi\, \sin\phi\, d\phi\, \sin^2\theta\, d\theta = 2\pi^2, \tag{3.14}$$

and so on. We have a recursion relation

$$\int d\Omega_n = \int d\Omega_{n-1}\, \sin^{n-1}\theta\, d\theta, \tag{3.15}$$

which is solved by

$$\Omega_n \equiv \int d\Omega_n = \frac{2\pi^{(n+1)/2}}{\Gamma[(n+1)/2]}. \tag{3.16}$$

We are now equipped to tackle the phase space integral, and see how the dimensional regularization succeeds in regularizing our integrals.

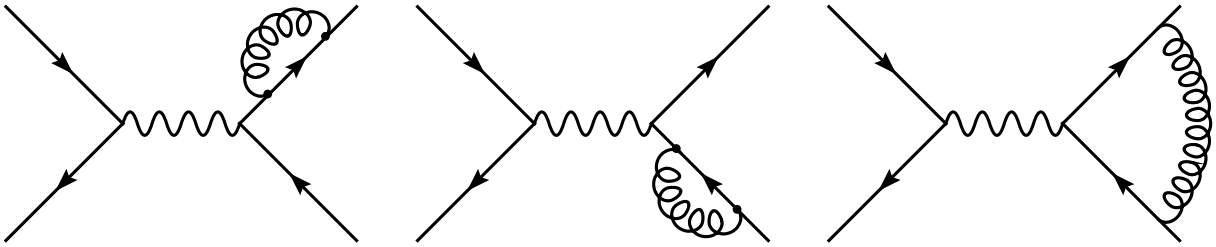


Fig. 3.1: One-loop diagrams for $e^+e^- \to q\bar{q}$

3.1.2.2 Regularization

Since the form of the propagator factors is unchanged in d dimensions, and it is these that dominate the singular region, it is straightforward to read off the behaviour in the regularized theory. In the soft region we have

$$\int_0 \omega^{1-2\epsilon} d\omega \, \frac{1}{\omega^2} = \int_0 \frac{d\omega}{\omega^{1+2\epsilon}} \sim -\frac{1}{2\epsilon}, \qquad \epsilon < 0, \tag{3.17}$$

and in the collinear region

$$\int_0 \sin^{1-2\epsilon}\theta \, d\theta \, \frac{1}{\theta^2} \sim \int_0 \frac{d\theta}{\theta^{1-2\epsilon}} \sim -\frac{1}{2\epsilon}, \qquad \epsilon < 0. \tag{3.18}$$

Since our cross section is divergent in both limits, and they can overlap, i.e., a radiated gluon can be both soft and collinear, we expect the total cross section to be of order $1/\epsilon^2$. Note, as a consistency check, that the integrands are positive definite and that, in the region in which they are well-defined, $\epsilon < 0$, the results are positive (and divergent as $\epsilon \to 0$).

3.1.2.3 Total $e^+e^- \to q\bar{q}g$ cross section

We now have all the ingredients we need to calculate the differential cross section for $e^+e^- \to q\bar{q}g$ and to integrate it over all phase space in dimensional regularization. We obtain

$$\sigma_{q\bar{q}g} = \sigma_0 \, C_F \frac{\alpha_s}{2\pi} H(\epsilon) \left(\frac{2}{\epsilon^2} + \frac{3}{\epsilon} + \frac{19}{2} - \pi^2 + \mathcal{O}(\epsilon) \right), \tag{3.19}$$

where σ_0 is the lowest order cross section and $H(\epsilon)$ is a smooth function, with $H(0) = 1$, that we will not ultimately need to know. Note that, as we anticipated from Eqs. (3.17) and (3.18), this result is positive, and divergent like $1/\epsilon^2$ as $\epsilon \to 0$.

So far, the regularization scheme has succeeded in quantifying the degree of divergence of the total three-parton cross section, but it has not helped us solve the problem of the divergence, by recovering a finite result for a physical cross section. As we already anticipated above, this will come by calculating the loop correction to $e^+e^- \to q\bar{q}$.

3.1.2.4 $\sigma(e^+e^- \to q\bar{q})$ at one loop

We already made the point that to calculate the total cross section for $e^+e^- \to$ hadrons, we must sum over all $e^+e^- \to$ partons processes. At this order of perturbation theory $q\bar{q}$ is the only other process that contributes. There are three diagrams, shown in Fig. 3.1. They are down by one power of $\alpha_{\rm S}$ relative to the tree-level diagram,

$$\mathcal{M}_1 \propto \alpha_{\rm S} \mathcal{M}_0. \tag{3.20}$$

Therefore $|\mathcal{M}_1|^2$ is two powers down and hence negligible at the order to which we are working. However, since the final state is the same as that of the tree-level diagram, the two interfere, and their interference, $\mathbb{R}\mathrm{e}\{\mathcal{M}_0^* \mathcal{M}_1\}$ does contribute at order $\alpha_{\rm S}$.

In quantum mechanics, you know that we must sum over all unobserved quantum numbers at the amplitude level. Since the gluon momentum is unconstrained by the outgoing quark momenta, we must sum over *all* gluon momenta,

$$\int d^d k. \tag{3.21}$$

Note that there is no mass-shell-constraining delta-function: the virtual integral is over all arbitrary on- and off-shell momenta.

We begin with the first two diagrams, which are proportional to the self-energy of a massless quark. It is actually easy to see that these have to be zero in dimensional regularization: the value of the integral has dimensions E^{d-4}, but by Lorentz invariance the result of the integral can only be a function of the square of the quark's momentum, $p^2 = 0$, so there is nothing that can provide this dimensionality[3.1]. The only way these two facts can be reconciled is if the integral is zero. However, if we examine the integrand somewhat closer, this is very surprising, because it is positive definite. How can a positive definite quantity integrate to zero?

The answer to this question comes from a subtle use of dimensional regularization. In fact this integral is divergent in both the infrared and ultraviolet. If we split the integral into two parts by introducing an arbitrary separation scale Λ, then we obtain an ultraviolet contribution $\sim \Lambda^{-2\epsilon}/\epsilon$ and an infrared contribution $\sim -\Lambda^{-2\epsilon}/\epsilon$. Each is positive in its domain of applicability ($\epsilon > 0$ and $\epsilon < 0$ respectively), but after analytically continuing each to arbitrary ϵ, they are exactly equal and opposite, giving a zero result for these diagrams.

Turning to the third diagram, the vertex correction, we find that it is also divergent in the infrared and ultraviolet regions. However, its ultraviolet divergence is exactly equal and opposite to the one from the sum of the two self-energy diagrams. Therefore the sum of the three diagrams is ultraviolet finite and no renormalization is needed at this order. This actually follows directly from the Ward identity of QED. Thus, one simply has to evaluate the vertex correction diagram in dimensional regularization, to obtain the complete order $\alpha_{\rm S}$ contribution to $e^+e^- \to q\bar{q}$. We find that the infrared divergences do not cancel, and we obtain

$$\sigma_{q\bar{q}} = \sigma_0 \, C_F \frac{\alpha_s}{2\pi} H(\epsilon) \left(-\frac{2}{\epsilon^2} - \frac{3}{\epsilon} - 8 + \pi^2 + \mathcal{O}(\epsilon) \right). \tag{3.22}$$

Dimensional regularization has succeeded in regularizing the divergence of this contribution as well. This time, however, the result is negative and divergent as $\epsilon \to 0$. This should not surprise us, as we already noted that this is an interference term, so there is no requirement that it be positive, as there was for $\sigma_{q\bar{q}g}$. In fact a quick glance at Eqs. (3.19) and 3.22) shows us that the divergences are going to cancel between them.

3.1.3 The total cross section

In the previous section we discussed how dimensional regularization provides finite results for the total cross sections for the $e^+e^- \to q\bar{q}$ and $e^+e^- \to q\bar{q}g$ processes, which each diverge as $\epsilon \to 0$. For the benefit of those who slept through it, I restate them here:

$$\sigma_{q\bar{q}} = \sigma_0 \, C_F \frac{\alpha_s}{2\pi} H(\epsilon) \left(-\frac{2}{\epsilon^2} - \frac{3}{\epsilon} - 8 + \pi^2 + \mathcal{O}(\epsilon) \right), \tag{3.23}$$

$$\sigma_{q\bar{q}g} = \sigma_0 \, C_F \frac{\alpha_s}{2\pi} H(\epsilon) \left(\frac{2}{\epsilon^2} + \frac{3}{\epsilon} + \frac{19}{2} - \pi^2 + \mathcal{O}(\epsilon) \right). \tag{3.24}$$

According to our earlier discussion, the total cross section for $e^+e^- \to$ hadrons is given by the sum of the two. It is finite, so the limit $\epsilon \to 0$ can be taken,

[3.1] In fact this statement relies on working in a covariant gauge. In a lightcone gauge for example, the self-energy can depend on $n \cdot p$. This diagram is not then zero, but of course the final answer for the sum of the three diagrams is gauge invariant.

$$\sigma_{e^+e^-\to\text{hadrons}} = \sigma_0\left(1+C_F\frac{\alpha_\mathrm{S}}{2\pi}\,\frac{3}{2}\right) \tag{3.25}$$

$$= \sigma_0\left(1+\frac{\alpha_\mathrm{S}}{\pi}\right). \tag{3.26}$$

Of course, this would be useless if it depended on the regularization procedure. The proof of its independence is beyond us here, but it is worth demonstrating it, by comparison with another scheme, the gluon mass regularization, in which we have

$$\sigma_{q\bar{q}} = \sigma_0\,C_F\frac{\alpha_\mathrm{S}}{2\pi}\left[-\log^2\frac{1}{\epsilon}+3\log\frac{1}{\epsilon}-\frac{11}{2}+\frac{\pi^2}{3}+\mathcal{O}\left(\epsilon\right)\right], \tag{3.27}$$

$$\sigma_{q\bar{q}g} = \sigma_0\,C_F\frac{\alpha_\mathrm{S}}{2\pi}\left[\ \log^2\frac{1}{\epsilon}-3\log\frac{1}{\epsilon}+7-\frac{\pi^2}{3}+\mathcal{O}\left(\epsilon\right)\right], \tag{3.28}$$

$$\sigma_\mathrm{had} = \sigma_0\left[1+\frac{\alpha_\mathrm{S}}{\pi}\right]. \tag{3.29}$$

Note that the individual cross sections have completely different forms in the different schemes, but that the sum of the two is scheme independent.

Equation (3.26) is one of the most fundamental quantities in QCD and is certainly one of the most well-calculated and measured. Despite the fact that it is a relative small correction to the total rate, experimental and theoretical systematic errors are so small that they can almost be neglected — even with the large statistics of τ decays and Z decays at LEP, the statistical errors dominate. This means that not only does it provide one of the most accurate measurements, but its quoted accuracy is rather easy to interpret and implement in global analyses for example, unlike measurements that are dominated by systematics.

Equation (3.26) is now known up to order α_S^3. As discussed in Section 1.8, renormalization introduces a renormalization scale dependence into α_s and the coefficient functions beyond the first one,

$$\sigma_{e^+e^-\to\text{hadrons}} = \sigma_0\left(1+\frac{\alpha_\mathrm{S}(\mu)}{\pi}+C_2\left(\frac{\mu^2}{s}\right)\left(\frac{\alpha_\mathrm{S}(\mu)}{\pi}\right)^2+C_3\left(\frac{\mu^2}{s}\right)\left(\frac{\alpha_\mathrm{S}(\mu)}{\pi}\right)^3\right). \tag{3.30}$$

Reducing this renormalization-scale dependence is one of the biggest reasons for going to higher orders. As can be seen in Fig. 3.2, the scale-dependence is indeed significantly smaller at each order, giving stability over a wider range of μ. It can also be seen that provided μ is of order Q, the higher order corrections are relatively small. We will see shortly that simply taking the leading order result with $\mu=\sqrt{s}$ does surprisingly well and is certainly sufficient to understand the phenomenology.

3.1.4 α_S measurements

As I mentioned above, the experimental measurement of $R_{e^+e^-}$ gives one of the best measurements of α_s. In fact the LEP combined value of R_{had} is

$$R(LEP) = 20.767 \pm 0.025, \tag{3.31}$$

while the tree-level prediction is

$$R_0(M_z) = 19.984. \tag{3.32}$$

Combining the two, and simply using the leading order result with $\mu = M_z$, we obtain our first measurement of α_S,

$$\alpha_\mathrm{S}(M_z) = 0.124 \pm 0.004, \tag{3.33}$$

surprisingly close to the value using the four-loop result [5], 0.119 ± 0.003.

As we discussed in Section 1.8.3, since QCD predicts the scale dependence of α_S, one measurement at any scale is sufficient to give a prediction for all scales. We can therefore phrase measurements

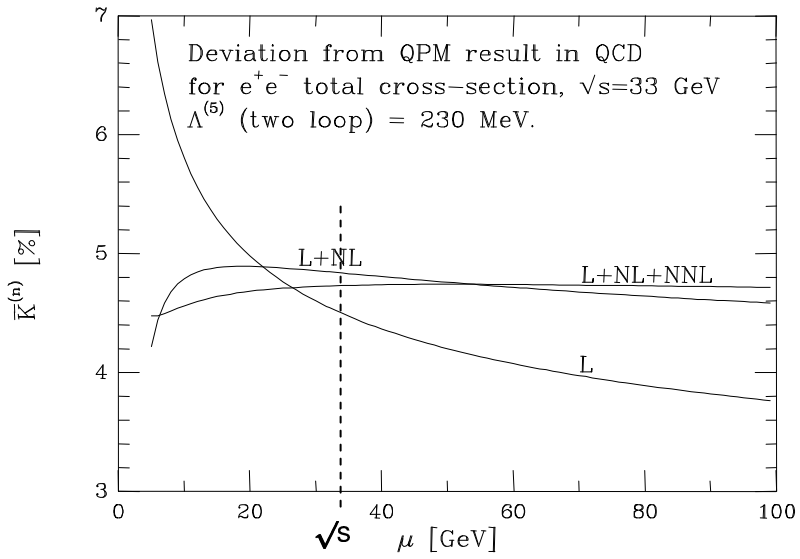


Fig. 3.2: The QCD prediction for the corrections to $R_{e^+e^-}$ at $\sqrt{s} = 33$ GeV as a function of renormalization scale at leading, next-to-leading, and next-to-next-to-leading order, taken from ESW [1]

at other scales either as tests of QCD throughout the intervening energy range or, by translating them all into measurements at a single scale, as different measurements of the same quantity that can be combined to give a better overall measurement.

As an example, the average measurement of R over several energy points around 34 GeV is

$$R(PETRA) = 3.88 \pm 0.03, \tag{3.34}$$

while the tree-level prediction is

$$R_0(34\text{ GeV}) = 3.69. \tag{3.35}$$

Again using the leading order result, we obtain

$$\alpha_{\rm S}(34\text{ GeV}) = 0.162 \pm 0.026. \tag{3.36}$$

Finally, using the one-loop renormalization group equation, we can convert this into a measurement of $\alpha_{\rm S}(M_z)$,

$$\alpha_{\rm S}(M_z) = 0.134 \pm 0.018. \tag{3.37}$$

This is in good agreement with the value from LEP, although with much larger uncertainties, simply due to the fact that the statistics of the PETRA experiments were much lower.

As a final example, we consider τ decays. The QCD corrections to the hadronic decay rate actually have two effects: on the ratio of branching fractions, R_τ, as discussed earlier, and also directly on the total decay rate of the τ. These can form the basis for two analyses in which the experimental errors are largely independent. The combined result for the two is

$$\alpha_{\rm S}(M_\tau = 1.77\text{ GeV}) = 0.34 \pm 0.01. \tag{3.38}$$

This time, because we are translating over such a wide energy range the one-loop renormalization group equation does not do quite such a good job,

$$\alpha_{\rm S}^{\text{(one-loop)}}(M_z) = 0.1272 \pm 0.0014, \tag{3.39}$$

compared to the four-loop value [5]

$$\alpha_{\rm S}(M_z) = 0.1212 \pm 0.0011, \tag{3.40}$$

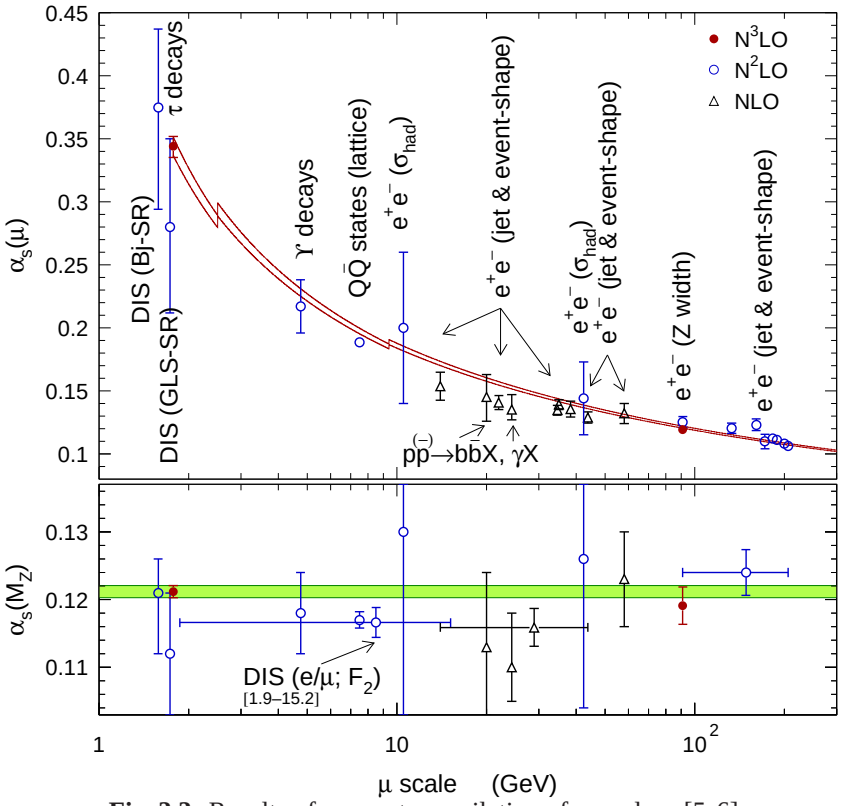


Fig. 3.3: Results of a recent compilation of $\alpha_{\rm S}$ values [5, 6]

but it is not so far out. Note in this case the phenomenon of the 'incredible shrinking error'. Although the measurement at the τ mass scale has a precision of about 3%, after evolving it to M_z the relative uncertainty gets scaled down by the ratio of the two α_s values, and τ decays give the best single measurement of $\alpha_s(M_z)$.

The results of a recent compilation [5, 6] are shown in Fig. 3.3. The scale dependence shows excellent agreement with the predictions of perturbative QCD over a wide energy range. When translated into measurements of $\alpha_{\rm S}(M_z)$, the separate measurements cluster strongly around the average value,

$$\alpha_{\rm S}^{\rm (average)}(M_z) = 0.1204 \pm 0.0009. \tag{3.41}$$

3.2 Deep inelastic scattering revisited

The parton model I described in the last lecture assumed that the partons are non-interacting. But we know that they do interact via QCD, so what will happen when we consider these interactions? We will discover that the structure functions do become slowly (logarithmically) varying with Q^2. We start by considering the next-to-leading order QCD corrections to quark scattering. We will find that these, if calculated naively, would be divergent, but that these divergences can be absorbed into the parton distribution functions. These will then become scale-dependent, giving rise to the Q^2-dependence of the structure functions.

3.2.1 NLO corrections to DIS

The next-to-leading order corrections come from three sources (recalling that we sum and integrate over all final states X, so we must sum over all contributions in which any kind of parton is scattered):

1. One-loop corrections to $eq \to eq$,

2. $eq \to eqg$,
3. $eg \to eq\bar{q}$.

The third contribution is completely new in QCD and is not present in the parton model. We come back to it in a later section. The other two can more genuinely be thought of as higher-order corrections to the parton model process. We start with the second.

There are two contributing diagrams. The matrix element squared can be obtained by crossing from $e^+e^- \to q\bar{q}g$ (1.66). Labeling the momenta as

$$e(k) + q(\eta p) \to e(k') + q(p_1) + g(p_2), \tag{3.42}$$

we obtain

$$\sum |\mathcal{M}|^2 = \frac{8C_F N_c e^4 e_q^2 g_s^2}{k\cdot k'\, p_1\cdot p_2\, \eta p\cdot p_2} \left((p_1\cdot k)^2 + (\eta p\cdot k)^2 + (p_1\cdot k')^2 + (\eta p\cdot k')^2\right). \tag{3.43}$$

As usual the phase space is given by (2.24),

$$dPS = \frac{Q^2}{16\pi^2 s x^2} dQ^2\, dx\, dPS_X. \tag{3.44}$$

This time X consists of two partons so is non-trivial,

$$dPS_X = \frac{d\cos\theta\, d\phi}{32\pi^2}, \tag{3.45}$$

where θ and ϕ refer to the direction of p_1 in the centre-of-mass system of $\eta p + q$. It is conventional to replace $\cos\theta$ by the manifestly Lorentz-invariant variable z,

$$z \equiv \frac{p_1\cdot p}{q\cdot p} = \tfrac{1}{2}(1 - \cos\theta), \tag{3.46}$$

with range $0 < z < 1$, giving

$$dPS_X = \frac{dz\, d\phi}{16\pi^2}. \tag{3.47}$$

It will later be instructive to know the transverse momentum of p_1 in this frame,

$$k_\perp^2 = Q^2\left(\frac{\eta}{x} - 1\right) z(1-z). \tag{3.48}$$

Note also that the case $\eta = x$ corresponds to a massless final state. Kinematically this can only happen if either p_1 or p_2 are infinitely soft (i.e., have zero energy), or if they are exactly collinear.

We therefore have

$$\frac{d\sigma^2(e+q)}{dx\, dQ^2} = \frac{1}{4N_c}\frac{1}{2\hat{s}}\frac{Q^2}{16\pi^2 s x^2}\int \frac{dz\, d\phi}{16\pi^2}\frac{8C_F N_c e^4 e_q^2 g_s^2}{k\cdot k'\, p_1\cdot p_2\, \eta p\cdot p_2}\left((p_1\cdot k)^2 + (\eta p\cdot k)^2 + (p_1\cdot k')^2 + (\eta p\cdot k')^2\right). \tag{3.49}$$

Rewriting in terms of our kinematic variables and averaging over ϕ, we have

$$\begin{aligned}\left\langle \frac{(p_1\cdot k)^2 + (\eta p\cdot k)^2 + (p_1\cdot k')^2 + (\eta p\cdot k')^2}{k\cdot k'\, p_1\cdot p_2\, \eta p\cdot p_2}\right\rangle_\phi \\ = \frac{1}{y^2Q^2}\left((1+(1-y)^2)\left[\frac{1+x_p^2}{1-x_p}\frac{1+z^2}{1-z} + 3 - z - x_p + 11x_p z\right] - y^2\left[8zx_p\right]\right),\end{aligned} \tag{3.50}$$

where $x_p = x/\eta$. Two things are already clear: at this order we will have a non-zero longitudinal structure function, $F_L(x, Q^2)$; and the z integration, which runs from 0 to 1, will give a divergent contribution

to F_2. This should worry us, since we are calculating a physical cross section, but let us continue for a while and see what happens.

Putting everything together we have

$$\frac{d\sigma^2(e+q)}{dx\,dQ^2} = \frac{C_F\alpha^2 e_q^2\alpha_{\rm S}}{2\eta x^2y^2s^2}\int_0^1 dz\left((1+(1-y)^2)\left[\frac{1+x_p^2}{1-x_p}\frac{1+z^2}{1-z}+3-z-x_p+11x_pz\right]-y^2\left[8zx_p\right]\right), \tag{3.51}$$

and hence

$$F_2(x,Q^2) = \sum_q\int_x^1 dx_p\,e_q^2\frac{x}{x_p}f_q\left(\frac{x}{x_p}\right)\frac{C_F\alpha_{\rm S}}{2\pi}\int_0^1 dz\left(\frac{1+x_p^2}{1-x_p}\frac{1+z^2}{1-z}+3-z-x_p+11x_pz\right). \tag{3.52}$$

The divergence at $z\to 1$ corresponds to kinematic configurations in which the outgoing gluon becomes exactly collinear with the incoming quark. Therefore in the Feynman diagram in which the gluon is attached to the incoming quark, the internal quark line becomes on-shell, causing the divergence. Note also that the coefficient of the divergence itself diverges at the point $x_p=1$, at which the gluon is infinitely soft.

In order to study the divergence, let us first regulate it by calculating the contribution from emission with $k_\perp^2>\mu^2$ (and assume $\mu^2\ll Q^2$ for simplicity). Since $k_\perp^2$ is proportional to $(1-z)$ this will give us finite integrals. At any time, the full result can be obtained by setting $\mu\to 0$. We therefore obtain

$$F_2(x,Q^2) = \sum_q\int_x^1 dx_p\,e_q^2\frac{x}{x_p}f_q\left(\frac{x}{x_p}\right)\frac{\alpha_{\rm S}}{2\pi}\left(\hat{P}(x_p)\log\frac{Q^2}{\mu^2}+R(x_p)\right), \tag{3.53}$$

where the function $R(x_p)$ is finite. In the following we will not keep track of this function, although it would be essential for quantitative analysis. The function $P(x_p)$ we introduced in (3.53) is called the splitting function (or more strictly speaking the unregularized splitting function),

$$\hat{P}(x) = C_F\frac{1+x^2}{1-x}. \tag{3.54}$$

It actually describes the probability distribution of quarks produced in a splitting process, $q\to qg$ in which the produced quark has a fraction x of the original quark's momentum. (We will quantify this statement slightly more shortly.)

Obviously by regulating the divergence we have not removed it: physical cross sections are still supposed to be obtained by setting $\mu\to 0$, in which case F_2 is logarithmically divergent. However, before discussing what happens to this divergence, let us consider the virtual one-loop correction to $eq\to eq$. Since this diagram contains two quark-gluon couplings, when squared it would give an $\mathcal{O}(\alpha_{\rm S}^2)$ correction. However, since it has the same final state as the lowest order diagram, we must consider the interference between the two, and this interference is $\mathcal{O}(\alpha_{\rm S})$, so we must include it.

We could obtain the result for the one-loop diagram by crossing from $e^+e^-\to q\bar{q}$. However, to illustrate the physics, it is sufficient to recall a few of its features. Firstly, since the external particles are the same as in the lowest-order process, the kinematics must be the same. In particular, it can only contribute at the point $\eta=x$. Secondly, as in the e^+e^- case, the interference of the one-loop and tree-level diagrams is divergent and negative. In fact the kinematic regions in which the one-loop integrand diverges are exactly the same as those of the $eq\to eqg$ contribution we have just considered: when the gluon is soft, or is collinear with either of the quarks.

It turns out that the divergence is exactly right to cancel the one we obtained above at $x_p\to 1$. In fact one finds that after including the one-loop contribution, one gets exactly the same formula as (3.53) except that the unregularized splitting function $\hat{P}(x_p)$ is replaced by the regularized one, $P(x_p)$,

$$P(x) = \hat{P}(x)+P_{\rm virtual}(x). \tag{3.55}$$

Since the one-loop contribution has the same kinematics as the lowest-order process, $P_{\rm virtual}(x)$ must be proportional to $\delta(1-x)$. $P(x)$ is therefore a distribution.

To define it, we will need to use a mathematical trick called the plus-distribution. Given some function $f(x)$, which is well-defined for all $0 \leq x < 1$, we define a distribution $f(x)_+$ on the region $0 \leq x \leq 1$, as

$$f(x)_+ = f(x) - \delta(1-x)\int_0^1 dx' f(x'). \tag{3.56}$$

The plus-distribution is most useful when the function $f(x)$ is divergent at $x \to 1$. This means that for any other function $g(x)$, which is smooth at $x = 1$, we have the property

$$\int_0^1 dx\, f(x)_+\, g(x) = \int_0^1 dx\, f(x)\left(g(x) - g(1)\right). \tag{3.57}$$

Provided that $g(x)$ goes to $g(1)$ sufficiently quickly, this integral is finite.

After including the virtual contribution, the splitting function is given by

$$P(x) = C_F\left[\frac{1+x^2}{(1-x)_+} + \frac{3}{2}\delta(1-x)\right]. \tag{3.58}$$

This is actually the first correction to a function that describes the momentum distribution of quarks within quarks,

$$\mathcal{P}(x) \equiv \delta(1-x) + \frac{\alpha_{\rm S}}{2\pi}\log\frac{Q^2}{Q_0^2}P(x) + \mathcal{O}(\alpha_{\rm S}^2\log^2), \tag{3.59}$$

where the distribution is defined to be a pure quark at scale Q_0 and probed at scale Q.

Inserting the full splitting function into (3.53), we find that the divergence at $x_p \to 1$ cancels between the real and virtual terms, but the divergence for $x_p < 1$ due to the region $z \to 1$ still remains.

3.2.2 Factorization of divergences

To understand why the results are still divergent even after including the virtual terms, and what ultimately happens to the divergences, we consider their physical origin. Like the e^+e^- annihilation case, we have singularities from regions in which the real gluon is collinear with either the incoming or outgoing quark, or is soft, and also from the virtual graph, as illustrated in Fig. 3.4. All these contributions were present in e^+e^- annihilation, but there we found that the divergences all cancelled to give a finite contribution. Why is the present situation different? In fact we find that here the magnitudes of the divergences are such as to cancel, but that the divergences arise in different regions of the x_p integral, so are prevented from cancelling.

In the e^+e^- case, we argued that the singular regions of real emission were indistinguishable from the lowest order process, since an infinitely soft gluon could not produce any hadrons and the jets produced by two collinear partons were indistinguishable from a single jet with their combined momentum. This statement is true here for the soft and final-state collinear contributions, but not the initial-state contribution. The final state of this contribution is indeed indistinguishable from the lowest order process (it has an additional jet collinear with the outgoing proton remnant, but this too gives a jet and the superposition of the two is indistinguishable from the proton remnant in the lowest order process). However, because we have used the parton model, we must convolute the partonic cross sections over an arbitrary (measured from experiment) probability distribution function, processes with different incoming momenta are effectively distinguishable. In all the singular regions, the final state of the process is massless, and this fact fixes the incoming momentum (to the value $Q/2x$ in the Breit frame), but in the initial-state singular process it is the internal line whose momentum gets fixed, as indicated in Fig. 3.4. Thus the incoming momentum in (c) is larger than in the other cases and its divergence, at $\eta > x$, cannot cancel the others, at $\eta = x$.

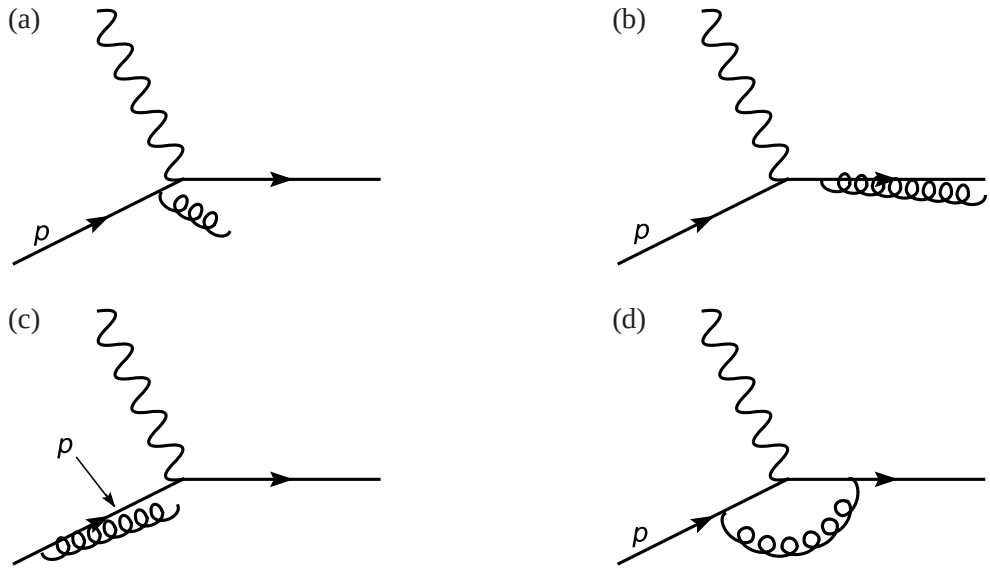


Fig. 3.4: Divergent contributions to DIS: (a) soft, (b) final-state collinear, (c) initial-state collinear, and (d) virtual. The label p shows which momentum in each contribution is fixed by the massless final-state condition.

As I mentioned earlier, these divergences come from the virtuality of the internal particle vanishing and hence the propagator diverging. Using the uncertainty principle, vanishingly small virtuality corresponds to arbitrarily long time-scales. This seems to be in direct contradiction with the assumption underlying the parton model, that the virtual photon takes an extremely rapid snapshot of the proton.

The problem is actually one of overcounting. We first introduced the pdfs, which are supposed to contain all information about the internal structure of the proton. Presumably this internal structure is the result of QCD interactions. We then tried to calculate the QCD corrections to the quark scattering cross sections, integrating over all final states, so all energy-scales. But these QCD corrections should somehow already be included in the internal dynamics of the proton.

To resolve this overcounting, we have to separate (or 'factorize') the different types of physics at different energy scales. Like in our discussion of renormalization, I will first try to give the physical picture in terms of a cutoff, before returning later to describe how factorization works in practice in dimensional regularization. We introduce the factorization scale μ, and call all physics at scales below μ part of the hadron wave function, and lump it into the parton distribution functions, and call all physics at scales above μ part of the partonic scattering cross section (or 'coefficient function').

Therefore we do in fact have a transverse momentum cutoff in the $eq \to eqg$ process and the form of (3.53) is correct.

Since physics at scales below μ is included in the pdfs and physics above is not, the pdfs themselves must become μ-dependent. We therefore have

$$F_2(x,Q^2) = \sum_q e_q^2 \int_x^1 dx_p \frac{x}{x_p} f_q\left(\frac{x}{x_p},\mu^2\right)\left\{\delta(1-x_p) + \frac{\alpha_{\rm S}}{2\pi}\left(P(x_p)\log\frac{Q^2}{\mu^2} + R(x_p)\right) + \mathcal{O}(\alpha_{\rm S}^2)\right\}, \tag{3.60}$$

where the function $R(x_p)$ is not necessarily the same one as earlier, as the virtual contributions could have introduced some additional finite terms.

Note that the structure functions are now Q^2-dependent, violating Bjorken scaling. However, they also appear to be μ^2 dependent, which should worry us: μ was introduced in a completely ad hoc theoretical way: it simply separates physical processes into two parts that are dealt with in different ways, and the final result, which is the sum of the two parts, should not depend on where the separation was

made. We return to discuss this point in more detail after calculating the μ^2-dependence of the pdfs.

It is important to emphasize that, although we have derived these formulae for the higher order corrections to DIS, the leading logarithmic behaviour is universal. In particular, for any quark-induced process with a hard scale Q, we expect a hadronic cross section of the form

$$\sigma_h(p_h) = \sum_q \int d\eta\, f_q\left(\eta, \mu^2\right) \left\{ \sigma_q(\eta p_h) + \frac{\alpha_{\rm S}}{2\pi} \log\frac{Q^2}{\mu^2} \int dz\, P(z)\, \sigma_q(z\eta p_h) \right\}, \tag{3.61}$$

where $\sigma_q(p)$ is the partonic cross section for a quark of flavour q and momentum p.

3.2.3 DGLAP evolution equation

Although the pdfs are fundamentally non-perturbative and cannot be predicted from first principles at present, physics at scales close to μ^2 can be described perturbatively. We can therefore calculate the μ^2-dependence of the pdfs so that, given their value at some starting scale μ_0, for example from experimental measurements, we can calculate their values at all higher scales μ.

To do this, we use the fact just noted, that physical cross sections should not depend on μ^2. Therefore we should have

$$\mu^2 \frac{dF_2(x,Q^2)}{d\mu^2} = 0, \tag{3.62}$$

or at least, since we are working at $\mathcal{O}(\alpha_{\rm S})$,

$$\mu^2 \frac{dF_2(x,Q^2)}{d\mu^2} = \mathcal{O}(\alpha_{\rm S}^2). \tag{3.63}$$

Applying this to (3.60), we obtain

$$\mu^2 \frac{d}{d\mu^2} f_q\left(x,\mu^2\right) = \frac{\alpha_{\rm S}}{2\pi} \int_x^1 \frac{dx_p}{x_p} f_q\left(\frac{x}{x_p}, \mu^2\right) P(x_p) + \mathcal{O}(\alpha_{\rm S}^2). \tag{3.64}$$

Equation (3.64) is called the Dokshitzer–Gribov–Lipatov–Altarelli–Parisi (or DGLAP, or GLAP, or Altarelli–Parisi for short) evolution equation. Note that the rate of change of the pdf at some x value depends on its value at all higher xs.

To understand its physical content, it is useful to rewrite the splitting function,

$$P(x) = C_F \left[\frac{1+x^2}{(1-x)_+} + \frac{3}{2}\delta(1-x) \right] = C_F \left(\frac{1+x^2}{1-x} \right)_+, \tag{3.65}$$

to give

$$\mu^2 \frac{d}{d\mu^2} f_q\left(x,\mu^2\right) = C_F \frac{\alpha_{\rm S}}{2\pi} \int_x^1 \frac{dx_p}{x_p} f_q\left(\frac{x}{x_p},\mu^2\right) \frac{1+x_p^2}{1-x_p} - C_F \frac{\alpha_{\rm S}}{2\pi} f_q\left(x,\mu^2\right) \int_0^1 dx_p \frac{1+x_p^2}{1-x_p}. \tag{3.66}$$

The first term represents the fact that the pdf at a given x value is increased by quarks with higher x's reducing their momentum fractions by radiating gluons. The second term represents the fact that it is decreased by the quarks at that x reducing their momentum fractions by radiating gluons. Each contribution is divergent due to emission with $x_p \to 1$, i.e., infinitely soft gluon emission, involving an infinitely small change in x. However the two divergences exactly cancel because the number of quarks being lost to this x value by infinitely soft gluon emission is equal to the number being gained.

The DGLAP equation is most easily solved in moment space. For any function $f(x)$, we define

$$f_N = \int_0^1 dx\, x^{N-1}\, f(x), \tag{3.67}$$

the Mellin transform. Taking moments of both sides of (3.64), we obtain

$$\mu^2\frac{d}{d\mu^2}f_{qN}\left(\mu^2\right) = \frac{\alpha_{\rm S}}{2\pi}\int_0^1 dx\,x^{N-1}\int_x^1\frac{dx_p}{x_p}f_q\left(\frac{x}{x_p},\mu^2\right)P(x_p)+\mathcal{O}(\alpha_{\rm S}^2) \tag{3.68}$$

$$= \frac{\alpha_{\rm S}}{2\pi}P_N f_{qN}(\mu^2). \tag{3.69}$$

It is common to introduce the notation

$$\gamma_N(\alpha_{\rm S}) = \frac{\alpha_{\rm S}}{2\pi}P_N+\mathcal{O}(\alpha_{\rm S}^2), \tag{3.70}$$

where γ_N is known as the anomalous dimension. If we assume that the coupling $\alpha_{\rm S}$ is fixed, we can easily solve (3.69) with the boundary condition of given values for f_{qN} at some fixed scale μ_0,

$$f_{qN}(\mu^2) = f_{qN}(\mu_0^2)\left(\frac{\mu^2}{\mu_0^2}\right)^{\gamma_N(\alpha_{\rm S})}. \tag{3.71}$$

However, as we have seen, the renormalization of QCD means that the coupling constant becomes scale dependent, $\alpha_{\rm S}(\mu^2)$, according to renormalization group equation

$$\mu^2\frac{d}{d\mu^2}\alpha_{\rm S}(\mu^2) = \beta(\alpha_{\rm S}(\mu^2)) = -\frac{\beta_0}{2\pi}\alpha_{\rm S}^2(\mu^2)+\mathcal{O}(\alpha_{\rm S}^3). \tag{3.72}$$

Inserting the solution of the running coupling, Eq. (1.73), into (3.69), we obtain

$$f_{qN}(\mu^2) = f_{qN}(\mu_0^2)\left(\frac{\alpha_{\rm S}(\mu_0)}{\alpha_{\rm S}(\mu)}\right)^{\frac{P_N}{\beta_0}}. \tag{3.73}$$

Having the solution for f_q in moment N-space, we have to convert it back to x-space. This is done by the Inverse Mellin Transform, where f_{qN} is continued to the complex plane,

$$f_q(x) = \frac{1}{2\pi i}\int_C dN\,f_{qN}\,x^{-N}, \tag{3.74}$$

where the contour C runs parallel to the imaginary axis to the right of all poles. Because of the complexity of this process, the DGLAP equation is often solved simply by 'brute force' numerical solution of (3.64).

Beyond $\mathcal{O}(\alpha_{\rm S})$ the general structure of (3.69) and (3.72) remains unchanged: the anomalous dimension and β function simply become power series in $\alpha_{\rm S}$.

3.2.4 Scheme/scale dependence

Factorization, as introduced above, may seem pretty arbitrary. However it can be proved to all orders in perturbation theory. The most convenient way to do this is to use, instead of the transverse momentum cutoff we used above, dimensional regularization. When we calculate the NLO cross section in d dimensions, the divergence shows up as a pole, $1/\epsilon$. The coefficient multiplying this pole turns out to be the same splitting function we encountered earlier.

In d dimensions, we obtain for the structure function up to $\mathcal{O}(\alpha_{\rm S})$,

$$F_2(x,Q^2) = \sum_q e_q^2\int_x^1 dx_p\frac{x}{x_p}\bar{f}_q\left(\frac{x}{x_p}\right)\left\{\delta(1-x_p)+\frac{\alpha_{\rm S}}{2\pi}\left(\left(\frac{4\pi\mu^2}{Q^2}\right)^{\epsilon}\frac{-1}{\epsilon}P(x_p)+R(x_p)\right)+\mathcal{O}(\epsilon)\right\}, \tag{3.75}$$

where μ is the scale introduced to make the coupling constant dimensionless. Note that I have sneakily added a bar to f_q and that it is scale independent. $\bar{f}_q$ is known as the bare pdf. We now note that the

distribution functions themselves are not physical observables, only their convolution with coefficient functions is. I can therefore define a modified set of distribution functions as follows:

$$x\,\bar{f}_q(x) \equiv \int_x^1 dx_p \frac{x}{x_p} f_q\left(\frac{x}{x_p},\mu_F^2\right)\left\{\delta(1-x_p) - \frac{\alpha_{\rm S}}{2\pi}\left(\left(\frac{4\pi\mu^2}{\mu_F^2}\right)^\epsilon \frac{-1}{\epsilon}P(x_p) + K(x_p)\right)\right\}, \quad (3.76)$$

where μ_F is the (completely arbitrary again) factorization scale, and $K(x_p)$ is a completely arbitrary finite function to be discussed shortly. (To fit in with the standard notation, I should really multiply all occurrences of $\alpha_{\rm S}$ by $1/\Gamma(1-\epsilon) = 1 - \gamma_E\epsilon + \mathcal{O}(\epsilon^2)$, but this will merely change the values of $R(x_p)$ and $K(x_p)$ which I do not specify anyway.) Combining (3.75) and (3.76), we end up with

$$F_2(x,Q^2) = \sum_q e_q^2 \int_x^1 dx_p \frac{x}{x_p} f_q\left(\frac{x}{x_p},\mu_F^2\right)\left\{\delta(1-x_p) + \frac{\alpha_{\rm S}}{2\pi}\left(P(x_p)\log\frac{Q^2}{\mu_F^2} + R(x_p) - K(x_p)\right) + \mathcal{O}(\alpha_{\rm S}^2)\right\}. \quad (3.77)$$

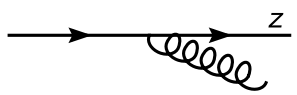
Note that this has the identical form to (3.60), except for the finite function. It is clear from (3.76) that $f_q(x,\mu_F^2)$ depends on the function $K(x_p)$. It therefore seems like we have no predictive power: the pdf and coefficient function each depend on the completely arbitrary function $K(x_p)$ and the completely arbitrary scale μ_F (note that all dependence on μ has again completely cancelled. As I said in the context of renormalization, many textbooks simply set it equal μ right from the start, but I consider this slightly confusing as they have quite different physical meaning. Having performed this manoeuvre, I henceforth drop the subscript $_F$). However, the factorization theorem proves, firstly that for any physical quantity, all dependence on $K(x_p)$ and μ will cancel and secondly that the scheme- and scale-dependent pdfs, $f_q(x,\mu^2)$ are universal (i.e., process-independent).

Two schemes are in common use, the $\overline{\rm MS}$ scheme in which $K(x_p)$ is zero, and the DIS scheme in which $K(x_p) = R(x_p)$, i.e. in which for $\mu = Q$ the parton model result is exact.

To understand the physical content of the scheme-dependence, it is worth while going back to the case with a cutoff. If, instead of a cut on transverse momentum we had used a cut on the virtuality of the internal quark line to separate the pdf from the coefficient function, we would have got exactly the same form as (3.60) except that $R(x_p)$ would have been a different function. In particular, it would differ by a $\log[(1-x_p)/x_p]$ term, together with some non-logarithmic terms. In fact, all logarithmic terms turn out to be the same with a p_t cutoff as in the $\overline{\rm MS}$ scheme, so for many purposes the two can be considered equivalent.

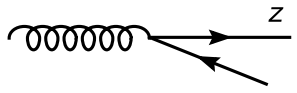
Although dependence on the scheme and scale must cancel in physical quantities, it is only guaranteed to do so after calculating to infinite orders of perturbation theory. At any finite order there can be some residual dependence. We must therefore have a procedure for choosing a value of μ. Essentially the identical discussion we had for the renormalization scale choice applies here. One can show that a structure like (3.60) continues to all orders of perturbation theory and that for every power of $\alpha_{\rm S}$, one gets a power of $\log Q^2/\mu^2$. Thus every order of perturbation theory contains terms like $\alpha_{\rm S}^n \log^m Q^2/\mu^2$, $m \le n$. It is clear that if μ is a long way from Q, the log can be large enough to compensate the smallness of $\alpha_{\rm S}$ and the perturbative series will not converge quickly. One should therefore choose μ 'not too far' from Q.

It is worth mentioning that one can set up DGLAP evolution equations for the Q^2-dependence of the structure functions, F_2 and F_L, themselves. These are then automatically scheme- and scale-independent even at finite orders of perturbation theory. This is sometimes known as the scheme-independent scheme.

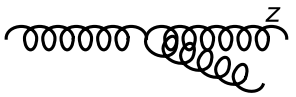


$$P_{qq}(x) = C_F\left[\frac{1+x^2}{(1-x)_+} + \frac{3}{2}\delta(1-x)\right]$$

$$P_{qg}(x) = T_R\left[x^2+(1-x)^2\right]$$



$$P_{gq}(x) = C_F\left[\frac{1+(1-x)^2}{x}\right]$$



$$P_{gg}(x) = C_A\left[\frac{2x}{(1-x)_+} + 2\frac{1-x}{x} + 2x(1-x)\right] + \beta_0\delta(1-x)$$

Fig. 3.5: The four DGLAP splitting functions of QCD

3.2.5 Initial-state gluons

As mentioned right at the start of this section, we also obtain $\mathcal{O}(\alpha_{\rm S})$ corrections from the process $eg \to eq\bar{q}$. Most of what we said above carries over in a straightforward way. Although there is no soft singularity or virtual term to cancel it, there is a collinear singularity. This corresponds to a two-step process in which a gluon splits to a q–$\bar{q}$ pair, one of which interacts with the photon. The singularity again corresponds to the virtuality of the internal quark line going to zero. This singularity can again be absorbed into a factorized universal pdf for the gluon. We end up with an additional contribution to the structure function of

$$F_2(x,Q^2) = \sum_q e_q^2 \int_x^1 dx_p \frac{x}{x_p} f_g\left(\frac{x}{x_p},\mu^2\right)\left\{\frac{\alpha_{\rm S}}{2\pi}\left(P_{qg}(x_p)\log\frac{Q^2}{\mu^2} + R_g(x_p) - K_{qg}(x_p)\right) + \mathcal{O}(\alpha_{\rm S}^2)\right\}, \tag{3.78}$$

where the sum over q is over all 'light' flavours. We now have four different types of splitting function, illustrated in Fig. 3.5. The DGLAP equation now becomes a set of coupled equations:

$$\mu^2\frac{d}{d\mu^2} f_a\left(x,\mu^2\right) = \sum_b \frac{\alpha_{\rm S}}{2\pi}\int_x^1 \frac{dx_p}{x_p} f_b\left(\frac{x}{x_p},\mu^2\right) P_{ab}(x_p) + \mathcal{O}(\alpha_{\rm S}^2). \tag{3.79}$$

In moment space, this can be conveniently written as a matrix equation (in general of $(2N_f{+}1)\times(2N_f{+}1)$ matrices, but for simplicity we show the case of only one flavour of quark):

$$\mu^2\frac{d}{d\mu^2}\begin{pmatrix} f_{qN} \\ f_{\bar{q}N} \\ f_{gN}\end{pmatrix} = \begin{pmatrix} \gamma_{qqN}(\alpha_{\rm S}(\mu)) & 0 & \gamma_{qgN}(\alpha_{\rm S}(\mu)) \\ 0 & \gamma_{qqN}(\alpha_{\rm S}(\mu)) & \gamma_{qgN}(\alpha_{\rm S}(\mu)) \\ \gamma_{gqN}(\alpha_{\rm S}(\mu)) & \gamma_{gqN}(\alpha_{\rm S}(\mu)) & \gamma_{ggN}(\alpha_{\rm S}(\mu))\end{pmatrix}\begin{pmatrix} f_{qN} \\ f_{\bar{q}N} \\ f_{gN}\end{pmatrix}. \tag{3.80}$$

Exactly the same solution is obtained, but in matrix notation,

$$\begin{pmatrix} f_{qN}(\mu^2) \\ f_{\bar{q}N}(\mu^2) \\ f_{gN}(\mu^2)\end{pmatrix} = \exp\int_{\mu_0^2}^{\mu^2}\frac{d\mu'^2}{\mu'^2}\begin{pmatrix} \gamma_{qqN}(\alpha_{\rm S}(\mu')) & 0 & \gamma_{qgN}(\alpha_{\rm S}(\mu')) \\ 0 & \gamma_{qqN}(\alpha_{\rm S}(\mu')) & \gamma_{qgN}(\alpha_{\rm S}(\mu')) \\ \gamma_{gqN}(\alpha_{\rm S}(\mu')) & \gamma_{gqN}(\alpha_{\rm S}(\mu')) & \gamma_{ggN}(\alpha_{\rm S}(\mu'))\end{pmatrix}\begin{pmatrix} f_{qN}(\mu_0^2) \\ f_{\bar{q}N}(\mu_0^2) \\ f_{gN}(\mu_0^2)\end{pmatrix}. \tag{3.81}$$

This is even more troublesome to do by the Inverse Mellin Transform, so the full set of DGLAP equations is almost always solved numerically.

Note that at higher orders of perturbation theory, even the zero entries in (3.80) become non-zero, as do contributions like $P_{qq'}(x)$.

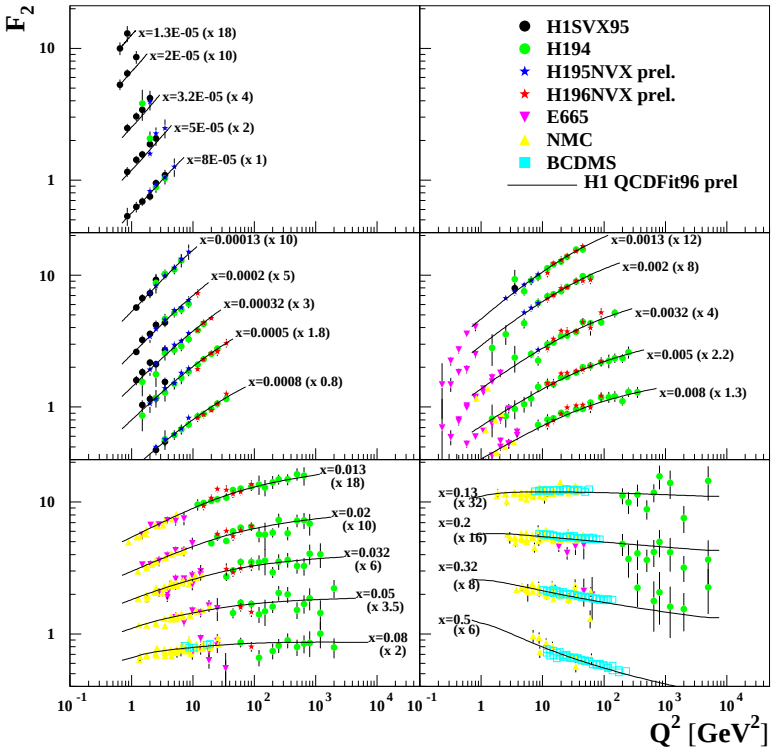


Fig. 3.6: Fit to the F_2 data over a wide range of Q^2 values, exhibiting violation of Bjorken scaling

3.2.6 Violation of Bjorken scaling

As we already noted, the factorization of initial-state singularities introduces a logarithmic Q^2 dependence into the structure functions and therefore a slow violation of Bjorken scaling. There is a close analogy with the renormalization of one-scale cross sections, where the energy-dependence was entirely due to the quantum corrections. Although the pdfs at some low scale are entirely non-perturbative and must be fit to data, the scale-dependence is entirely predicted by QCD and provides a stringent test over a wide range of energy scales. The result is impressive, see Fig. 3.6.

3.3 Summary

NLO calculations are hard! This is mainly because the real and virtual corrections are each divergent and must be regularized in some self-consistent way, for example with dimensional regularization. Unlike the ultraviolet divergences, which are isolated in well-localized pieces of the loop calculation and can effectively be removed by a redefinition of the Feynman rules, these divergences arise in different partonic contributions to physical observables. They must therefore be kept explicit until the very end of the calculation when all the partonic contributions are combined. Only then, provided our observable is infrared safe, will the real and virtual divergences cancel to yield a finite result.

Processes with incoming partons have extra divergences, arising from a miscancellation of the initial-state-collinear real and virtual contributions, which appear at different points in the integral over incoming momentum fraction. (It is worth mentioning that the same argument applies to the final-state distributions of identified hadrons, for example the momentum distribution of pions produced in e^+e^- annihilation.) These divergences have to be factorized into the non-perturbative, but universal, parton distribution functions at some factorization scale μ_F. This extra scale in the structure functions allows them to be Q^2-dependent. This Q^2-dependence is entirely driven by the μ_F^2-dependence of the parton distribution functions, which is predicted by the DGLAP evolution equations. Thus structure function data over a wide range of Q^2 provide a stringent test of perturbative QCD.

4 Summary

In this short course on QCD phenomenology, I have resisted the temptation to review the many important tests and studies of QCD that have been made over the years and have instead tried to concentrate on the key ideas that underpin them. These are:

- The gauge invariance of the theory, which allows us to write down the Lagrangian and which predicts some of the most important features of the theory: the universality of the coupling constant and the self-coupling of gluons, which ultimately leads to the negative β function and hence to asymptotic freedom at high energies and strong interactions at low energies.
- Renormalization and decoupling, which allow us to make predictive calculations at finite energy, without knowing the full structure of the theory to arbitrarily high energy and without the introduction of arbitrarily many input parameters. Renormalization is related to the quantum structure of the theory and introduces a dimensionful scale into even the scaleless Lagrangian of massless QCD, giving rise to energy-dependence of one-scale observables that would be energy-independent in the classical theory.
- Factorization and evolution, which allow us to use perturbation theory to calculate the interactions of hadrons, since all the non-perturbative physics gets factorized, into universal functions that can be measured in one process, like DIS, and then used to predict the cross sections for any other process. Again, this introduces a scale dependence into the parton model so that the structure functions of DIS, and other one-scale observables such as the Drell–Yan cross section, become scale dependent.
- Infrared safety, which ensure that the infrared singularities associated with soft and collinear emission cancel between real and virtual contributions, allowing the perturbative calculation of jet cross sections, without a detailed understanding of the mechanism by which partons become jets.

Together, these allow us to make sense of QCD, without having to solve the theory at all possible scales: unknown or uncalculable high- and low-energy effects can be renormalized, factorized and cancelled away. After all this, it is remarkable that most QCD phenomenology can be understood at least qualitatively from leading order perturbation theory with the one-loop renormalization group and DGLAP evolution equations. Higher order corrections, while essential for quantitative analysis, do not change this simple picture dramatically.

Acknowledgements

It is a pleasure to acknowledge the organizers, lecturers, tutors and students of the Latin American School of High Energy Physics at Recinto Quirama for making giving these lectures such an enjoyable experience.

References

[1] R.K. Ellis, W.J. Stirling, and B.R. Webber, *QCD and Collider Physics*, Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology, Volume 8 (Cambridge University Press, 1996).

[2] M.E. Peskin and D.V. Schroeder, *An Introduction to Quantum Field Theory* (Addison-Wesley, 1995).

[3] M. Gockeler, R. Horsley, A. C. Irving, D. Pleiter, P. E. L. Rakow, G. Schierholz, and H. Stuben, *A determination of the Lambda parameter from full lattice QCD*, Phys. Rev. D **73** (2006) 014513 [arXiv:hep-ph/0502212].

[4] M. Davier, S. Eidelman, A. Höcker, and Z. Zhang, *Confronting spectral functions from e+ e- annihilation and tau decays: Consequences for the muon magnetic moment*, Eur. Phys. J. C **27** (2003) 497 [arXiv:hep-ph/0208177].

[5] M. Davier, S. Descotes-Genon, A. Höcker, B. Malaescu, and Z. Zhang, *The determination of α_s from τ decays revisited*, Eur. Phys. J. C **56** (2008) 305 [arXiv:0803.0979 [hep-ph]].

[6] S. Bethke, *Experimental tests of asymptotic freedom*, Prog. Part. Nucl. Phys. **58** (2007) 351 [arXiv:hep-ex/0606035].

Beyond the Standard Model for Montañeros*

M. Bustamante[1], *L. Cieri*[2] *and J. Ellis*[3]
[1] Pontificia Universidad Católica del Perú, Lima, Peru
[2] Universidad de Buenos Aires, Buenos Aires, Argentina
[3] CERN, Geneva, Switzerland

Abstract

These notes cover (i) *electroweak symmetry breaking* in the Standard Model (SM) and the Higgs boson, (ii) *alternatives to the SM Higgs boson* including an introduction to composite Higgs models and Higgsless models that invoke extra dimensions, (iii) the theory and phenomenology of *supersymmetry*, and (iv) various *further beyond topics*, including Grand Unification, proton decay and neutrino masses, supergravity, superstrings and extra dimensions.

1 The Standard Model, electroweak symmetry breaking and the Higgs boson

In this first Lecture, we review the electroweak sector of the Standard Model (SM) (for more detailed accounts, see, e.g., [1–3]), with particular emphasis on the nature of electroweak symmetry breaking. The theory grew out of experimental information on charged-current weak interactions, and of the realisation that the four-point Fermi description ceases to be valid above $\sqrt{s} = 600$ GeV [3]. Electroweak theory was able to predict the existence of neutral-current interactions, as discovered by the Gargamelle Collaboration in 1973 [4]. One of its greatest subsequent successes was the detection in 1983 of the $W^{\pm}$ and Z^0 bosons [5–8], whose existences it had predicted. Over time, thanks to the accumulating experimental evidence, the $SU(2)_L \otimes U(1)_Y$ electroweak theory and $SU(3)_C$ quantum electrodynamics, collectively known as the Standard Model, have come to be regarded as the correct description of electromagnetic, weak and strong interactions up to the energies that have been probed so far. However, although the SM has many successes, it also has some shortcomings, as we also indicate. In subsequent Lectures we discuss ideas for rectifying (at least some of) these defects: see also [9–11].

The particle content of the SM is summarized in Table 1. Within the SM, the electromagnetic and weak interactions are described by a Lagrangian that is symmetric under local weak isospin and hypercharge gauge transformations, described using the $SU(2)_L \otimes U(1)_Y$ group (the L subindex refers to the fact that the weak $SU(2)$ group acts only the left-handed projections of fermion states; Y is the hypercharge). We can write the $SU(2)_L \otimes U(1)_Y$ part of the SM Lagrangian as

$$\begin{aligned} \mathcal{L} &= -\frac{1}{4}\mathbf{F}^a_{\mu\nu}\mathbf{F}^{a\mu\nu} \\ &+ i\overline{\psi}\not{D}\psi + h.c. \\ &+ \psi_i y_{ij}\psi_j\phi + h.c. \\ &+ |D_\mu\phi|^2 - V(\phi) \ . \end{aligned} \tag{1}$$

This is short enough to write on a T-shirt!

The first line is the kinetic term for the gauge sector of the electroweak theory, with a running over the total number of gauge fields: three associated with $SU(2)_L$, which we shall call B^1_μ, B^2_μ, B^3_μ, and one with $U(1)_Y$, which we shall call $\mathcal{A}_\mu$. Their field-strength tensors are

$$F^a_{\mu\nu} = \partial_\nu B^a_\mu - \partial_\mu B^a_\nu + g\varepsilon_{bca}B^b_\mu B^c_\nu \ \text{ for a} = 1,2,3 \tag{2}$$

$$f_{\mu\nu} = \partial_\nu \mathcal{A}_\mu - \partial_\mu \mathcal{A}_\nu \ . \tag{3}$$

*Based on lectures by John Ellis at the 2009 CERN–CLAF School of High-Energy Physics, Medellín, Colombia.

Table 1: Particle content of the Standard Model with a minimal Higgs sector.

Bosons	**Scalars**
γ, W^+, W^-, Z^0, $g_{1\ldots8}$	ϕ (Higgs)
Fermions	
Quarks (each with 3 colour charges)	Leptons
$2/3:$ $\begin{pmatrix} u \\ d \end{pmatrix}, \begin{pmatrix} c \\ s \end{pmatrix}, \begin{pmatrix} t \\ b \end{pmatrix}$	neutral: $\begin{pmatrix} \nu_e \\ e^- \end{pmatrix}, \begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}, \begin{pmatrix} \nu_\tau \\ \tau^- \end{pmatrix}$
$-1/3:$	$-1:$

In Eq. (2), g is the coupling constant of the weak-isospin group $SU(2)_L$, and the ε_{bca} are its structure constants. The last term in this equation stems from the non-Abelian nature of $SU(2)$. At this point, all of the gauge fields are massless, but we will see later that specific linear combinations of the four electroweak gauge fields acquire masses through the Higgs mechanism.

The second line in Eq. (1) describes the interactions between the matter fields ψ, described by Dirac equations, and the gauge fields.

The third line is the Yukawa sector and incorporates the interactions between the matter fields and the Higgs field, ϕ, which are responsible for giving fermions their masses when electroweak symmetry breaking occurs.

The fourth and final line describes the scalar or Higgs sector. The first piece is the kinetic term with the covariant derivative defined here to be

$$D_\mu = \partial_\mu + \frac{ig'}{2}\mathcal{A}_\mu Y + \frac{ig}{2}\tau \cdot \mathbf{B}_\mu \,, \tag{4}$$

where g' is the $U(1)$ coupling constant, and Y and $\tau \equiv (\tau_1, \tau_2, \tau_3)$ (the Pauli matrices) are, respectively, the generators of $U(1)$ and $SU(2)$. The second piece of the final line of (1) is the Higgs potential $V(\phi)$.

Whereas the first two lines of (1) have been confirmed in many different experiments, there is **no** *experimental evidence for the last two lines and one of the main objectives of the LHC is to discover whether it is right, needs modification, or is simply wrong.*

1.1 The Higgs mechanism in $U(1)$

To explain the Higgs mechanism of mass generation, we first apply it to the gauge group $U(1)$, and then extend it to the full electroweak group $SU(2)_L \otimes U(1)_Y$. Thus, we first consider the following Lagrangian for a single complex scalar field:

$$\mathcal{L} = (\partial_\mu \phi)^* (\partial^\mu \phi) - V(\phi^* \phi) \,, \tag{5}$$

with the potential defined as

$$V(\phi^* \phi) = \mu^2 (\phi^* \phi) + \lambda (\phi^* \phi)^2 \,, \tag{6}$$

where μ^2 and $\lambda > 0$ are real constants. This Lagrangian is clearly invariant under global $U(1)$ phase transformations

$$\phi \to e^{i\alpha}\phi \,, \tag{7}$$

for α some rotation angle. Equivalently, it is invariant under a $SO(2)$ rotational symmetry, which is made evident by writing $\mathcal{L}$ in terms of the decomposition of the complex scalar field into two real fields ϕ_1 and ϕ_2: $\phi \equiv \phi_1 + i\phi_2$.

If we choose $\mu^2 > 0$ in (8), the sole vacuum state has $\langle\phi\rangle = 0$. Perturbing around this vacuum reveals that, in this case, the scalar-sector Lagrangian simply factors into two Klein–Gordon Lagrangians, one for ϕ_1 and the other for ϕ_2, with a common mass. The symmetry of the original Lagrangian is preserved in this case.

However, when $\mu^2 < 0$, the Lagrangian (5) exhibits spontaneous breaking of the $U(1)$ global symmetry, which introduces a massless scalar particle known as a Goldstone boson, as we now show. In order to make manifest this breaking of the $U(1)$ symmetry present in Eq. (5), we first minimize the potential (6) so as to identify the vacuum expectation value, or v.e.v., of the scalar field. To do this, we first write the Higgs potential as

$$V\left(\phi^*\phi\right) = \mu^2\left(\phi_1^2 + \phi_2^2\right) + \lambda\left(\phi_1^2 + \phi_2^2\right)^2 \ , \tag{8}$$

and note that minimization with respect to $\phi^*\phi$ yields the value

$$\phi_1^2 + \phi_2^2 = -\mu^2/\left(2\lambda\right) \ , \tag{9}$$

i.e., there is a set of equivalent minima lying around a circle of radius $\sqrt{-\mu^2/\left(2\lambda\right)}$, when $\mu^2 < 0$ as assumed. The quanta of the Higgs field arise when a particular ground state is chosen and perturbed. Reflecting the appearance of spontaneous symmetry breaking we may, without loss of generality, choose for instance

$$\phi_{1,\text{vac}} = \sqrt{-\mu^2/\left(2\lambda\right)} \equiv v/\sqrt{2} \quad , \quad \phi_{2,\text{vac}} = 0 \ . \tag{10}$$

Perturbations around this vacuum may be parametrized by

$$\eta/\sqrt{2} \equiv \phi_1 - v/\sqrt{2} \quad , \quad \xi/\sqrt{2} \equiv \phi_2 \ , \tag{11}$$

so that the perturbed complex scalar is $\phi = \left(v + \eta + i\xi\right)/\sqrt{2}$, where η and ξ are real fields. In terms of these, the Lagrangian becomes

$$\begin{aligned} \mathcal{L} &= \left[\frac{1}{2}\left(\partial^\mu\eta\right)\left(\partial_\mu\eta\right) - \frac{\mu^2}{2}\eta^2\right] + \frac{1}{2}\left(\partial^\mu\xi\right)\left(\partial_\mu\xi\right) \\ &- \frac{\lambda}{2}\left[\left(v+\eta\right)^2 + \xi^2\right]^2 - \mu^2 v\eta - \frac{\mu^2}{2}\xi^2 - \frac{1}{2}\mu^2 v^2 \ . \end{aligned} \tag{12}$$

The first and second terms describe two scalar particles: the first, η, is massive with $m_\eta^2 = -\mu^2 > 0$ (we recall that $\mu^2 < 0$), and the second, ξ, is massless, the Goldstone boson.

We now discuss how this spontaneous symmetry breaking manifests itself in the presence of a $U(1)$ gauge field. For this purpose, we make the Lagrangian (5) invariant under local $U(1)$ phase transformations, i.e.,

$$\phi \to e^{i\alpha(x)}\phi \ . \tag{13}$$

This requires the introduction of a gauge field $\mathcal{A}_\mu$ that transforms as follows under $U(1)$:

$$\mathcal{A}'_\mu \to \mathcal{A}_\mu + \left(1/q\right)\partial_\mu\alpha\left(x\right) \ , \tag{14}$$

and replacing the space-time derivatives by covariant derivatives

$$D_\mu = \partial_\mu + iq\mathcal{A}_\mu \ , \tag{15}$$

where q is the conserved charge. Replacing the derivatives in Eq. (5) and adding a kinetic term for the $\mathcal{A}_\mu$ field, the Lagrangian becomes

$$\mathcal{L} = \left[\left(\partial_\mu - iq\mathcal{A}_\mu\right)\phi^*\right]\left[\left(\partial^\mu + iq\mathcal{A}^\mu\right)\phi\right] - V\left(\phi^*\phi\right) - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} \ . \tag{16}$$

The last term in this equation, $\left(1/4\right)F^{\mu\nu}F_{\mu\nu}$, with $F_{\mu\nu} \equiv \partial_\nu\mathcal{A}_\mu - \partial_\mu\mathcal{A}_\nu$, is the kinetic term, which is separately invariant under the transformation (14) of the gauge field.

We now repeat the minimization of the potential $V(\phi)$ and write the Lagrangian in terms of the perturbations around the ground state, Eqs. (11):

$$\begin{aligned}\mathcal{L} &= \left\{\frac{1}{2}\left[(\partial^\mu\eta)(\partial_\mu\eta) - \mu^2\eta^2\right] + \frac{1}{2}(\partial^\mu\xi)(\partial_\mu\xi) - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} + \frac{1}{2}q^2v^2\mathcal{A}^\mu\mathcal{A}_\mu\right\} \\ &+ vq^2A^\mu\mathcal{A}_\mu\eta + \frac{q^2}{2}\mathcal{A}^\mu\mathcal{A}_\mu\eta^2 + q(\partial^\mu\xi)\mathcal{A}_\mu(v+\eta) - q(\partial^\mu\eta)\mathcal{A}_\mu\xi \\ &- \mu^2v\eta - \frac{\mu^2}{2}\xi^2 - \frac{\lambda}{2}\left[(v+\eta)+\xi^2\right]^2 - \frac{\mu^2v}{2}\,. \end{aligned} \tag{17}$$

The first three terms again describe a (real) scalar particle, η, of mass $\sqrt{-\mu^2}$ and a massless Goldstone boson, ξ. The fourth term describes the free gauge field. However, whereas previously the Lagrangian described a massless boson field [see Eq. (12)], now it contains a term proportional to $\mathcal{A}_\mu\mathcal{A}^\mu$, which gives the gauge field a mass of

$$m_\mathcal{A} = qv\,, \tag{18}$$

from which we see that the boson field has acquired a mass that is proportional to the vacuum expectation value of the Higgs field. Indeed, the last two terms in the first line of Eq. (12) are identical with the Proca Lagrangian for a $U(1)$ gauge boson of mass m.

The rest of the terms in Eq. (12) define couplings between the fields A^μ, η and ξ, among which is a bilinear interaction coupling A^μ and $\partial_\mu\xi$. In order to give the correct propagating particle interpretation of (12), we must diagonalize the bilinear terms and remove this term. This is easily done by exploiting the gauge freedom of the $\mathcal{A}_\mu$ field to replace

$$\mathcal{A}_\mu \to \mathcal{A}'_\mu = \mathcal{A}_\mu + \frac{1}{qv}\partial_\mu\xi\,, \tag{19}$$

which is accompanied by the local phase transformation

$$\phi \to \phi' = e^{-i\xi(x)/v}\phi = (v+\eta)/\sqrt{2}\,. \tag{20}$$

After making this transformation, the field ξ no longer appears, and the Lagrangian (12) takes the simplified form

$$\mathcal{L} = \frac{1}{2}\left[(\partial^\mu)(\partial_\mu) - \mu^2\eta^2\right] - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} + \frac{q^2v^2}{2}\mathcal{A}^{\mu\,\prime}\mathcal{A}'_\mu + \ldots\,. \tag{21}$$

where the $\ldots$ represent trilinear and quadrilinear interactions.

The interpretation of (21) is that the Goldstone boson ξ that appeared when the global $U(1)$ symmetry was broken by the choice of an asymmetric ground state when $\mu^2 < 0$ has been absorbed (or 'eaten') by the gauge field $\mathcal{A}_\mu$, with the effect of generating a mass. Another way to understand this is to recall that, whereas a massless gauge boson has only two degrees of freedom, or polarization states (which are transverse), a massive gauge boson must have a third (longitudinal) polarization state. In the Higgs mechanism, this is supplied by the Goldstone boson of the spontaneously-broken $U(1)$ global symmetry.

At first sight, the Higgs mechanism may seem somewhat artificial. From one point of view, it is merely a description of the breaking of electroweak symmetry, rather than an explanation of how a massless gauge boson may become massive. As Quigg says [12], the electroweak symmetry is broken because $\mu^2 < 0$, and we must choose $\mu^2 < 0$, because otherwise electroweak symmetry is not broken. From another point of view, the *only* consistent formulation of an interacting massive gauge boson is *via* the Higgs mechanism, and the spontaneous breaking of symmetry is a mathematical ruse for describing this phenomenon.

1.2 The Higgs mechanism in $SU(2)_L \otimes U(1)_Y$

Following closely in both spirit and notation the book by Quigg [12], we now consider the weak-isospin doublet

$$\mathrm{L} = \begin{pmatrix} \nu \\ e \end{pmatrix}_L , \tag{22}$$

with the left-handed neutrino and electron states defined by

$$\nu_L = \frac{1}{2}\left(1-\gamma_5\right)\nu \quad , \quad e_L = \frac{1}{2}\left(1-\gamma_5\right)e \, . \tag{23}$$

The operator $\left(1-\gamma_5\right)/2$ is of course the left-handed helicity projector, and ν, e are solutions of the free-field Dirac equation. Within the SM, we consider the neutrino to be massless, and it does not have a corresponding right-handed component, i.e.,

$$\nu_R = \frac{1}{2}\left(1+\gamma_5\right)\nu = 0 \, . \tag{24}$$

Hence, the only right-handed lepton, e_R, constitutes a weak-isospin singlet, i.e.,

$$\mathrm{R} = e_R = \frac{1}{2}\left(1+\gamma_5\right)e \, . \tag{25}$$

We write initially the Lagrangian as

$$\mathcal{L} = \mathcal{L}_{\text{gauge}} + \mathcal{L}_{\text{leptons}} \tag{26}$$

$$\mathcal{L}_{\text{gauge}} = -\frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu} - \frac{1}{4}f_{\mu\nu}f^{\mu\nu} \tag{27}$$

$$\mathcal{L}_{\text{leptons}} = \overline{\mathrm{R}}\left(\partial_\mu + i\frac{g'}{2}\mathcal{A}_\mu Y\right)\mathrm{R} + \overline{\mathrm{L}}i\gamma^\mu\left(\partial_\mu + i\frac{g'}{2}\mathcal{A}_\mu Y + i\frac{g}{2}\tau\cdot\mathbf{B}_\mu\right)\mathrm{L} \, , \tag{28}$$

where the field-strength tensors, $F_{\mu\nu}$ and $f_{\mu\nu}$, were defined in Eqs. (2) and (3), respectively. Here, $g'/2$ is the coupling constant associated to the hypercharge group $U(1)_Y$, and $g/2$ is the coupling to the weak-isospin group $SU(2)_L$. So far, we are presented with four massless bosons ($\mathcal{A}_\mu$, B^1_μ, B^2_μ, B^3_μ); the Higgs mechanism will select linear combinations of these to produce three massive bosons ($W^\pm$, Z^0) and a massless one (γ).

The Higgs field is now a complex $SU(2)$ doublet

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} , \tag{29}$$

with ϕ^+ and ϕ^0 scalar fields. We need to add the Lagrangian

$$\mathcal{L}_{\text{Higgs}} = \left(D_\mu\phi\right)^\dagger\left(D^\mu\phi\right) - V\left(\phi^\dagger\phi\right) \, , \tag{30}$$

with the Higgs potential given by analogy to Eq. (6) as

$$V\left(\phi^\dagger\phi\right) = \mu^2\left(\phi^\dagger\phi\right) + \lambda\left(\phi^\dagger\phi\right)^2 \, , \tag{31}$$

with $\lambda > 0$. We should also include the interaction Lagrangian between this scalar field and the fermionic matter fields, which occurs through Yukawa couplings,

$$\mathcal{L}_{\text{Yukawa}} = -G_e\left[\overline{\mathrm{R}}\phi^\dagger\mathrm{L} + \overline{\mathrm{L}}\phi\mathrm{R}\right] \, . \tag{32}$$

As we see later, these terms give rise to masses for the matter fermions.

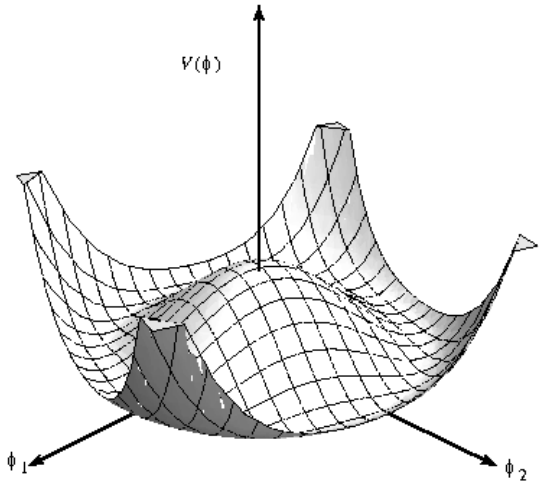


Fig. 1: Scalar potential $V\left(\phi^\dagger\phi\right)$ with $\lambda > 0$ and $\mu^2 < 0$

A plot of the Higgs potential is presented in Fig. 1.2, where we see that $\langle\phi\rangle = 0$ is an unstable local minimum of the effective potential if $\mu^2 < 0$, and that the minimum is at some $\langle\phi\rangle \neq 0$ with an arbitrary phase, leading to spontaneous symmetry breaking. Minimizing the Higgs potential, we obtain

$$\frac{\partial}{\partial\left(\phi^\dagger\phi\right)}V\left(\phi^\dagger\phi\right) = \mu^2 + 2\lambda\langle\phi\rangle_0 = \mu^2 + 2\lambda\left[\left(\phi^+_{\text{vac}}\right)^2 + \left(\phi^0_{\text{vac}}\right)^2\right] = 0\,. \tag{33}$$

Choosing $\phi^+_{\text{vac}} = 0$ and $\phi^0_{\text{vac}} = \sqrt{-\mu^2/\left(2\lambda\right)}$, the v.e.v. of the scalar field becomes

$$\langle\phi\rangle_0 = \begin{pmatrix} 0 \\ v/\sqrt{2} \end{pmatrix}\,, \tag{34}$$

with $v \equiv \sqrt{-\mu^2/\lambda}$. Selecting a particular v.e.v. breaks, of course, both $SU(2)_L$ and $U(1)_Y$ symmetries. Nevertheless, an invariance under the $U(1)_{\text{EM}}$ symmetry is preserved, with the charge operator as the generator. In the preceding section, we saw one example of the general theorem that, for every broken generator (i.e., every generator that does not leave the vacuum invariant), there would (in the absence of the Higgs mechanism) be a Goldstone boson.

In general, a generator $\mathcal{G}$ leaves the vacuum invariant if

$$e^{i\alpha\mathcal{G}}\langle\phi\rangle_0 \simeq \left(1 + i\alpha\mathcal{G}\right)\langle\phi\rangle_0 = \langle\phi\rangle_0\,, \tag{35}$$

which is satisfied when $\mathcal{G}\langle\phi\rangle_0 = 0$. Let's test whether the generators of $SU(2)_L \otimes U(1)_Y$ satisfy this condition:

$$\tau_1\langle\phi\rangle_0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ v/\sqrt{2} \end{pmatrix} = \begin{pmatrix} v/\sqrt{2} \\ 0 \end{pmatrix} \tag{36}$$

$$\tau_2\langle\phi\rangle_0 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}\begin{pmatrix} 0 \\ v/\sqrt{2} \end{pmatrix} = \begin{pmatrix} -iv/\sqrt{2} \\ 0 \end{pmatrix} \tag{37}$$

$$\tau_3\langle\phi\rangle_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 0 \\ v/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 0 \\ -v/\sqrt{2} \end{pmatrix} \tag{38}$$

$$Y\langle\phi\rangle_0 = \langle\phi\rangle_0\,. \tag{39}$$

Thus, none of the generators leave the vacuum invariant. However, we note that

$$Q\langle\phi\rangle_0 = \frac{1}{2}\left(\tau_3 + Y\right)\langle\phi\rangle_0 = 0\,, \tag{40}$$

which is what we expected: the linear combination of generators corresponding to electric charge remains unbroken. Correspondingly, as we shall now see, whilst the photon remains massless, the other three gauge bosons acquire mass.

To see this, we now consider perturbations around the choice of vacuum. The full perturbed scalar field is

$$\phi = \exp\left(\frac{i\xi\cdot\tau}{2v}\right)\begin{pmatrix} 0 \\ (v+\eta)/\sqrt{2}\end{pmatrix} . \tag{41}$$

However, in analogy to what we did for the $U(1)$ Higgs in the previous section to rotate the Goldstone boson ξ away, we are also able here to gauge-transform the scalar ϕ and the gauge and matter fields, i.e.,

$$\phi \rightarrow \phi' = \exp\left(\frac{-i\xi\cdot\tau}{2v}\right)\phi = \begin{pmatrix} 0 \\ (v+\eta)/\sqrt{2}\end{pmatrix} . \tag{42}$$

$$\tau\cdot\mathbf{B}_\mu \rightarrow \tau\cdot\mathbf{B}'_\mu \tag{43}$$

$$\mathrm{L} \rightarrow \mathrm{L}' = \exp\left(\frac{-i\xi\cdot\tau}{2v}\right)\mathrm{L} , \tag{44}$$

while the $\mathcal{A}_\mu$ and R remain invariant. It is possible to show that $\tau\cdot\mathbf{B}'_\mu = \tau\cdot\mathbf{B}_\mu - \xi\times\mathbf{B}_\mu\cdot\tau - (1/g)\,\partial_\mu(\xi\cdot\tau)$.

In the unitary gauge, we can write the perturbed state as

$$\langle\phi\rangle_0 \rightarrow \phi = \begin{pmatrix} 0 \\ (v+\eta)/\sqrt{2}\end{pmatrix} , \tag{45}$$

and the Lagrangian in the Yukawa sector, Eq. (32), becomes

$$\mathcal{L}_{\text{Yukawa}} = -G_e\left[\overline{e}_R\phi^\dagger\begin{pmatrix}\nu_L \\ e_L\end{pmatrix} + (\overline{\nu}_L\ \overline{e}_L)\,\phi e_R\right] = -G_e\frac{v+\eta}{\sqrt{2}}\left(\overline{e}_R e_L + \overline{e}_L e_R\right) . \tag{46}$$

Defining $\overline{e}\equiv(\overline{e}_R,\overline{e}_L)$ and $e\equiv(e_L,e_R)^T$ yields

$$\mathcal{L}_{\text{Yukawa}} = -\frac{G_e v}{\sqrt{2}}\overline{e}e - \frac{G_e\eta}{\sqrt{2}}\overline{e}e , \tag{47}$$

so that the electron has acquired a mass

$$m_e = G_e v/\sqrt{2} . \tag{48}$$

Clearly, this mechanism may be applied to all the SM fermions, with the general feature that their masses are proportional to their Yukawa couplings to the Higgs field [1]. This implies that the preferred decays of a Higgs boson into generic fermions f are into heavier species, as long as the Higgs mass $> 2m_f$.

To see the effect of spontaneous symmetry breaking on the scalar-sector Lagrangian, $\mathcal{L}_{\text{Higgs}}$ in Eq. (30), it is useful to calculate first

$$\phi^\dagger\phi = \left(\frac{v+\eta}{\sqrt{2}}\right)^2 , \tag{49}$$

so that

$$V\left(\phi^\dagger\phi\right) = \mu^2\left(\frac{v+\eta}{\sqrt{2}}\right)^2 + \lambda\left(\frac{v+\eta}{\sqrt{2}}\right)^4 , \tag{50}$$

and we also need

$$D_\mu\phi = \partial_\mu\phi + \frac{ig'}{2}\mathcal{A}_\mu Y\phi + \frac{ig}{2}\tau\cdot\mathbf{B}_\mu\phi , \tag{51}$$

[1] The Higgs couplings to quarks also induce their Cabibbo–Kobayashi–Maskawa mixing — see Eq. (93) below.

whose first term is simply

$$\partial_\mu \phi = \begin{pmatrix} 0 \\ \partial_\mu \eta/\sqrt{2} \end{pmatrix} . \tag{52}$$

Using Eqs. (36)–(39), we calculate the second and third terms, i.e.,

$$\frac{ig'}{2}\mathcal{A}_\mu Y \phi = \frac{ig'}{2}\mathcal{A}_\mu \phi = \frac{ig'}{2}\mathcal{A}_\mu \begin{pmatrix} 0 \\ (v+\eta)/\sqrt{2} \end{pmatrix}, \tag{53}$$

$$(\tau \cdot \mathbf{B}_\mu)\,\phi = B^1_\mu \begin{pmatrix} (v+\eta)/\sqrt{2} \\ 0 \end{pmatrix} + B^2_\mu \begin{pmatrix} -i\,(v+\eta)/\sqrt{2} \\ 0 \end{pmatrix} + B^3_\mu \begin{pmatrix} 0 \\ -(v+\eta)/\sqrt{2} \end{pmatrix} \tag{54}$$

Hence,

$$D_\mu \phi = \begin{pmatrix} \frac{ig}{2}\left(\frac{v+\eta}{\sqrt{2}}\right)\left(B^1_\mu - iB^2_\mu\right) \\ \frac{1}{\sqrt{2}}\partial_\mu \eta + \left(\frac{v+\eta}{\sqrt{2}}\right)\frac{i}{2}\left(ig'\mathcal{A}_\mu - igB^3_\mu\right) \end{pmatrix} \tag{55}$$

and

$$(D^\mu \phi)^\dagger (D_\mu \phi) = \frac{g^2}{8}(v+\eta)^2 |B^1_\mu - iB^2_\mu|^2 + \frac{1}{2}(\partial_\mu \eta)(\partial^\mu \eta) + \frac{1}{8}(v+\eta)^2 \left(g'\mathcal{A}_\mu - gB^3_\mu\right)^2 . \tag{56}$$

With this, the scalar-sector Lagrangian becomes

$$\begin{aligned} \mathcal{L}_{\rm Higgs} &= \left\{ \frac{1}{2}(\partial_\mu \eta)(\partial^\mu \eta) - \frac{\mu^2}{2}\eta^2 + \frac{v^2}{8}\left[g^2 |B^1_\mu - iB^2_\mu|^2 + \left(g'\mathcal{A}_\mu - gB^3_\mu\right)^2 \right] \right\} \\ &+ \left\{ \frac{1}{8}\left(\eta^2 + 2v\eta\right)\left[g^2 |B^1_\mu - iB^2_\mu|^2 + \left(g'\mathcal{A}_\mu - gB^3_\mu\right)^2 \right] \right. \\ &- \left. \frac{1}{4}\eta^4 - \lambda v \eta^3 - \frac{3}{2}\lambda v^2 \eta^2 - \left(\lambda v^3 + \mu^2 v\right)\eta - \left(\frac{\lambda v^4}{4} + \frac{\mu^2 v^2}{2}\right) \right\} . \end{aligned} \tag{57}$$

From the second term inside the first curly brackets, we see that the η field has acquired a mass; indeed, it is the Higgs boson, with non-zero mass. The terms inside the second curly brackets either describe interactions between the gauge and Higgs fields, or are constants that do not affect the physics.

It is convenient to define the charged gauge fields $W^\pm_\mu$ as linear combinations of the massless fields B^1_μ and B^2_μ, i.e.,

$$W^\pm_\mu \equiv \frac{B^1_\mu \mp iB^2_\mu}{\sqrt{2}} , \tag{58}$$

and, analogously,

$$Z_\mu \equiv \frac{-g'\mathcal{A}_\mu + gB^3_\mu}{\sqrt{g^2 + g'^{\,2}}} , \tag{59}$$

$$A_\mu \equiv \frac{g\mathcal{A}_\mu + g'B^3_\mu}{\sqrt{g^2 + g'^{\,2}}} . \tag{60}$$

Writing the original fields $\mathcal{A}_\mu$, B^i_μ in terms of the new fields, we have

$$B^1_\mu = \frac{\sqrt{2}}{2}\left(W^-_\mu + W^+_\mu\right) \quad , \quad B^2_\mu = \frac{\sqrt{2}}{2}\left(W^-_\mu - W^+_\mu\right) , \tag{61}$$

$$B^3_\mu = \frac{g'}{\sqrt{g^2 + g'^{\,2}}}\left(A_\mu + \frac{g}{g'}Z_\mu\right) \quad , \quad \mathcal{A}_\mu = \frac{g}{\sqrt{g^2 + g'^{\,2}}}\left(A_\mu - \frac{g'}{g}Z_\mu\right) . \tag{62}$$

Making these replacements in the broken scalar-sector Lagrangian, Eq. (57), leads to

$$\mathcal{L}_{\rm Higgs} = \left[\frac{1}{2}(\partial^\mu \eta)(\partial_\mu \eta) - \frac{\mu^2}{2}\eta^2\right] + \frac{v^2 g^2}{8} W^{+\,\mu} W^+_\mu + \frac{v^2 g^2}{8} W^{-\,\mu} W^-_\mu + \frac{\left(g^2 + g'^{\,2}\right)v^2}{8} Z^\mu Z_\mu$$

$$+ \quad \ldots \,, \tag{63}$$

and it is evident now that while the photon field $\mathcal{A}_\mu$ is massless due to the unbroken $U(1)_{\rm EM}$ symmetry (i.e., the symmetry under $e^{iQ\alpha(x)}$ rotations), the vector bosons $W^\pm$ and Z^0 have masses

$$m_W = gv/2 \quad , \quad m_Z = (v/2)\,\sqrt{g^2 + g'^{\,2}}\,. \tag{64}$$

We see again that the Higgs couplings to other particles, in this case the $W^\pm$ and Z^0, are related to their masses.

We also see that the masses of the neutral and charged weak-interaction bosons are related through

$$m_Z = m_W \sqrt{1 + g'^{\,2}/g^2}\,. \tag{65}$$

Experimentally, the weak gauge boson masses are known to high accuracy to be [13]

$$m_W = 80.399 \pm 0.023 \text{ GeV} \quad , \quad m_Z = 91.1875 \pm 0.0021 \text{ GeV}\,, \tag{66}$$

which can be compared in detail with (65) only after the inclusions of radiative corrections. Meanwhile, the current experimental upper limit on the photon mass, based on plasma physics, is very stringent: $m_\gamma < 10^{-18}$ eV [14]. For the Higgs mass, we see from (57) that

$$m_H = -2\mu^2\,. \tag{67}$$

A priori, however, there is no theoretical prediction within the Standard Model, since μ is not determined by any of the known parameters of the Standard Model. Later we will see various ways in which experiments constrain the Higgs mass.

We can introduce a weak mixing angle θ_W to parametrize the mixing of the neutral gauge bosons, defined by

$$\tan(\theta_W) = g'/g\,, \tag{68}$$

so that

$$\cos(\theta_W) = \frac{g}{\sqrt{g^2 + g'^{\,2}}} \quad , \quad \sin(\theta_W) = \frac{g'}{\sqrt{g^2 + g'^{\,2}}}\,. \tag{69}$$

With this, we can write, from Eqs. (59) and (60),

$$Z_\mu = -\sin(\theta_W)\,\mathcal{A}_\mu + \cos(\theta_W)\,B^3_\mu\,, \tag{70}$$

$$A_\mu = \cos(\theta_W)\,\mathcal{A}_\mu + \sin(\theta_W)\,B^3_\mu\,. \tag{71}$$

The relation (65) between the masses of $W^\pm$ and Z^0 becomes

$$m_W = m_Z \cos(\theta_W)\,, \tag{72}$$

and it is common practice to define the ratio

$$\rho = \frac{m_W^2}{m_Z^2 \cos^2(\theta_W)}\,. \tag{73}$$

According to the Standard Model, this is equal to unity at the tree level, a prediction that has been well tested by experiment, including radiative corrections. The value of $\sin^2(\theta_W)$ is obtained from measurements of the Z pole and neutral-current processes, and depends on the renormalization prescription. The 2008 Particle Data Group review [13] states values of $\sin^2(\theta_W) = 0.2319(14)$ and $\rho = 1.0004^{+0.0008}_{-0.0004}$.

Therefore, after the spontaneous breaking of the electroweak $SU(2)_L \otimes U(1)_Y$ symmetry, we have ended up with what we desired: three massive gauge bosons ($W^\pm$, Z^0) that mediate weak interactions, one massless gauge boson (A) corresponding to the photon, and an extra, massive, Higgs boson (H).

1.3 QCD

The QCD Lagrangian has a structure similar to that of the electroweak Lagrangian [13], being also a gauge theory, but based on the group $SU(3)$ and without spontaneous symmetry breaking:

$$\mathcal{L}_{\rm QCD} = -\frac{1}{4}F^a_{\mu\nu}F^{a\,\mu\nu} + i\sum_q \overline{\psi}^i_q \gamma^\mu \left(D_\mu\right)_{ij} \psi^j_q - \sum_q m_q \psi^i_q \psi_{qi} \,, \tag{74}$$

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - g_s f_{abc} A^b_\mu A^c_\nu \,, \tag{75}$$

$$\left(D_\mu\right)_{ij} = \delta_{ij}\partial_\mu + i g_s \sum_a \frac{\lambda^a_{i,j}}{2} A^a_\mu \,, \tag{76}$$

with g_s the strong coupling constant, f_{abc} the $SU(3)$ structure constants, and λ_i ($i = 1, \ldots, 8$) the generators of $SU(3)$ (which can be taken to be the eight traceless Gell-Mann matrices). Note also that ψ^i_q is the free-field Dirac spinor representing a quark of colour i and flavour q and the A^a_μ ($a = 1, \ldots, 8$) are the eight gluon fields. As is well known, QCD and non-Abelian gauge theories possess the property of asymptotic freedom: $\alpha_s \equiv g_s^2/4\pi$ obeys the renormalization-group equation (RGE) that determines its evolution as a function of the effective scale Q:

$$Q\frac{d\alpha_s}{dQ} = 2\beta_0\alpha_s + \ldots \,, \tag{77}$$

where

$$\beta_0 = 11 - \frac{2}{3}n_q \tag{78}$$

and n_q is the number of quark flavours with masses $\ll Q$. In addition to (76), which specifies QCD at the perturbative level, its full specification of its vacuum at the non-perturbative level requires an additional angle parameter, θ_{QCD}, that violates both parity P and CP [15] [2].

1.4 Parameters of the Standard Model

The transformation from being one of the possible explanations of electromagnetic, weak and strong phenomena into a description in outstanding agreement with experiments is reflected in the dozens of electroweak precision measurements available today [13, 16, 17]. These are sensitive to quantum corrections at and beyond the one-loop level, which are essential for obtaining agreement with the data. The calculations of these corrections rely upon the renormalizability (calculability) of the SM [3], and depend on the masses of heavy virtual particles, such as the top quark and the Higgs boson and possibly other particles beyond the SM. The consistency with the data may be used to constrain the masses of these particles.

Many of these observables have quadratic sensitivity to the mass of the top quark, e.g.,

$$s^2_W \equiv 1 - m^2_W/m^2_Z \ni -\frac{2\alpha}{16\pi\sin^2\left(\theta_W\right)}\frac{m^2_t}{m^2_Z} \,. \tag{79}$$

This effect was used before the discovery of the top quark to predict successfully its mass [18], and the consistency of the prediction with experiment can be used to constrain possible new physics beyond the SM, particularly mass-squared differences between isospin partner particles, that would contribute analogously to (79). Many electroweak observables are also logarithmically sensitive to the mass of the Higgs boson, e.g.,

$$s^2_W \ni \frac{5\alpha}{24\pi}\ln\left(\frac{m^2_H}{m^2_W}\right) \tag{80}$$

[2] The upper limit on the electric dipole moment of the neutron tells us that $|\theta_{QCD}| < \mathcal{O}(10^{-9})$ [13].

[3] A crucial aspect of this is cancellation of anomalous triangle diagrams between quarks and leptons, which may be a hint of an underlying Grand Unified Theory — see Lecture 4.

when $m_H \gg m_W$. If there were no Higgs boson, or nothing to do its job [4], radiative corrections such as (80) would diverge, and the SM calculations would become meaningless. Two examples of precision electroweak observables, namely the coupling of the Z^0 boson to leptons and the mass of the W boson, are shown in Fig. 2.

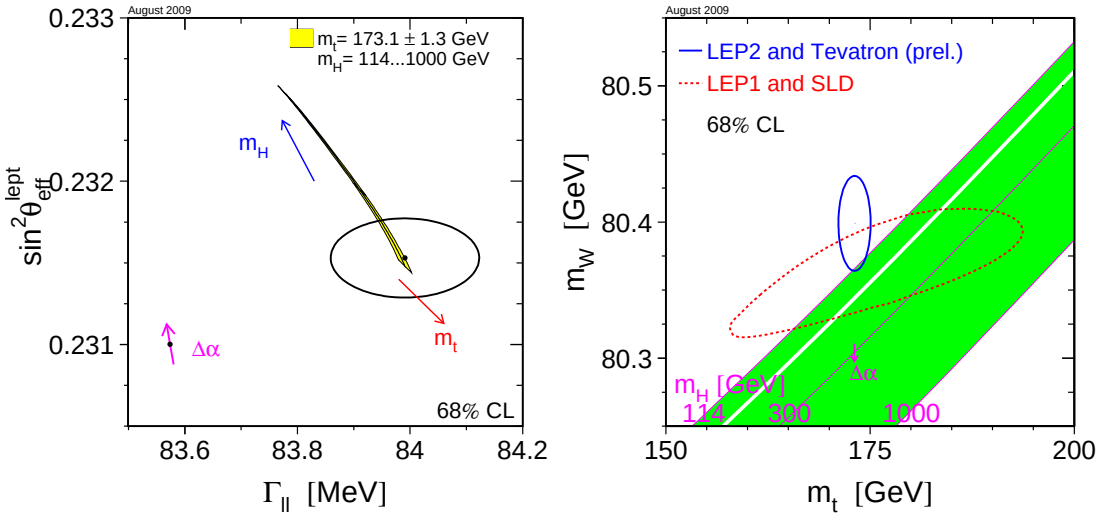


Fig. 2: Left: LEP and SLD measurements of $\sin^2\theta_W$ and the leptonic decay width of the Z^0, Γ_{ll}, compared with the SM prediction for different values of m_t and m_H. Right: The predictions for m_t and m_W made in the SM using LEP1 and SLD data (dotted mango-shaped contour) for different values of m_H, compared with the LEP2 and Tevatron measurements (ellipse). The arrows show the additional effects of the uncertainty in the value of α_{em} at the Z^0 peak [16].

Table 2 and Fig. 1.4 [17] compare the predicted (fitted) and experimentally measured values for several parameters of the Standard Model; the agreement is usually better than 1σ. This is a remarkable success for a theory that, as we have seen, can be written down in only a few lines.

The agreement of the precision electroweak observables with the SM can be used to predict m_H, just as it was used previously to predict m_t. Since the early 1990s [19], this method has been used to tighten the vise on the Higgs, providing ever-stronger indications that it is probably relatively light, as hinted in Fig. 4. The latest estimate of the Higgs mass is [16]

$$m_H \;=\; 89^{+35}_{-26}\ \text{GeV}. \tag{81}$$

Although the central value is somewhat below the lower limit of 114.4 GeV set by direct searches at LEP [20], there is consistency at the 1-σ level, and no significant discrepancy. *A priori*, the relatively light mass range (81) suggests that the Higgs boson interacts relatively weakly, with a small quartic coupling λ, though there is no theoretical consensus on this: see the discussion in the next Lecture.

This success is very impressive. However, our rejoicing is muted by the fact that to specify the SM we need at least 19 input parameters in order to calculate physical processes, namely:

- three coupling parameters, which we can choose to be the strong coupling constant, α_s, the fine structure constant, $\alpha_{\rm em}$, and the weak mixing angle, $\sin^2(\theta_W)$;
- two parameters that specify the shape of the Higgs potential, μ^2 and λ (or, equivalently, m_H and m_W or m_Z);
- six quark masses (or the six Yukawa couplings for the quarks);

[4] See Lecture 2 for a discussion of possible alternatives.

Table 2: Fit and experimental values of some SM quantities, as obtained using the `Gfitter` package [17]. For all the observables listed, except A_l (LEP) and A_l (SLD), the fit values shown are the results of 'complete fits', i.e., the results of using all the inputs, including the input value of the parameter that is being fit, to calculate the result. For the two exceptions, the fit values shown were calculated using all inputs except their own. Consult [17] for a description of each observable.

Parameter	Input value	Fit value
M_Z [GeV]	91.1875 ± 0.0021	91.1876 ± 0.0021
Γ_Z [GeV]	2.4952 ± 0.0023	2.4956 ± 0.0015
$\sigma^0_{\rm had}$	41.540 ± 0.037	41.478 ± 0.014
R^0_l	20.767 ± 0.025	20.741 ± 0.018
$A^{0,l}_{\rm FB}$	0.0171 ± 0.0010	0.01624 ± 0.0002
A_l (LEP)	0.1465 ± 0.0033	0.1473 ± 0.0009
A_l (SLD)	0.1513 ± 0.0021	$0.1465^{+0.0007}_{-0.0010}$
$\sin^2\phi^l_{\rm eff}\,(Q_{\rm FB})$	0.2324 ± 0.0012	$0.23151^{+0.00010}_{-0.00012}$
$A^{0,c}_{\rm FB}$	0.0707 ± 0.0035	0.0737 ± 0.0005
$A^{0,b}_{\rm FB}$	0.0992 ± 0.0016	$0.1032^{+0.0007}_{-0.0006}$
A_c	0.670 ± 0.027	$0.6679^{+0.00042}_{-0.00036}$
A_b	0.923 ± 0.020	$0.93463^{+0.00007}_{-0.00008}$
R^0_c	0.1721 ± 0.0030	0.17225 ± 0.00006
R^0_b	0.21629 ± 0.00066	0.21577 ± 0.00005
$\Delta\alpha^{(5)}_{\rm had}\left(M_Z^2\right)$	2768 ± 22	2764^{+22}_{-21}
M_W [GeV]	80.399 ± 0.023	$80.371^{+0.008}_{-0.011}$
Γ_W [GeV]	2.098 ± 0.048	2.092 ± 0.001
$\overline{m}_c$ [GeV]	1.25 ± 0.09	1.25 ± 0.09
$\overline{m}_b$ [GeV]	4.20 ± 0.07	4.20 ± 0.07
m_t [GeV]	173.1 ± 1.3	173.6 ± 1.2

- four parameters (three mixing angles and one weak CP-violating angle) for the Cabibbo-Kobayashi-Maskawa matrix [see Eq. (93) below];
- three charged-lepton masses (or the corresponding Yukawa couplings);
- one parameter to allow for non-perturbative CP violation in QCD, $\theta_{\rm QCD}$.

Moreover, because we now know that neutrinos have mass and that they mix (see, e.g., [21, 22]), the Standard Model must be extended to incorporate this fact. Therefore, we also need to specify three neutrino masses and three mixing angles plus a CP-violating phase for the neutrino mixing matrix, bringing the grand total to 26 parameters. Additionally, if neutrinos turn out to be Majorana particles, so that they are their own antiparticles, two more CP-violating phases need to be specified. Notice that at least 20 of the parameters relate to flavour physics.

Many of the ideas for physics beyond the SM that are discussed later have been motivated by attempts to reduce the number of its parameters, or understand their origins, or at least to make them seem less unnatural, as discussed in subsequent Lectures.

1.5 Bounds on the Standard Model Higgs boson mass

1.5.1 Upper bounds from unitarity

As already emphasized, if there were no Higgs boson, and nothing analogous to replace it, the Standard Model would not be a calculable, renormalizable theory. This incompleteness is reflected in the behaviours of physical quantities as the Higgs mass increases. The most basic example of this is W^+W^-

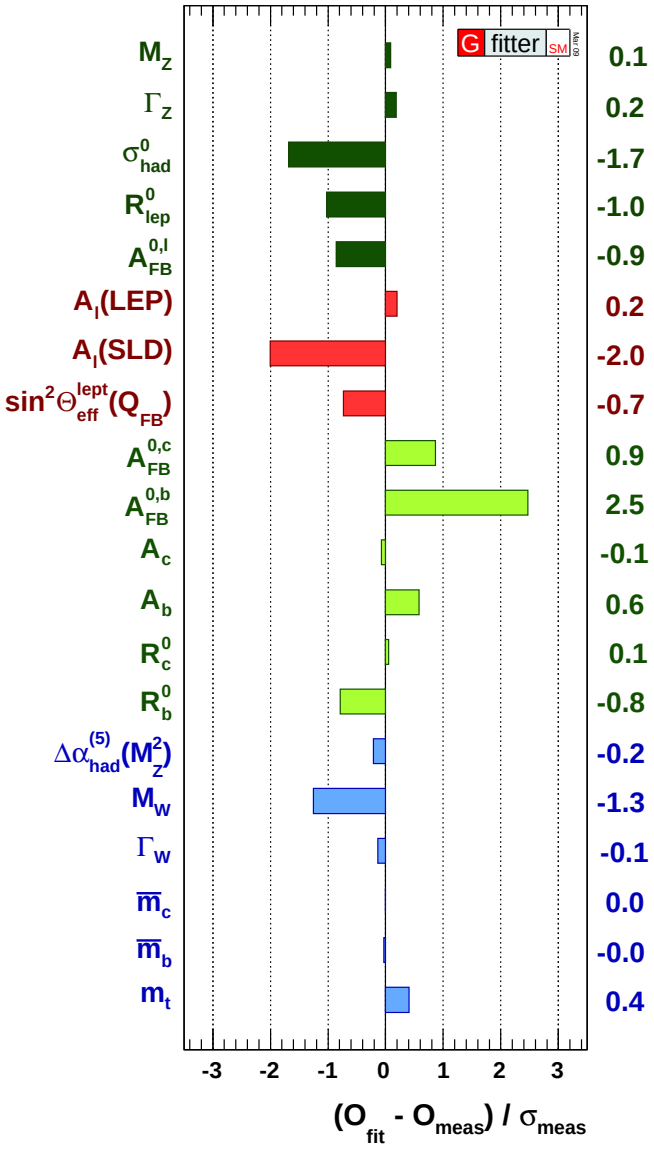


Fig. 3: Comparison between direct measurements and the results of a fit using the `Gfitter` package [17]

scattering [23], whose high-energy s-wave amplitude grows with m_H:

$$T \sim -\frac{4G_F}{\sqrt{2}} m_H^2. \tag{82}$$

Imposing the unitarity bound $|T| < 1$, one finds the upper limit $M_H^2 < 4\pi\sqrt{2}/G_F$, which is strengthened to

$$M_H^2 < \frac{8\pi\sqrt{2}}{3G_F} \sim 1 \text{ TeV}^2 \tag{83}$$

when one makes a coupled analysis including the $Z^0 Z^0$ channel.

A related effect is seen in the behaviour of the quartic self-coupling λ of the Higgs field. Like any of the Standard Model parameters, λ is subject to renormalization *via* loop corrections. Loops of fermions, most importantly the top quark, tend to *decrease* λ as the renormalization scale Λ increases,

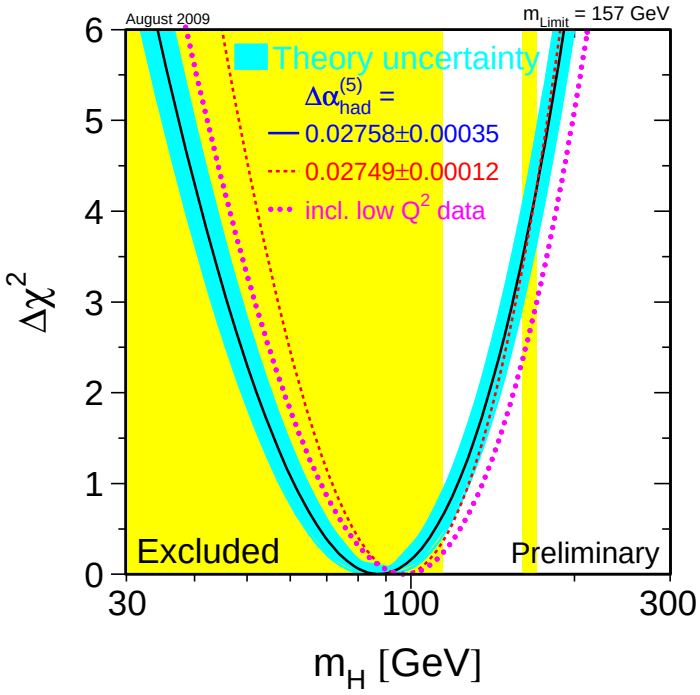


Fig. 4: The χ^2 likelihood function for m_H in a global electroweak fit. The blue band around the (almost) parabolic solid curve represents the theoretical uncertainty: the other curves indicate the effects of different calculations of the renormalization of α_{em} and of including low-energy data. The shaded regions are those excluded by LEP and by the Tevatron [16].

whereas loops of bosons tend to *increase* λ. In particular, if the Higgs mass $\gtrsim m_t$, the positive renormalization due to the Higgs self-coupling itself is dominant, and λ increases uncontrollably with Λ. The larger the value of m_H, the larger the low-energy value of λ, and the smaller the value of Λ at which λ blows up. In general, one should regard the limiting value of Λ, also for smaller m_H, as a scale where novel non-perturbative dynamics must set in. This behaviour is seen in the upper part of Fig. 5, where we see, for example, that if $m_H = 170$ GeV, then $\Lambda \sim 10^{19}$ GeV, whereas if $m_H = 300$ GeV, the coupling λ blows up at a scale $\Lambda \sim 10^6$ GeV. One may ask: under what circumstances does $m_H \sim \Lambda$ itself? The answer is when $m_H \sim 700$ GeV: if the Higgs boson were heavier than this mass, the Higgs self-coupling would blow up at a scale smaller than its mass. In this case, Higgs physics would necessarily be described by some new strongly-interacting theory, cf., the technicolour theories described in Lecture 2.

1.5.2 Lower bounds from vacuum stability

Looking at lower values of m_H in Fig. 5, we see an uneventful range of m_H extending down to $m_H \sim 130$ GeV, where (as far as we know) the SM could continue to be valid all the way to the Planck scale. At lower m_H, there is a band below which the present electroweak vacuum becomes unstable at some scale $\Lambda < 10^{19}$ GeV. For example, if the Higgs is slightly above the present experimental lower limit from LEP, $m_H \sim 115$ GeV, the present electroweak vacuum is unstable against decay into a vacuum with $\langle|\phi|\rangle \sim 10^7$ GeV. This instability is due to the negative renormalization of λ by the top quark, which overcomes the positive renormalization due to λ itself, and drives $\lambda < 0$ [5].

If m_H is only slightly below the top band, and above the middle band, it is true that the present

[5] The widths of the boundary bands indicate the uncertainties in these calculations.

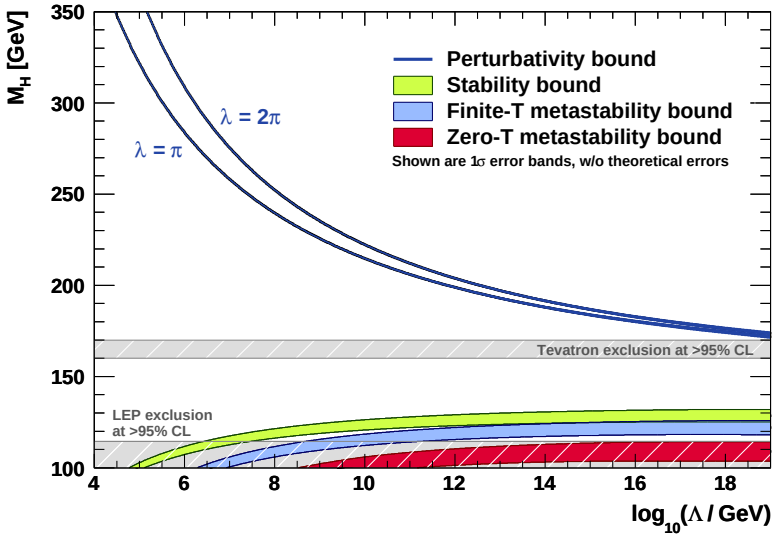


Fig. 5: The scale Λ at which the two-loop RGEs drive the quartic SM Higgs coupling non-perturbative (upper curves), and the scale Λ at which the RGEs create an instability in the electroweak vacuum (lower curves). The widths of the bands reflect the uncertainties in m_t and $\alpha_s(m_Z)$ (added quadratically). The perturbativity upper bound (sometimes referred to as 'triviality' bound) is given for $\lambda = \pi$ (lower bold line [blue]) and $\lambda = 2\pi$ (upper bold line [blue]). Their difference indicates the theoretical uncertainty in this bound. The absolute vacuum stability bound is displayed by the light shaded [green] band, while the less restrictive finite-temperature and zero-temperature metastability bounds are medium [blue] and dark shaded [red], respectively. The grey hatched areas indicate the LEP [20] and Tevatron [24] exclusion domains. Figure taken from [25].

electroweak vacuum is in principle unstable against decay into a state with $\langle|\phi|\rangle > \Lambda$, but it would not have decayed during the conventional thermal expansion of the Universe at finite temperatures. Below the middle band but above the lowest band, the vacuum would have decayed to a correspondingly large value of $\langle|\phi|\rangle$ at some finite temperature, but its present-day (low-temperature) lifetime is longer than the age of the Universe. Below the lowest band, the lifetime for decay to a vacuum with $\langle|\phi|\rangle > \Lambda$ would be less than the present age of the Universe at low temperatures, and we should really watch out!

In fact, as we see shortly, such low values of m_H are almost excluded by LEP searches for the SM Higgs boson, as also seen in Fig. 5.

One could in principle avoid this vacuum instability by introducing some new physics at an energy scale $< \Lambda$: what type of physics [26]? One needs to overcome the negative effects of renormalization of λ by loops with the top quark circulating. The sign of renormalization could be reversed by loops with some boson circulating, potentially restoring the stability of the electroweak vacuum. However, then one should consider the renormalization of the quartic coupling between the Higgs and the new boson. It turns out that the renormalization of this coupling is in turn very unstable, and that the best way to stabilize this coupling would be to introduce a new fermion.

These new scalars and fermions look very much like the partners of the top quark and Higgs bosons, respectively, that are predicted by supersymmetry [26]. In Lecture 3 we will study in more detail the renormalization of mass and vacuum parameters in a supersymmetric theory.

1.5.3 Results from searches at LEP and the Tevatron

As seen in Fig. 2, searches for the reaction $e^+e^- \to Z^0 + H$ at LEP established a lower limit on the possible mass of a SM Higgs boson [20]:

$$m_H \; > \; 114.4 \text{ GeV} \tag{84}$$

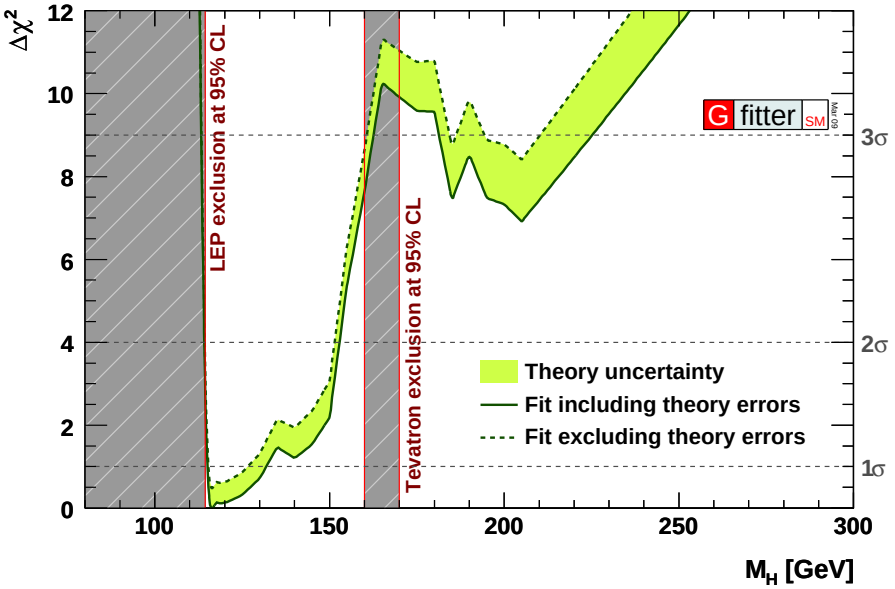


Fig. 6: Dependence on M_H of the $\Delta\chi^2$ function obtained from the global fit of the SM parameters to precision electroweak data [25], excluding (left) or including (right) the results from direct searches at LEP and the Tevatron

at the 95% confidence level. The lower limit (84) is somewhat higher than the central value of the SM Higgs mass preferred by the global precision electroweak fit (81), but there is no significant tension between these two pieces of information. Figure 6 shows the χ^2 likelihood function obtained by combining the LEP search and the global electroweak fit. At the 95% confidence level, one finds [20]

$$m_H \; < 157 \text{ GeV}, \; 186 \text{ GeV}, \tag{85}$$

depending whether one uses precision electroweak data alone, or includes also the lower limit (84) from the direct search at LEP. The χ^2 function obtained by combining the LEP limit (84) with the precision electroweak fit is shown in Fig. 6. Notice the little blip at $m_H \sim 115$ GeV, reflecting the hint of a signal found in the last run at the highest LEP energies: this was only at the 1.7-σ level, insufficient to claim any evidence.

Searches at the Fermilab Tevatron collider have recently started to exclude a region of mass for the SM Higgs boson, as also seen in Figs. 2, 5 and 6. At the time of writing, these searches exclude [24]

$$163 \text{ GeV} \; < \; m_H \; < \; 166 \text{ GeV} \tag{86}$$

at the 95% confidence level, as seen in Fig. 7. At smaller masses, the Tevatron 95% confidence level upper limit on Higgs production and decay is only a few times bigger than the SM expectations, and the integrated luminosity is expected to double over the next couple of years.

Figure 6 also includes the effect on the χ^2 likelihood function of combining the Tevatron search with the global electroweak fit and the LEP search. We see from this that the 'blow-up' region $m_H > 180$ GeV is strongly disfavoured: above the 99% confidence level if the Tevatron data are included, compared with 96% if they are dropped [25]. The combination of all the data yields a 68% confidence level range [17]

$$m_H \; = \; 116^{+16}_{-1.3} \text{ GeV}. \tag{87}$$

The Tevatron is expected to continue running until late 2011, accumulating $\mathcal{O}(10)$/fb of integrated luminosity. That could be sufficient to exclude a SM Higgs boson over all the mass range between (84) and (86), which would exclude all the preferred range (85) — a very intriguing possibility! Alternatively, perhaps the Tevatron will find some evidence for a Higgs boson with a mass within this range?

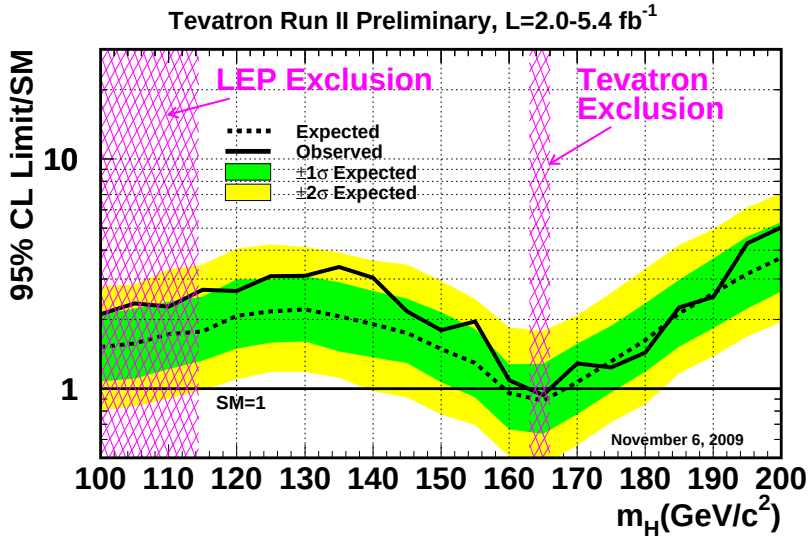


Fig. 7: Combined 95% confidence level upper limit from searches by CDF and D0 for the Higgs boson at the Tevatron collider [24], compared with the SM expectation

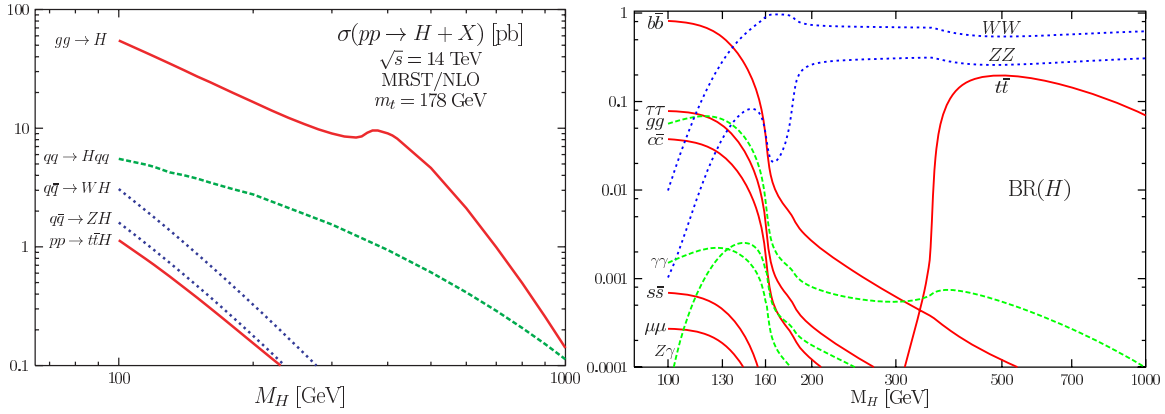


Fig. 8: Left: the dominant mechanisms for producing a SM Higgs boson at the LHC at 14 TeV, and right: the most important branching ratios for a SM Higgs boson, taken from [27]

1.5.4 LHC prospects

The search for the Higgs boson is one of the main raisons d'être of the LHC. Many mechanisms may make important contributions to SM Higgs production at the LHC. If the Higgs boson is relatively light, as suggested above, the dominant production mechanisms are expected to be $gg \to H$ and $W^+W^- \to H$, where the $W^\pm$ are radiated off incoming quarks: $q \to Wq'$.

As already mentioned, the fact that Higgs couplings to other particles are proportional to their masses implies that the Higgs prefers to decay into the heaviest particles that are kinematically accessible. As seen in Fig. 8, this means that a Higgs lighter than $\sim$ 130 GeV prefers to decay into $\bar{b}b$, whereas a heavier Higgs prefers to decay into W^+W^- and Z^0Z^0. However, couplings to lighter particles can become important under certain circumstances. For example, whilst there is no tree-level coupling to gluons because they are massless, one is induced by loops of heavy particles such as the top quark. For

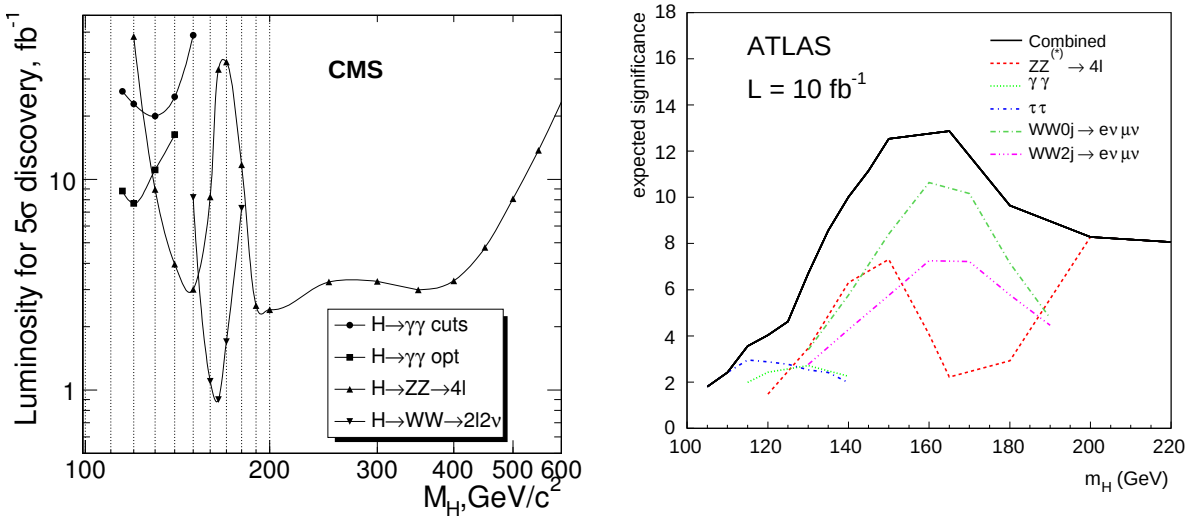


Fig. 9: Left: the amount of integrated luminosity that would be required by CMS [28] to discover a SM Higgs boson as a function of m_H. Right: the significance expected by ATLAS [29] for a SM Higgs boson, assuming 10/fb of data at 14 TeV.

the same reason, there is no tree-level Higgs coupling to photons, but the Higgs boson may decay into $\gamma\gamma$ *via* top and $W^{\pm}$ loops. Although this decay has a very small branching ratio, it is very distinctive experimentally, and may be detectable at the LHC if the SM Higgs weighs < 130 GeV.

Figure 9 displays estimates of the sensitivities of CMS (left) [28] and ATLAS (right) [29] to a SM Higgs boson. A fraction of an inverse femtobarn per experiment may suffice to exclude a Higgs boson over a large range of masses from ~ 150 GeV to ~ 400 GeV. An integrated luminosity ~ 1/fb per experiment would be needed to discover a Higgs boson with a mass in a similar range, but more luminosity would be required if $m_H < 150$ GeV. Indeed, a luminosity ~ 5/fb per experiment would be needed for discovery over all the displayed range of m_H, down to the LEP limit. One way or another, the LHC will be able determine whether or not there is a SM Higgs boson.

1.6 Issues beyond the Standard Model

The Standard Model, however, is not expected to be the final description of the fundamental interactions, but rather an effective low-energy (up to a few TeV) manifestation of a more complete theory.

Some of the outstanding questions in the Standard Model are:

- **How is electroweak symmetry broken?** In other words, how do gauge bosons acquire mass? We have seen that the Standard Model incorporates the Higgs mechanism in the form of a single weak-isospin doublet with a non-zero v.e.v. in order to generate the gauge boson masses, but this is not the only possible way in which the electroweak symmetry can be broken. For instance, there could be more than one Higgs doublet, the Higgs could be a pseudo-Goldstone boson (with a low mass relative to the mass scale of some new interaction) or electroweak symmetry could be broken by a condensate of new particles bound by a new strong interaction. We cover a few of the possibilities in Lecture 2.
- **How do fermions acquire mass?** Electroweak symmetry breaking is a necessary, but not a sufficient, condition to generate the fermion masses. There also needs to be a mechanism that generates the required Yukawa couplings [see Eq. (46)] between the fermions and the (effective) Higgs field. The separation between electroweak symmetry breaking and the generation of fermion masses is made evident in models of dynamical symmetry breaking, such as technicolour (see Section 2),

where the breaking is carried out by the formation of a condensate of particles associated to a new interaction, a process which, while breaking electroweak symmetry and giving masses to the gauge bosons, does not necessarily give masses to the fermions. This situation is resolved by adding new interactions which are responsible for generating the fermion masses. Within the Standard Model, there are no predictions for the values of the Yukawa couplings. Moreover, the values required to generate the correct masses for the three charged leptons and the six quarks span six orders of magnitude, which presumably makes the mechanism for the generation of the couplings highly non-trivial.

- **The hierarchy problem.** Why should the Higgs mass remain low, $m_H \lesssim 1$ TeV, in the face of divergent quantum loop corrections? Following [3], the Higgs mass can be expanded in perturbation theory as

$$m_H^2\left(p^2\right) = m_{0,H}^2 + \mathcal{C}g^2 \int_{p^2}^{\Lambda^2} dk^2 + \dots \,, \tag{88}$$

where $m_{0,H}^2$ is the tree-level (classical) contribution to the Higgs mass squared, g is the coupling constant of the the theory, $\mathcal{C}$ is a model-dependent constant, and Λ is the reference scale up to which the Standard Model is assumed to remain valid. The integrals represent contributions at loop level and are apparently quadratically divergent. If there is no new physics, the reference scale is high, like the Planck scale, $\Lambda \sim M_{\rm Pl} \approx 10^{19}$ GeV or, in Grand Unified Theories (GUTs), $\Lambda \sim M_{GUT} \approx 10^{15} - 10^{16}$ GeV (see Lecture 4). Clearly, both choices result in large corrections to the Higgs mass. In order for these to be small, there are two alternatives: either the relative magnitudes of the tree-level and loop contributions are finely tuned to yield a net contribution that is small (a feature that is disliked by physicists, but which Nature might have implemented), or there is a new symmetry, like supersymmetry, that protects the Higgs mass, as discussed in Lecture 3.

- **The vacuum energy problem**. The value of the scalar potential, Eq. (31), at the v.e.v. $\langle\phi\rangle_0$ of the Higgs boson is

$$V\left(\langle\phi^\dagger\phi\rangle_0\right) = \frac{\mu^2 v^2}{4} < 0 \,. \tag{89}$$

Hence, because the Higgs mass is $m_H^2 = -2\mu^2$, this corresponds to a uniform vacuum energy density

$$\rho_H = -\frac{m_H^2 v^2}{8} \,. \tag{90}$$

Taking $v = \left(G_F\sqrt{2}\right)^{-1/2} \approx 246$ GeV for the Higgs v.e.v. and using the current experimental lower bound on the Higgs mass [13], $m_H \gtrsim 114.4$ GeV, we have

$$-\rho_H \gtrsim 10^8 \text{ GeV}^4 \,. \tag{91}$$

On the other hand, if the apparent accelerated expansion of the Universe — originally inferred from observations of type 1A supernovae [30] — is attributed to a non-zero cosmological constant corresponding to $\sim 70\%$ of the total energy density of the Universe [13], the required energy density should be

$$\rho_{\rm vac} \sim 10^{-46} \text{ GeV}^4 \,, \tag{92}$$

which is at least 54 orders of magnitude lower than the corresponding density from the Higgs field, and of the opposite sign! The character of this dark energy remains unexplained [31, 32], and will probably remain so until we have a full quantum theory of gravity.

- **How is flavour symmetry broken?** Part of the flavour problem in the Standard Model is, of course, related to the widely different mass assignments of the fermions ascribed to the Yukawa

couplings, which also set the mixing angles between flavour and mass eigenstates. Mixing occurs both in the quark and the lepton sectors, the former being parametrized by the Cabibbo–Kobayashi–Maskawa (CKM) matrix and the latter, by the Maki–Nakagawa–Sakata (MNS) matrix. These are complex rotation matrices, and can each be written in terms of three mixing angles and one CP-violating phase (δ) [13]:

$$V = \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta} & c_{23}c_{13} \end{pmatrix} , \tag{93}$$

where $c_{ij} \equiv \cos(\theta_{ij})$, $s_{ij} \equiv \sin(\theta_{ij})$. While the off-diagonal elements in the quark sector are rather small (of order 10^{-1} to 10^{-3}), so that there is little mixing between quark families, in the lepton sector the off-diagonal elements (except for $[V_{\rm MNS}]_{e3}$, which is close to zero) are of order 1, so that the mixing between neutrino families is large. The Standard Model does not provide an explanation for this difference.

- **What is dark matter?** The observation that galaxy rotation curves do not fall off with radial distance from the galactic centre can be explained by postulating the existence of a new type of weakly-interacting matter, *dark matter*, in the halos of galaxies. Supporting evidence from the cosmic microwave background (CMB) indicates that the dark matter makes up $\sim 25\%$ of the energy density of the Universe [33]. Dark matter is usually thought to be composed of neutral relic particles from the early Universe. Within the Standard Model, neutrinos are the only candidate massive neutral relics. However, they contribute only with a normalized density of $\Omega_\nu \gtrsim 1.2\,(2.2) \times 10^{-3}$ if the mass hierarchy is normal (inverted), or no more than 10% if the lightest mass eigenstate lies around 1 eV, that is, if the hierarchy is degenerate [3]. On top of that, structure formation indicates that dark matter should be cold, i.e., non-relativistic at the time of structure formation, whereas neutrinos would have been relativistic particles. Within the Minimal Supersymmetric extension of the Standard Model (MSSM), the lightest supersymmetric partner, called a *neutralino*, is a popular dark matter candidate [34].
- **How did the baryon asymmetry of the Universe arise?** The antibaryon density of the Universe is negligible, whilst the baryon-to-photon ratio has been determined, using WMAP data [6] of the CMB [35] to be

$$\eta = \frac{n_b - \overline{n}_b}{n_\gamma} \simeq \frac{n_b}{n_\gamma} = 6.12\,(19) \times 10^{-10} , \tag{94}$$

where n_b, $\overline{n}_b$, and n_γ are the number densities of baryons, antibaryons, and photons, respectively. The fact that the ratio is not zero is intriguing considering that, in a cosmology with an inflationary epoch, conventional thermal equilibrium processes would have yielded an equal number of particles and antiparticles. In 1967, Sakharov [36] established three necessary conditions (more fully explained in [37]) for the particle–antiparticle asymmetry of the Universe to be generated:

1. violation of the baryon number, B;
2. microscopic C and CP violation;
3. loss of thermal equilibrium.

Otherwise, the rate of creation of baryons equals the rate of destruction, and no net asymmetry results. In the perturbative regime, the Standard Model conserves B; however, at the non-perturbative level, B violation is possible through the triangle anomaly [15]. The loss of thermal equilibrium may occur naturally through the expansion of the Universe, and CP violation enters the Standard Model through the complex phase in the CKM matrix [13]. However, the CP violation observed so far, which is described by the Kobayashi–Maskawa mechanism of the Standard

[6] We use here values from the three-year WMAP analysis [35], rather than the five-year analysis [33], in order to be consistent with the values quoted by the Particle Data Group [13] summary tables.

Model, is known to be insufficient to explain the observed value of the ratio η, and new physics is needed. One possible solution lies in leptogenesis scenarios, where the baryon asymmetry is a result of a previously existing lepton asymmetry generated by the decays of heavy sterile neutrinos [38].

- **Quantization of the electric charge**. It is an experimental fact that the charges of all observed particles are simple multiples of a fundamental charge, which we can take to be the electron charge, e. Dirac [39–41] proved that the existence of even a single magnetic monopole (a magnet with only one pole) is sufficient to explain the quantization of the electric charge, but the particle content of the Standard Model (see Table 1) does not include magnetic monopoles. Hence, in the absence of any indication for a magnetic monopole, the explanation of charge quantization must lie beyond the Standard Model. Indeed, so far there has only been one candidate monopole detection event in a single superconducting loop [42], in 1982, and the monopole interpretation of the event has now been largely discounted. One expects monopoles to be very massive and non-relativistic at present, in which case time-of-flight measurements in the low-velocity regime ($\beta \equiv v/c \ll 1$) become important. The best current direct upper limit on the supermassive monopole flux comes from cosmic-ray observations [13]:

$$\Phi_{\text{1pole}} < 1.0 \times 10^{-15}\ \text{cm}^{-2}\text{sr}^{-1}\text{s}^{-1}\ , \tag{95}$$

  for $1.1 \times 10^{-4} < \beta < 0.1$. An alternative route towards charge quantization is *via* a Grand Unified Theory (GUT) (see Lecture 4). Such a theory implies the existence of magnetic monopoles that would be so massive that their cosmological density would be suppressed to an unobservably small value by cosmological inflation.

- **How to incorporate gravitation?** One of the most obvious shortcomings of the Standard Model is that it does not incorporate gravitation, which is described on a classical level by general relativity. However, the consistency of our physical theories requires a quantum theory of gravity. The main difficulty in building a quantum field theory of gravity is its non-renormalizability. String theory [43] and loop quantum gravity [44] constitute attempts at building a quantized theory of gravity. If one could answer this question, one would surely also be able to solve the dark energy problem. Conversely, solving the dark energy problem presumably requires a complete quantum theory of gravity.

2 Electroweak symmetry breaking beyond the Standard Model

2.1 Theorists are getting cold feet

After so many years, it seems that we will soon know whether a Higgs boson exists in the way predicted by the Standard Model, or not. Closure at last!

Like the prospect of an imminent hanging, the prospect of imminent Higgs discovery concentrates wonderfully the minds of theorists, and many theorists with cold feet are generating alternative models, as prolifically as monkeys on their laptops. These serve the invaluable purpose of providing benchmarks that can be compared and contrasted with the SM Higgs. Experimentalists should be ready to search for reasonable alternatives, already at the Tevatron and also at the LHC once it is up and running, and they should be on the look-out for tell-tale deviations from the SM predictions if a Higgs boson should appear.

Even within the SM with a single elementary Higgs boson, questions are being asked. As discussed in the previous section, within this framework the experimental data seem to favour a light Higgs boson. However, the interpretation of the precision electroweak data has been challenged. Even if one accepts the data at face value, the SM fit may need to take into account non-renormalizable, higher-dimensional interactions that could conspire to permit a heavier SM Higgs boson? In this section, in addition to these possibilities, we explore several mechanisms of electroweak symmetry breaking be-

yond the minimal Higgs, i.e., a single elementary $SU(2)$ Higgs doublet whose potential is arranged to have a non-zero v.e.v.

Any successful model of electroweak symmetry breaking must give masses to the matter fermions as well as the weak gauge bosons. This could be achieved using either a single boson, as in the SM, or two of them, as in the Minimal Supersymmetric extension of the Standard Model (MSSM) [7], or by some composite of new fermions with new strong interactions that generate a non-zero v.e.v. as in (extended) technicolour models, or by some Higgsless mechanism.

We do know, however, that the energy scale at which EWSB must occur is $\mathcal{O}(1)$ TeV [45]. This scale is set by the decay constant of the three Goldstone bosons that, through the Higgs mechanism, are transformed into the longitudinal components of the weak gauge bosons:

$$F_\pi = \left(G_F\sqrt{2}\right)^{-1/2} \approx 246 \text{ GeV} . \tag{96}$$

If there is any new physics associated to the breaking of electroweak symmetry, it must occur near this energy scale. Another way to see how this energy scale emerges is to consider s-wave WW scattering. In the absence of a direct-channel Higgs pole, this amplitude would violate the unitarity limit at an energy scale ~ 1 TeV (82).

It is the scale of 1 TeV, and the typical values of QCD and electroweak cross sections at this energy, $\sigma \simeq 1$ nb–1 fb, that set the energy and luminosity requirements of the LHC: $\sqrt{s} = 14$ TeV and $\mathcal{L} = 10^{34}$ cm^{-2} s^{-1} for pp collisions [13]. This energy scale is to be contrasted with the energy scale of the other unexplained broken symmetry in the SM, namely flavour symmetry, which is completely unknown: it may lie anywhere from 1 TeV up to the Planck scale, $M_P = 1.22 \times 10^{19}$ GeV.

There are some general constraints that any proposed model of electroweak symmetry breaking must satisfy [46]. First, the model must predict a value of the ρ parameter, Eq. (73), that agrees with the value $\rho \approx 1$ found experimentally. The desired value $\rho = 1$ is found automatically in models that contain only Higgs doublets and singlets, but would be violated in models with scalar fields in larger $SU(2)$ representations. A second constraint comes from the strict upper limits on flavour-changing neutral currents (FCNCs). These are absent at tree level in the minimal Higgs model, a fact that is in general not true in non-minimal models.

2.2 Interpretation of the precision electroweak data

It is notorious that the two most precise measurements at the Z^0 peak, namely the asymmetries measured with leptons (particularly $A_\ell(SLD)$) and hadrons (particularly $A_{FB}^{0,b}$), do not agree very well [47], as seen in Table 2 and Fig. 1.4 [8]. Within the SM, they favour different values of m_H, around 40 and 500 GeV, respectively, as seen in Fig. 10. Most people think that this discrepancy is just a statistical fluctuation, since the total χ^2 of the global electroweak fit is acceptable ($\chi^2 = 17.3$ for 13 d.o.f., corresponding to a probability of 18% [16]), but it may also reflect the existence of an underestimated systematic error. However, if there were a big error in $A_{FB}^{0,b}$, the preferred value of m_H would be pulled uncomfortably low by the other data, whereas if there was a big error in the interpretation of the leptonic data m_H would be pulled towards much higher values. On the other hand, if we take both pieces of data at face value, perhaps the discrepancy is evidence for new physics at the electroweak scale. In this case there would be no firm basis for the prediction of a light Higgs boson, which is based on a Standard Model fit, and no fit value of m_H could be trusted?

[7] We leave the treatment of the Higgs sector within the MSSM for a later section.

[8] Another anomaly is exhibited by the NuTeV data on deep-inelastic $\nu - N$ scattering [48], but this is easier to explain away as due to our lack of understanding of hadronic effects.

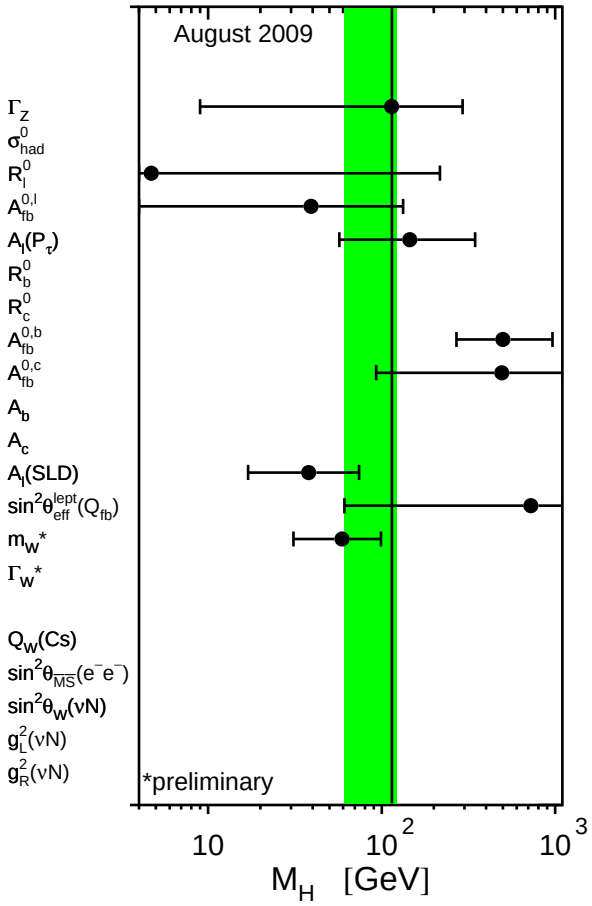


Fig. 10: The 68% confidence level ranges for m_H that are indicated by various individual electroweak measurements [16]

2.3 Higher-dimensional operators within the SM

The Standard Model should be regarded simply as an effective low-energy theory, to be embeded within some more complete and satisfactory theory. Therefore, one should anticipate that the renormalizable dimension-four interactions of the SM could be supplemented by higher-dimensional operators of the general form:

$$\mathcal{L}_{eff} \;=\; \mathcal{L}_{SM} + \Sigma_i \frac{c_i}{\Lambda_i^p} \mathcal{O}_i^{4+p}, \tag{97}$$

where Λ_i is a scale at which the supplementary interaction $\mathcal{O}_i^{4+p}$ of dimension $4+p$ appears to be generated. A global fit to the precision electroweak data suggests that, if the Higgs is indeed light, the coefficients of these additional interactions are small:

$$\Lambda_i \;>\; \mathcal{O}(10)\ \text{TeV} \tag{98}$$

for $c_i = \pm 1$. It is then a problem to understand the 'little hierarchy' between the electroweak scale and Λ_i.

However, conspiracies are in principle possible, which could allow m_H to be large, even if one takes the precision electroweak data at face value [49]. Examples are shown in Fig. 11, where one sees corridors of allowed parameter space extending up to a heavy Higgs mass, if $\Lambda_i \ll 10$ TeV. A theory that predicts a heavy Higgs boson but remains consistent with the precision electroweak data should predict a correlation of the type seen in Fig. 11. At the moment, this may seem unnatural to us, but Nature may

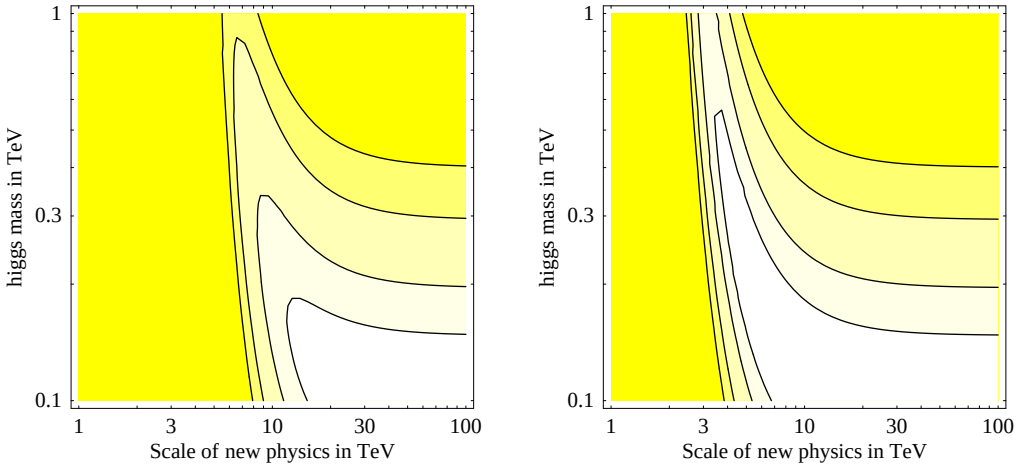


Fig. 11: The 68%, 90%, 99% and 99.9% confidence levels fit for global electroweak fits including two different types of higher-dimensional operators, demosntrating that they might conspire with a relatively heavy Higgs boson to yield and acceptable fit [49]

know better. In any case, any theory beyond the SM must link the value of m_H and the scales of these higher-dimensional effective operators in some way.

2.4 Little Higgs

One way to address the 'little hierarchy problem' and explain the lightness of the Higgs boson (if it is light) is by treating it as a pseudo-Goldstone boson corresponding to a spontaneously broken approximate global symmetry of a new strongly-interacting sector at some higher mass scale, the 'little Higgs' scenario [50]. Such a theory would work by analogy with the pions in QCD, which have masses far below the generic mass scale of the strong interactions ~ 1 GeV.

If the Higgs is a pseudo-Goldstone boson, its mass is protected from acquiring quadratically-divergent loop corrections [51]. This occurs as a result of the particular manner in which the gauge and Yukawa couplings break the global symmetries: more than one couplng must be turned on at a time in order for the symmetry to be broken, a feature known as 'collective symmetry breaking' [52, 53]. As a consequence, the quadratic divergences that would normally appear in the SM are cancelled by new particles, sometimes in unexpected ways. For example, the top-quark loop contribution to the Higgs mass-squared has the general form

$$\delta m^2_{H,top}(SM) \ \sim \ (115GeV)^2 \left(\frac{\Lambda}{400 \text{ GeV}}\right)^2. \tag{99}$$

As illustrated in Fig. 12, in little Higgs models this is cancelled by the loop contribution due to a new heavy top-like quark T with charge +2/3 that is a singlet of $SU(2)_L$, leaving a residual logarithmic divergence:

$$\delta m^2_{H,top}(LH) \ \sim \ \frac{6G_F m_t^2}{\sqrt{2\pi^2}} m_T^2 \log\frac{\Lambda}{m_T}. \tag{100}$$

Analogously, the quadratic loop divergences associated with the gauge bosons and the Higgs boson of the Standard Model are cancelled by loops of new gauge bosons and Higgs bosons in little Higgs models.

The net result is a spectrum containing a relatively light Higgs boson and other new particles that may be somewhat heavier:

$$M_T < 2 \text{ TeV} \left(\frac{m_H}{200 \text{ GeV}}\right)^2, M_{W'} < 6 \text{ TeV} \left(\frac{m_H}{200 \text{ GeV}}\right)^2, M_{H^{++}} < 10 \text{ TeV}. \tag{101}$$

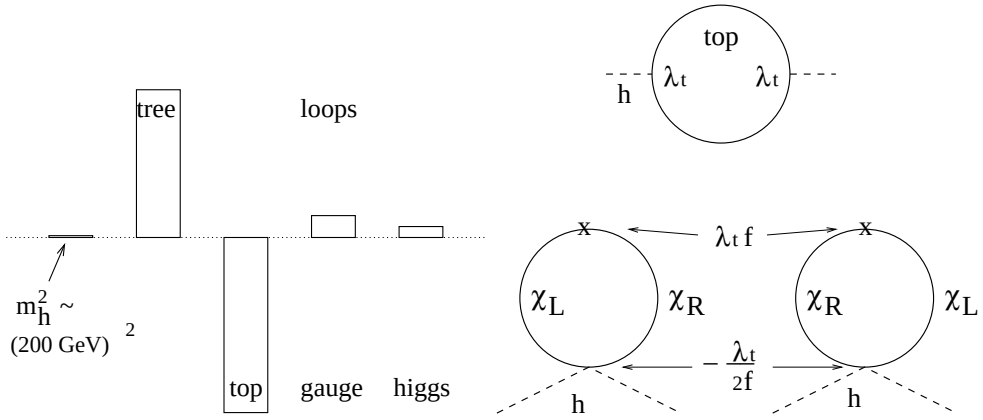


Fig. 12: Left: If the Standard Model Higgs boson weighs around 200 GeV, the top-quark loop contribution to its physical mass (calculated here with a loop momentum cutoff of 10 TeV) must cancel delicately against the tree-level contribution. Right: In 'little Higgs' models, the top-quark loop is cancelled by loops containing a heavier charge-2/3 quark [50].

The extra T quark, in particular, should be accessible to the LHC. In addition, there should be more new strongly-interacting physics at some energy scale at or above 10 TeV, to provide the ultra-violet completion of the theory.

2.5 Technicolour

Little Higgs models are particular examples of composite Higgs models, of which the prototypes were technicolour models [54, 55]. In these models, electroweak symmetry is broken dynamically, by the introduction of a new non-Abelian gauge interaction [56–58] that becomes strong at the TeV scale. The building blocks are massless fermions called technifermions and new force-carrying fields called technigluons. As in the SM, the left-handed components of the technifermions are assigned to electroweak doublets, while the right-handed components form electroweak singlets, and both components carry hypercharge. At $\Lambda_{\rm EW} \sim 1$ TeV the technicolour coupling becomes strong, which leads to the formation of condensates of technifermions with v.e.v.'s

$$\langle \phi \rangle = \langle \overline{f}_L f_R \rangle \equiv v \; . \tag{102}$$

Because the left-handed technifermions carry electroweak quantum numbers, but the right-handed ones do not, the formation of this technicondensate breaks electroweak symmetry.

The massless technifermions have the chiral symmetry group

$$G_\chi = SU(2N_D)_L \otimes SU(2N_D)_R \supset SU(2)_L \otimes SU(2)_R \; , \tag{103}$$

where N_D is the number of technifermion doublets. When the condensate forms, this large global symmetry is broken down to

$$S_\chi = SU(2N_D) \supset SU(2)_V \; , \tag{104}$$

where V refers to the vector combination of left and right currents, and $4N_D^2 - 1$ massless Goldstone bosons appear, with decay constant $F_\pi^{\rm TC}$. Similarly to the Higgs mechanism in the SM, three of these bosons are 'eaten' and become the longitudinal components of the $W^\pm$ and Z^0 weak bosons, which acquire masses [45]

$$m_W = \frac{g}{2}\sqrt{N_D} F_\pi^{\rm TC} \quad , \quad m_Z = \frac{1}{2}\sqrt{g^2 + g'^{\,2}}\sqrt{N_D} F_\pi^{\rm TC} = \frac{m_W}{\cos(\theta_W)} \; . \tag{105}$$

The scale $\Lambda_{\rm TC}$ at which technicolour interactions become strong is related to the magnitude of electroweak symmetry breaking, namely to the weak scale, by:

$$\Lambda_{\rm TC} = \text{few} \times F_\pi^{\rm TC}, F_\pi^{\rm TC} = F_\pi/\sqrt{N_D}\,, \tag{106}$$

where $F_\pi = v \approx 246$ GeV. The breaking of the chiral symmetry in technicolour is reminiscent of chiral symmetry in QCD, which provides a working precedent for the model [9]. Technicolour guarantees $\rho = m_W^2/\left(m_Z^2 \cos\left(\theta_W\right)\right) = 1 + \mathcal{O}\left(\alpha\right)$ through a custodial $SU(2)_R$ flavour symmetry in G_χ [45], which is traceable to the quantum numbers assigned to the technifermions.

Dynamical symmetry breaking addresses the problem of quadratic divergences in the Higgs mass-squared, such as (99), by introducing a composite Higgs boson that 'dissolves' at the scale $\Lambda_{\rm TC}$. In this way, it makes loop corrections to the electroweak scale 'naturally' small. Moreover, technicolour has a plausible mechanism for stabilizing the weak scale far below the Planck scale. The idea is that technicolour, being an asymptotically-free theory, couples weakly at very high energies $\sim 10^{16}$ GeV, and then evolves to become strong at lower energies ~ 1 TeV [54]. However, writing down an explicit GUT scenario based on this scenario has proved elusive.

As described above, the simplest technicolour models could provide masses for the gauge bosons $W^\pm$ and Z^0, but not to the matter fermions. Additions to technicolour could allow for quark and lepton masses by introducing new interaction with technifermions, as in 'extended technicolour' models [55, 60]. However, these had severe problems with flavour-changing neutral interactions [61] and a proliferation of relatively light pseudo-Goldstone bosons that have not been seen by experiment [62].

Moreover, a generic problem with technicolour models is presented by the global electroweak fit discussed in the first Lecture. The preference within the SM for a relatively light Higgs boson (81) may be translated into constraints on the possible vacuum polarization effects due to generic new physics models. QCD-like technicolour models have many strongly-interacting dynamical scalar resonances in the TeV range, e.g., a scalar analogous to the σ meson of QCD that corresponds naively to a relatively heavy Higgs boson, which is disfavoured by the data [63]. Such a model can be reconciled with the electroweak data only if some other effect is postulated to cancel the effects of its large mass. One strategy for evading this problem is offered by 'walking technicolour' theories [64], where the coupling strength evolves slowly, i.e., walks. However, the loss of the close analogy with QCD makes it more difficult to calculate so reliably in such models: lattice techniques may come to the rescue here.

2.6 Interpolating models

So far, we have examined two extreme scenarios: the orthodox interpretation of the SM in which the Higgs is elementary and relatively light, and hence interacts only weakly, and strongly-coupled models exemplified by technicolour. The weakly-coupled scenario would require additional TeV-scale particles to stabilize the Higgs mass by cancelling out the quadratic divergences such as (99). A prototype for such models is provided by supersymmetry, as discussed in the next Lecture. On the other hand, strongly-coupled models such as technicolour introduce many resonances that are required by unitarity and generate important contributions to the oblique radiative corrections, e.g., a vector resonance ρ in W^+W^- scattering would induce

$$\delta\rho \sim \frac{m_W^2}{m_\rho^2} \tag{107}$$

where ρ was defined in (73), and the experimental upper limit $|\rho| < 10^{-3}$ at the 95% confidence level imposes $m_\rho > 2.5$ TeV.

One way to interpolate between these two extreme scenarios, and provide a basis for determining how far from the light-SM-Higgs scenario the data permit us to go, is to consider models in which

[9] The condensation phenomenon also occurs in solid-state physics: dynamical symmetry breaking in superconductors is achieved by the formation of Cooper pairs [59], which are condensates of electron pairs with charge $-2e$.

the unitarization of the W^+W^- scattering amplitude is shared between a light Higgs boson with modified couplings and a vector resonance with mass m_ρ and coupling g_ρ, whose relative importance is parametrized by the combination

$$\xi \equiv v\frac{g_\rho}{m_\rho} . \tag{108}$$

The SM is recovered in the limit $\xi \to 0$, but its decay branching ratios may differ considerably as ξ increases towards the strong-coupling limit $\xi = 1$, as seen in Fig. 13. Thus, one signature for such models at the LHC may be the observation of a Higgs boson with couplings that differ from those of the SM.

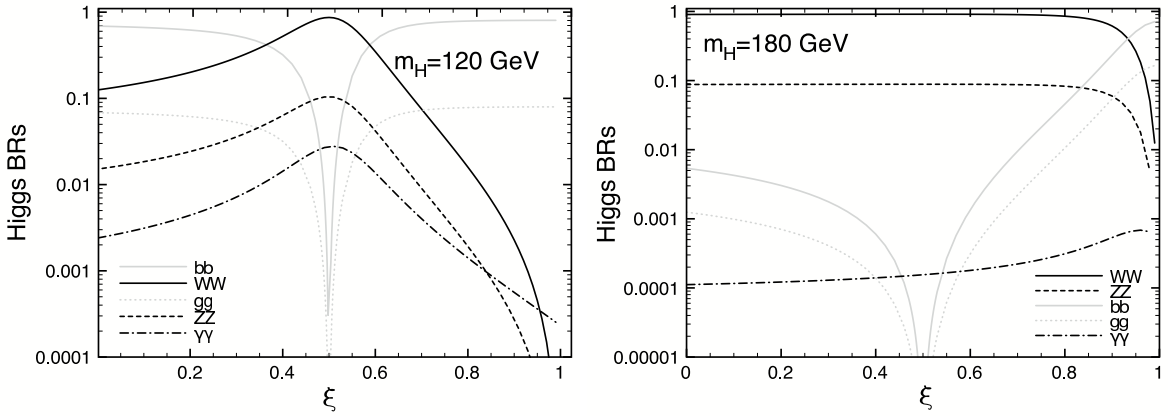


Fig. 13: The dependences of Higgs branching ratios on the parameter ξ (108), for $m_H = 120$ GeV (left) and 180 GeV (right) [65]

Another way to probe such models is to look for effects in $W_L^+W_L^+$ scattering. Unfortunately, at the LHC the $W^\pm$ bosons that are flashed off from incoming energetic quarks: $q \to Wq'$ have predominantly transverse polarizations, so that $\sigma(W_T^+W_T^+ \to W_T^+W_T^+) \gg \sigma(W_L^+W_T^+ \to W_T^+LW_T^+)$ and $\sigma(W_L^+W_L^+ \to W_L^+W_L^+)$ for all $m_{W^+W^+}$ in the SM, and there is an accidental very small factor [65]:

$$\frac{d\sigma^{LL}/dt}{d\sigma^{TT}/dt} = \frac{1}{2304}\left(\frac{m_{W^+W^+}}{m_W}\right)^4 \xi^2 , \tag{109}$$

which implies that, even for $\xi = 1$, $\sigma(W_L^+W_L^+ \to W_L^+W_L^+) > \sigma(W_T^+W_T^+ \to W_T^+W_T^+)$ only for $m_{W^+W^+} > 1.2$ TeV, which is unlikely to be accessible at the LHC, as seen in Fig. 14. An alternative possibility for the LHC may be double-Higgs production *via* the reaction $W^+W^- \to HH$, which may be greatly enhanced as compared with its rate in the SM, as also seen in Fig. 14 — though its observability may be a different matter.

2.7 Higgsless models and extra dimensions

As has already been discussed, if there is nothing like a SM Higgs boson, s-wave WW scattering reaches the unitarity limit at $m_{W^+W^-} \sim 1$ TeV (83). An immediate reaction might be: Who cares? Some non-perturbative strong dynamics will necessarily restore unitarity, even in the absence of a Higgs boson. However, more detailed study in specific models has shown that this strong dynamics is apparently incompatible with the precision data: one needs some perturbative mechanism to break the electroweak symmetry.

How can one break a gauge symmetry? Breaking it explicitly would destroy the renormalizability (calculability) of the gauge theory, whereas breaking the symmetry spontaneously by the v.e.v. of some field everywhere in space does retain the renormalizability (calculability) of the gauge symmetry. But

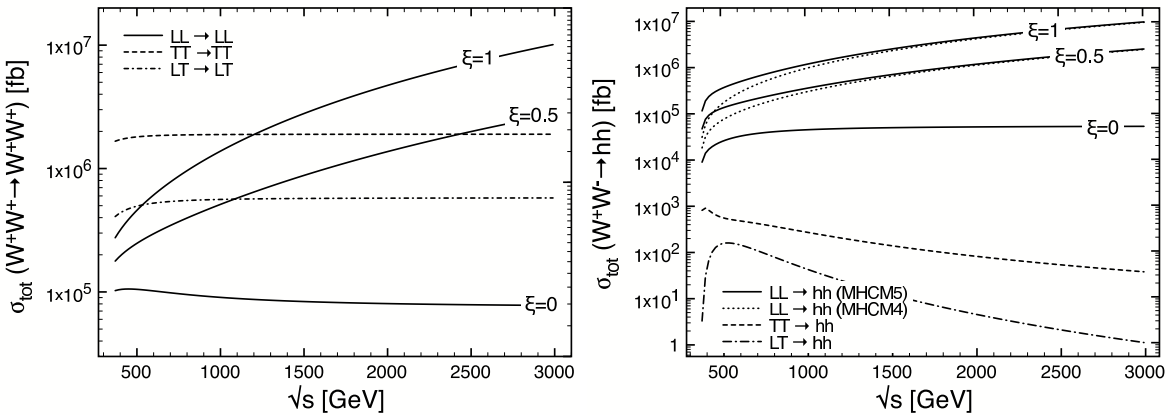


Fig. 14: Left; the cross sections $\sigma(W_T^+W_T^+ \to W_T^+W_T^+)$, $\sigma(W_L^+W_T^+ \to W_T^+LW_T^+)$, and $\sigma(W_L^+W_L^+ \to W_L^+W_L^+)$, as functions of ξ (108). Right: cross sections for double Higgs production [65].

that is the Higgs approach that we are trying to escape: Is there another way? The alternative is to break the electroweak symmetry *via* boundary conditions. This is impossible in conventional $3+1$-dimensional space-time, because it has no boundaries. However, it becomes an option if we postulate finite-size (small) extra space dimensions [66–68].

To see how this works, let us first consider the particle spectrum in the simplest possible model with one extra dimension compactified on a circle S^1 of radius R with internal coordinate (fifth dimension) y, as illustrated in Fig. 15. In this case, the wave function of a boson ϕ at y and $y + 2\pi R$ must be identified:

$$\phi(y + 2\pi R) \; = \; \phi(y) \; , \tag{110}$$

so that one can expand the five-dimensional field as follows:

$$\phi(x, y) \; = \; \sum_n \frac{1}{\sqrt{2^{\delta_{n0}}\pi R}} \left(\cos\left(\frac{ny}{R}\right) \phi_n^+(x) + \sin\left(\frac{ny}{R}\right) \phi_n^-(x) \right) . \tag{111}$$

The $\phi_n^\pm$ are the four-dimensional Kaluza–Klein [69, 70] modes of the field, which appear in four dimensions as particles with masses

$$m_n \; = \; p_y^n \; = \; \frac{n}{R} \, , \tag{112}$$

and the functions $\cos, \sin(ny/R)$ describe the localizations of these modes along the extra dimension. the lowest-lying mode has a flat wave function ($n = 0$), and the excitations have $n > 0$.

We now consider what happens if we 'fold' the circle by identifying $y \sim -y$. Mathematically, this is the simplest *orbifold* S^1/Z_2, also illustrated in Fig. 15. At the same time as identifying $y \sim -y$, we can also identify the field ϕ up to a sign:

$$\phi(-y) \; = \; U\phi(y) \; : \; U^2 \; = \; 1. \tag{113}$$

This has the effect of projecting out half the Kaluza–Klein wave functions (111). If we choose $U = +1$, we select the even wave functions $\cos(ny/R)$ and hence the Kaluza–Klein modes $\phi_n^+(x)$ whereas, if we choose $U = -1$, we select the odd wave functions $\sin(ny/R)$ and hence the Kaluza–Klein modes $\phi_n^-(x)$. The 'even' particles include the massless mode with $n = 0$ whereas all the 'odd' particles are massive. The projection U serves to give masses to all the states that are asymmetric.

This mechanism can be extended to break gauge symmetry [66–68]. Let us consider a five-dimensional theory with a gauge field $A_{\mu,5}$, and let us identify it on the orbifold $y \sim -y$ up to a discrete

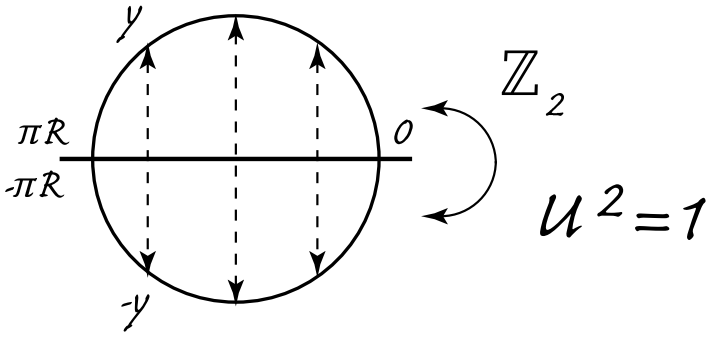


Fig. 15: Compactification on a circle S^1 of radius R with internal coordinate (fifth dimension) y, illustrating the possible orbifolding of this model *via* the identification S^1/Z_2

gauge transformation $U : U^2 = 1$:

$$A_\mu = +UA_\mu(y)U^\dagger, \quad (114)$$
$$A_5 = -UA_5(y)U^\dagger. \quad (115)$$

The gauge symmetry group is broken at the end-points of the orbifold $y = 0, \pi R$: the surviving subgroup is the one that commutes with U, and asymmetric particles acquire masses as described above. In this way, one could imagine breaking $SU(2) \otimes U(1) \to U(1)$ with a suitable orbifold construction.

It is a general feature of this construction that a vector resonance should appear in WZ scattering, corresponding to the lowest-lying Kaluza–Klein excitation. The production of such a particle at the LHC has been considered in the context of a Higgsless model, and could well be observable, as seen in Fig. 16.

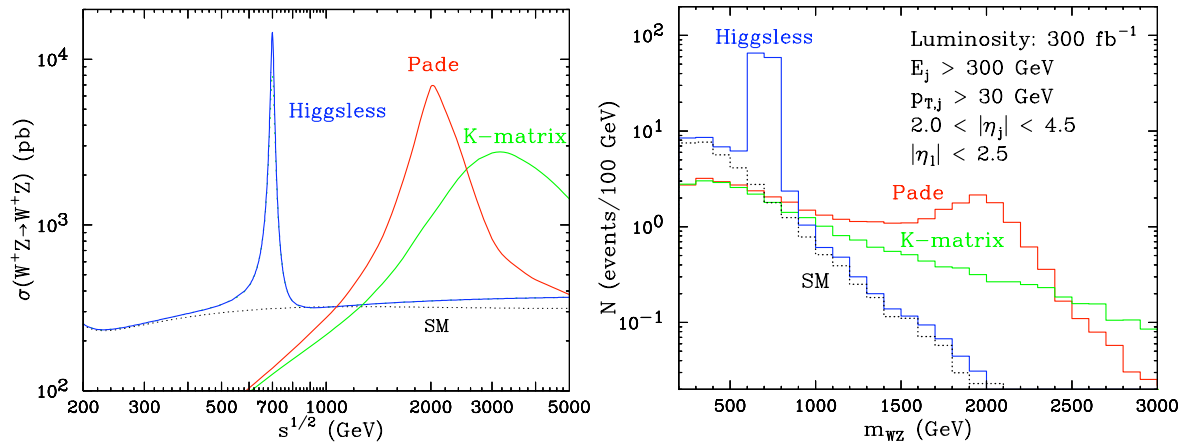


Fig. 16: Left: calculations of the possible modifications of $\sigma(W^+Z^0 \to W^+Z^0)$. Right: simulations of the possible numbers of events at the LHC [65].

You might wonder whether this type of vector resonance bears any relation to the vector resonances discussed previously in the context of new strong dynamics. The answer is yes: as was first emphasized in the context of string theory, a strong coupling is equivalent to a new compactified dimension, and there is in general a 'holographic' relation between four- and five-dimensional theories, the former being considered as boundaries of the five-dimensional 'bulk' theory. These ideas enable the strongly-interacting models of electroweak symmetry breaking discussed in this Lecture, and many others, to be related through a unified description à la M-theory [71], as seen in Fig. 17 [72]. The alternative is a weakly-interacting model of electroweak symmetry breaking, which is favoured, naively, by

the indications from precision electroweak data of a light Higgs boson. In the next Lecture we discuss supersymmetry, which is the most developed such alternative.

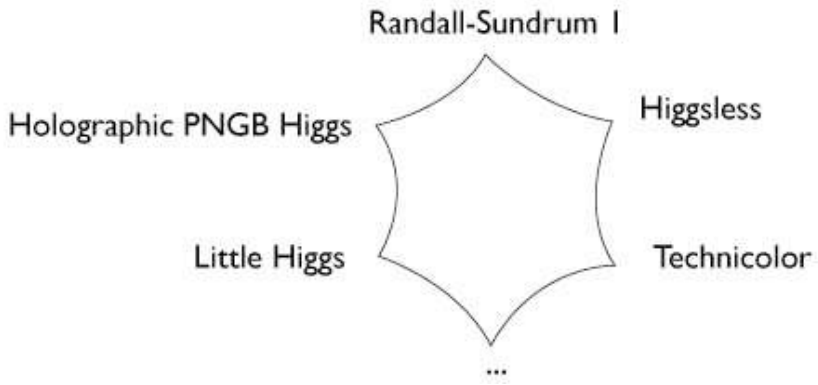


Fig. 17: Relations between different models of electroweak symmetry breaking [72]

3 Supersymmetry

We have seen that the Standard Model is a valid description of physical phenomena at energies lower than a few hundreds of GeV. However, there are various reasons to think that supersymmetry might appear at the TeV scale, and hence play an important role in new discoveries at the LHC, which will explore energies of the order of a TeV. In this Lecture we present and discuss supersymmetric models, with a focus on the phenomenological consequences of supersymmetry.

We first give a brief historical introduction and summarize the motivations for supersymmetry in particle physics. Subsequently we discuss the general formal structure of a physical supersymmetric theory. We then continue with some theoretical notions and applications to 'low-energy' particle physics around the TeV scale. Among the possible models, we focus on the Minimal Supersymmetric Standard Model (MSSM), which provides a basis for analysing supersymmetric phenomenology. Within the context of the MSSM, we discuss the principal experimental constraints on supersymmetry, and then discuss possible aspects of the detection of supersymmetry.

3.1 History and motivations

3.1.1 What is supersymmetry?

Supersymmetry is a radically new type of symmetry that transforms a bosonic state into a fermionic state, or vice versa, with $\Delta S = \pm 1/2$, where S is the spin. Denoting the supersymmetry generator by Q, we may write schematically:

$$Q|Boson\rangle = |Fermion\rangle \quad (116)$$

$$Q|Fermion\rangle = |Boson\rangle. \quad (117)$$

Formally, supersymmetry is an extension of the space-time symmetry reflected in the Poincaré group, and this was a principal motivation leading to its discovery. Initially, it was also hoped that one could use supersymmetry to combine the external space-time symmetries with internal symmetries. However, this prospect seems more distant, as discussed below.

3.1.2 Milestones

There were several attempts in the 1960s to combine internal and external symmetries, but Coleman and Mandula [73] showed in 1967 that it is impossible to combine these types of symmetry, *via* a fa-

mous no-go theorem that is discussed later in more detail. However, their proof assumed that the new symmetry should be generated by bosonic charges of integer spin. In 1971, Golfand and Likhtman [74] discovered an extension of the Poincaré group using fermionic charges of half-integer spin. In the same year, Ramond [75], Neveu and Schwarz [76] proposed supersymmetric models in two dimensions, with the aim of obtaining strings with fermionic states that could accommodate baryons. A few years later, in 1973, Volkov and Akulov [77] tried to apply a nonlinear realization of supersymmetry to neutrinos in four dimensions, but their theory did not describe correctly the low-energy interactions of neutrinos.

In the same year, Wess and Zumino [78, 79] proposed the first four-dimensional supersymmetric field theories of interest from the phenomenological point of view. Specifically, they showed how to construct supersymmetric field theories linking scalars with fermions of spin $1/2$ [78], and also fermions of spin $1/2$ with gauge particles of spin 1 [79]. Then, together with Iliopoulos and Ferrara, Zumino discovered that supersymmetry would eliminate many of the divergences present in other field theories [80, 81]. At first, these ultraviolet properties were regarded as curiosities, in particular because not all logarithmic divergences were eliminated, but attempts were made to construct phenomenological supersymmetric models, for example theories unifying matter particles and Higgs fields in the same supermultiplet. Subsequently, in 1976, two groups [82, 83] found a local version of supersymmetry in which the supersymmetry transformation depends on the space-time coordinates. This theory necessarily includes a description of gravitation, and hence has been called supergravity.

3.1.3 Why supersymmetry?

Following these formal developments, the phenomenology of supersymmetry has been studied intensively, and models based on supersymmetry are considered to be among the most serious candidates for physics beyond the SM [84–86]. Why introduce supersymmetry in particle physics? What makes it so attractive for particle physicists?

The reasons for its introduction in particle physics are principally physical, and quite diverse in nature, as we now discuss.

• The very special properties of supersymmetric field theories are helpful in addressing the naturalness of a (relatively) light Higgs boson. In the previous Lectures we have discussed the existence of enormous radiative corrections to the Higgs mass-squared, m_H^2, which feels the virtual effects of any particle that couples directly or indirectly to the Higgs field. For example, the correction due to a fermionic loop such as that in Fig. 18(a) yields [10]:

$$\Delta m_H^2 = -\frac{y_f^2}{8\pi^2}[2\Lambda^2 + 6m_f^2 \ln(\Lambda/m_f) + ...], \tag{118}$$

where Λ is an ultraviolet cutoff used to represent the scale up to which the SM remains valid, at which new physics appears. We see that the mass of the Higgs diverges quadratically with Λ and, if we suppose that the SM remains valid up to the Planck scale, $M_P \simeq 10^{19}$ GeV, then $\Lambda = M_P$ and this correction is 10^{30} times bigger than the reasonable value of the mass-squared of the Higgs, namely (10^2) GeV)^2! Moreover, there is a similar correction coming from a loop of a scalar field S, such as that in Fig. 18(b):

$$\Delta m_H^2 = \frac{\lambda_S}{16\pi^2}[\Lambda^2 - 2m_S^2 \ln(\Lambda/m_S) + ...], \tag{119}$$

where Λ_S is the quartic coupling to the Higgs boson.

Comparing (118) and (119), we see that the divergent contributions terms $\propto \Lambda^2$ are cancelled if, for every fermionic loop of the theory there is also a scalar loop with $\lambda_S = 2y_f^2$. *We will see later that supersymmetry imposes exactly this relationship!* Thus supersymmetric field theories have no quadratic divergences, at both the one- and multi-loop levels, which enables a large hierarchy between different

[10] For this calculation, we define the Yukawa coupling of the Higgs boson to a fermion, as usual, *via*: $y_f H\overline{\psi}\psi$.

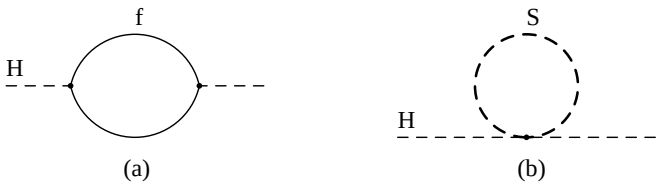


Fig. 18: One-loop quantum corrections to the mass-squared of the Higgs boson due to (a) a fermionic loop, (b) a scalar boson loop

physical mass scales to be maintained in a natural way. In addition, other logarithmic corrections to couplings also vanish in a supersymmetric theory [87].

• A second circumstantial hint in favour of supersymmetry is the fact, discussed in the previous Lecture, that precision electroweak data prefer a relatively light Higgs boson weighing less than about 150 GeV [16]. This is perfectly consistent with calculations in the minimal supersymmetric extension of the Standard Model (MSSM), in which the lightest Higgs boson weighs less than about 130 GeV [88].

• A third motivation for supersymmetry is provided by the astrophysical necessity of cold dark matter, which has a density of $\Omega_{CDM}h^2 = 0.1099 \pm 0.0062$ according to the recent measurements of WMAP [33]. This dark matter could be provided by a neutral, weakly-interacting particle weighing less than about 1 TeV, such as the lightest supersymmetric particle (LSP) χ [34]. In many supersymmetric models, a conserved quantum number called R parity guarantees that the LSP is stable. As the Universe expanded and cooled, all the particles present at high energies and densities would have annihilated, disintegrated, or combined to form baryons, atoms, etc., except for stable weakly-interacting particles such as the neutrinos and the LSP. The latter would be present in the Universe as a relic from the Big Bang, and could have the right density to constitute the majority of the cold dark matter favoured by cosmologists.

• Fourthly, let us consider the couplings that characterize each of the fundamental forces. As seen in the left panel of Fig. 19, it has been known for a long time now that if we evolve them with energy according to the renormalization-group equations of the Standard Model, we find that they never quite become equal at the same scale. However, as seen in the right panel of Fig. 19, when we include supersymmetric particles in the evolution of the couplings, they appear to intersect at exactly the same energy scale (about 2×10^{16} GeV) [89]. Nobody is forced to believe in such a 'Grand Unification' on the basis of this possible unification of the couplings, but it is very intriguing that supersymmetry favours unification with high precision.

• Fifthly, supersymmetry seems to be essential for the consistency of string theory [90], although this argument does not really restrict the mass scale at which supersymmetric particles should appear.

• A final hint for supersymmetry may be provided by the anomalous magnetic moment of the muon, $g_\mu - 2$, whose experimental value [91] seems to differ from that calculated in the SM, in a manner that could be explained by contributions from supersymmetric particles. The amount of this discrepancy depends on how one calculates the SM contributions to $g_\mu - 2$, in particular that due to low-energy hadronic vacuum polarization, and to a lesser extent that due to light-by-light scattering. The most direct way to calculate the hadronic vacuum polarization contribution is to use low-energy data on $e^+e^- \to$ hadrons: these do not agree perfectly, but may be combined to yield a discrepancy [92]

$$\delta a_\mu \equiv \delta \left(\frac{g_\mu - 2}{2} \right) = (24.6 \pm 8.0) \times 10^{-10}, \tag{120}$$

a discrepancy of 3.1 σ, as illustrated in Fig. 20. Alternatively, and less directly, one may use τ decay data, in which case the discrepancy is reduced to about 2 σ.

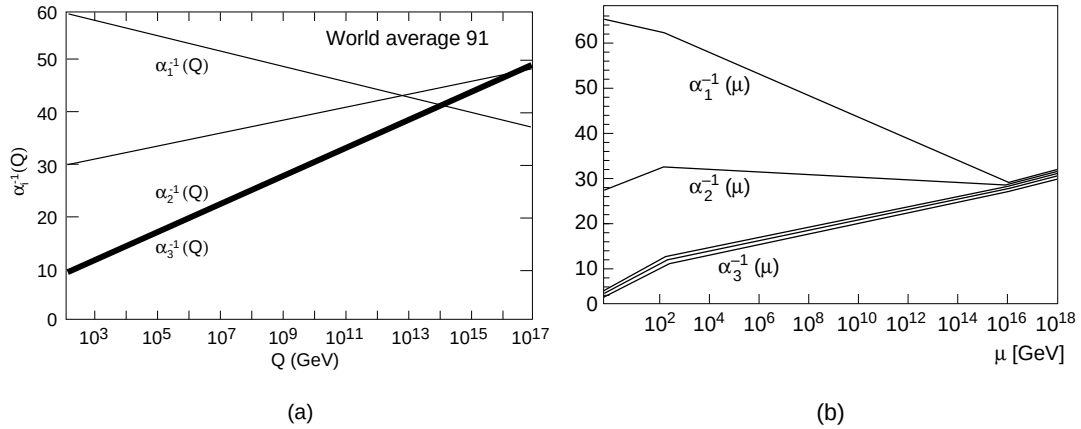


Fig. 19: The measurements of the gauge coupling strengths at LEP (a) do not evolve to a unified value if there is no supersymmetry but do (b) if supersymmetry is included [89]

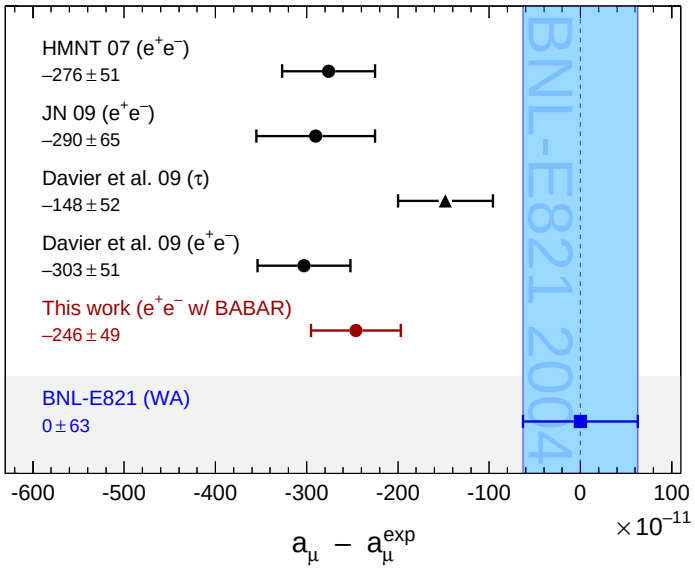


Fig. 20: SM calculations of $a_\mu \equiv (g_\mu - 2)/2$ disagree with the experimental measurement [91], particularly if they are based on low-energy e^+e^- data [73].

As we have seen, there are several arguments that motivate the study of supersymmetry [11]. Although there are no experimental proofs of its existence, supersymmetry combines so many attractive and useful characteristics that it deserves to be studied in detail.

3.2 The structure of a supersymmetric theory

3.2.1 Interlude on 'spinorology'

In order to lay the basis for the theoretical description of supersymmetry [84], we first present the notations and conventions that we use in the rest of the section [11, 87].

[11] Other extensions of the SM also address some of these issues, though perhaps none do so as naturally as supersymmetry.

• We choose the *Weyl representation* for the γ matrices:

$$\gamma^\mu = \begin{pmatrix} 0 & \sigma^\mu \\ \overline{\sigma}^\mu & 0 \end{pmatrix}, \tag{121}$$

with $\sigma^\mu = (\mathbf{1}_2, \sigma^i)$, $\overline{\sigma}^\mu = (\mathbf{1}_2, -\sigma^i)$ where σ_i are the Pauli matrices, and $\gamma_5 = i\gamma^0\gamma^1\gamma^2\gamma^3 = \mathrm{diag}(-\mathbf{1}_2, \mathbf{1}_2)$. We also use $\{\gamma^\mu, \gamma^\nu\} = 2\eta_{\mu\nu}$, where $\eta_{\mu\nu} = \mathrm{diag}(+1, -1, -1, -1)$ is the Minkowski metric, that may be used to lower or to raise Lorentz indexes.

• A *Weyl spinor* describes a particle of spin $1/2$ and given chirality. It has two components, which we label with Greek letters, ψ_α, ξ_β, ... where $\alpha,\ \beta, ... = 1, 2$. A spinor ψ_α or ψ_L will denote a particle with left chirality, whereas we denote by $\overline{\psi}^{\dot\alpha}$ or ψ_R a spinor with right chirality. These are related by complex conjugation:

$$(\psi_\alpha)^* = \overline{\psi}_{\dot\alpha} \quad , \tag{122}$$

$$(\overline{\psi}^{\dot\alpha})^* = \psi^\alpha \quad . \tag{123}$$

We also use the matrix $\varepsilon_{\alpha\beta} = \varepsilon_{\dot\alpha\dot\beta} \equiv i\sigma_2$ and $\varepsilon^{\alpha\beta} = \varepsilon^{\dot\alpha\dot\beta} \equiv -i\sigma_2$, which allows us to raise and lower the spinorial indices α and β.

• A *Dirac spinor* is constructed out of two Weyl spinors, and describes a particle with both chiralities. It is a spinor of four components, which we denote here using capital Greek letters: Ψ, χ, Φ, ... In terms of Weyl spinors, we have

$$\Psi = \begin{pmatrix} \psi_L \\ \psi_R \end{pmatrix} = \begin{pmatrix} \psi_\alpha \\ \overline{\eta}^{\dot\alpha} \end{pmatrix}. \tag{124}$$

The projection operators $P_{R,L} = \frac{1}{2}(1 \pm \gamma_5)$ allow us to select the right or left chiralty, respectively: $\Psi_{R,L} = P_{R,L}\Psi$.

• A *charge conjugate spinor* is a spinor to which charge conjugation has been applied. It describes the antiparticle of a given particle, with opposite internal opposite charge.

$$\Psi^c = C\overline{\Psi}^T = \begin{pmatrix} \eta_\alpha \\ \overline{\psi}^{\dot\alpha} \end{pmatrix}, \tag{125}$$

where the charge conjugation matrix C can be written:

$$C = i\gamma^0\gamma^2. \tag{126}$$

• A *Majorana spinor* is constructed out of a single Weyl spinor, but possesses four components that are interrelated by charge conjugation, so that $\Psi_M = \Psi_M^c$:

$$\Psi_M = \begin{pmatrix} \psi_L \\ -i\sigma_2(\psi_L)^* \end{pmatrix} = \begin{pmatrix} \psi_\alpha \\ \overline{\psi}^{\dot\alpha} \end{pmatrix}. \tag{127}$$

3.2.2 The supersymmetry algebra and supermultiplets

As was described before, supersymmetry combines the space-time transformations of the Poincaré group with transformations of an internal symmetry. Prior to the advent of supersymmetry, there had been many previous attempts to combine internal and external symmetries, but they had always failed, for a reason demonstrated by Coleman and Mandula [73]. All the previous attempts used bosonic charges, scalar (or vector) such as the electromagnetic charge (or momentum operator):

$$\langle \mathrm{Spin}J|Q|\mathrm{Spin}J\rangle \quad = \quad q, \tag{128}$$

$$\langle \mathrm{Spin}J|P_\mu|\mathrm{Spin}J\rangle \quad = \quad p_\mu. \tag{129}$$

Conservation of momentum in any $2 \to 2$ collision implies

$$p_\mu^{(1)} + p_\mu^{(2)} \;=\; p_\mu^{(3)} + p_\mu^{(4)}. \tag{130}$$

Consider now a tensor charge $\Sigma_{\mu\nu}$: by Lorentz invariance, its diagonal matrix elements in any particle state $|a\rangle$ must be of the form

$$\langle a|\Sigma_{\mu\nu}|a\rangle \;=\; \alpha g_{\mu\nu} + \beta p_\mu p_\nu. \tag{131}$$

Conservation of the tensor charge during a $2 \to 2$ collision would require

$$p_\mu^{(1)} p_\nu^{(1)} + p_\mu^{(2)} p_\nu^{(2)} \;=\; p_\mu^{(3)} p_\nu^{(3)} + p_\mu^{(4)} p_\nu^{(4)}. \tag{132}$$

This is compatible with the linear relation (130) of conventional momentum conservation iff

$$p_\mu^{(1)} \;=\; p_\mu^{(3)} \text{ or } p_\mu^{(4)}, \tag{133}$$

implying that only exactly forward and backward scattering are allowed: no need to place any detectors at large angles! This proof can easily be extended to bosonic charges with any number of indices. However, it makes the crucial assumption that the diagonal matrix element $\langle a|Q|a\rangle \neq 0$, which is not true in supersymmetry, enabling it to evade the Coleman–Mandula no-go theorem.

Supersymmetry is generated by *spinorial* charges Q_α which have vanishing diagonal matrix elements: $\langle a|Q_\alpha|a\rangle = 0$. Being spinors, the Q_α anti-commute in the same way as other fermionic fields. It is possible to introduce more generators, but in the simplest version of supersymmetry there is just a pair of generators, Q_α and $\bar{Q}^{\dot{\alpha}}$, that are complex spinors transforming inequivalently under the Lorentz group. This is $\mathcal{N} = 1$ supersymmetry, which is essentially the only case that we consider in these notes. The initial reason for this choice is pedagogical, but in the following section we give some physical reasons for such a choice.

The algebra of the supersymmetry (like that of any other symmetry) is summarized in the commutation (and anticommutation) relations of its generators, i.e., its Lie (super)algebra. In addition to the commutation relations of the Poincaré algebra, the supersymmetry algebra includes the following relations for the generators Q_α y $\bar{Q}^{\dot{\alpha}}$:

$$[P^\mu, Q_\alpha] \quad = 0 = [P^\mu, \bar{Q}^{\dot{\alpha}}], \tag{134}$$

$$\{Q_\alpha, \bar{Q}_{\dot{\beta}}\} \quad = 2(\sigma_\mu)_{\alpha\dot{\beta}} P^\mu, \tag{135}$$

$$\{Q_\alpha, Q_\beta\} \quad = \{\bar{Q}^{\dot{\alpha}}, \bar{Q}^{\dot{\beta}}\} = 0, \tag{136}$$

$$\{M_{\mu\nu}, Q_\alpha\} \quad = \frac{1}{2}(\sigma_{\mu\nu})_\alpha^\beta Q_\beta, \tag{137}$$

$$\{M_{\mu\nu}, \bar{Q}_{\dot{\alpha}}\} \quad = \frac{1}{2}(\overline{\sigma}_{\mu\nu})_{\dot{\alpha}}^{\dot{\beta}} \bar{Q}_{\dot{\beta}}. \tag{138}$$

What is the significance of Q_α? First, Q is a charge in the sense of Noether's theorem, i.e, it is the charge conserved by the symmetry. As a conserved charge, it commutes with the Hamiltonian of the system and is invariant under translations, see (134). Since it possesses spin 1/2 and has two complex components, it can be written as a Weyl spinor, or alternatively as a Majorana spinor with 4 components: as such, its commutation relations with the Lorentz generators are completely determined, see (137) and (138). The non-trivial anticommutation relation above is (135): schematically $\{Q, \bar{Q}\} \sim P$, which means that Q is the 'square root' of a space-time translation.

If we want to apply supersymmetry to particle physics, we must know how to arrange particles in irreducible representations (supermultiplets), and their transformation properties. Therefore, we now study the supermultiplets and detail their contents. We recall that the Poincaré group has two Casimir

invariant elements, the spin invariant $W^2 = W^\mu W_\mu$, where $W^\mu = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma} P_\nu M_{\rho\sigma}$ is the Pauli-Lubanski vector, and the mass invariant $P^2 = P^\mu P_\mu$, where P^μ is the four-momentum. In a multiplet of the Poincaré group, the particles have the same masses and the same spins. However, in the case of supersymmetry, W^2 is not an invariant of the algebra, so only mass is conserved, not spin:

$$[P^2, Q_\alpha] = 0, \tag{139}$$

$$[W^2, Q_\alpha] \neq 0. \tag{140}$$

Thus, in a supermultiplet, the particles have the same mass but different spins. We can nevertheless modify W to obtain a new invariant whose eigenvalues are of the form $2j(j+1)m^4$ with $j = 0, \frac{1}{2}, 1, ...$ the quantum number of this 'superspin'. This modified W is an invariant, so every irreductible representation can be characterized by a pair $[m, j]$, and the relation between the spin S and j is deduced from the relation: $M_S = M_j, M_j + \frac{1}{2}, M_j - \frac{1}{2}, M_j$. Within a given supermultiplet, there are particles of the same mass and the same superspin. In addition, an important property of any supermultiplet is that there are equal numbers of bosonic and fermionic degrees of freedom: $n_B = n_F$.

We can construct now two different supermultiplets:

▷ The fundamental representation $[m, 0]$ is called a chiral supermultiplet. The value $j = 0$ implies $M_S = 0, +\frac{1}{2}, -\frac{1}{2}, 0$, and this supermultiplet Ψ contains two real scalar fields described by a single complex scalar field (the sfermion), ϕ, and a two-component Weyl fermionic field of spin 1/2, ψ with the same mass:

$$\Psi = (\phi, \psi_\alpha, F). \tag{141}$$

What is F? In order that the supersymmetry be preserved in loops, where the particles are not on-shell, i.e., $P^2 \neq M^2$, it is necessary that the fermionic and bosonic degrees of freedom be balanced also *off-shell*. This is an issue because an off-shell Weyl fermion possesses 4 spin degrees of freedom, as opposed to 2 on-shell. It is necessary to add to the on-shell content of this representation another scalar complex field F that does not propagate, and does not correspond to a physical particle. This is termed an auxiliary field, and does not have a kinetic term, and the equation of motion $F = F^* = 0$ may be used to eliminate it when on-shell.

▷ The second representation we use later is the vector (or gauge) supermultiplet $[m, 1/2]$, denoted by Φ. Its field content is obtained in the same way: a Weyl fermion (or, equivalently, a Majorana fermion), called the gaugino λ^a_α, a gauge boson (of zero mass) A^μ_a, and in the presence of any chiral supermultiplet, an auxiliar real scalar field, D^a:

$$\Phi = (\lambda^a_\alpha, A^a_\mu, D^a), \tag{142}$$

where a is an index of the gauge group.

These two representations may be used to accommodate the particles of the SM and their superpartners. However, before doing so, we first construct with these two representations generic supersymmetric field theories.

3.3 Supersymmetric field theories

Before discussing supersymmetric models in general, and particularly the minimal supersymmetric extension of the SM (the MSSM), we first present, without detailed derivations, the general structure of a field theory with supersymmetry. We first introduce the model of Wess and Zumino [78] without interactions to see how the fields transform. Then we introduce the interactions, which will lead us to the new notion of the superpotential. Finally, we discuss gauge fields in a supersymmetric theory. At the end of this section, we will have accumulated enough theoretical baggage to understand the structure of the MSSM, and be able to study concretely its experimental predictions.

3.3.1 The action for free bosons and fermions is globally supersymmetric

The simplest supersymmetric action is the combination of actions for a non-interacting massless complex scalar $\phi$and a spin-$1/2$ fermion ψ:

$$S = \int d^4x \, (\mathcal{L}_{scalar} + \mathcal{L}_{fermion}) : \tag{143}$$

$$\mathcal{L}_{scalar} = -\partial^\mu \phi \, \partial_\mu \phi^*, \tag{144}$$

$$\mathcal{L}_{fermion} = -i\psi^\dagger \bar{\sigma}^\mu \, \partial_\mu \psi. \tag{145}$$

If we introduce an infinitesimal supersymmetric global transformation parameter ϵ_α, which is a Weyl fermion independent of the space-time coordinates ($\partial^\mu \epsilon_\alpha = 0$), and apply it to the scalar field ϕ, the result must be proportional to the fermionic field ψ:

$$\delta\phi = \epsilon^\alpha \psi_\alpha \text{ and } \delta\phi^* = \bar{\epsilon}_{\dot{\alpha}} \, \bar{\psi}^{\dot{\alpha}}, \tag{146}$$

leading to

$$\delta\mathcal{L}_{scalar} = -\epsilon^\alpha \, (\partial^\mu \psi_\alpha) \, \partial_\mu \phi^* - \partial^\mu \phi \, \bar{\epsilon}_{\dot{\alpha}} \, (\partial_\mu \bar{\psi}^{\dot{\alpha}}). \tag{147}$$

Since the mass dimensions of free boson and fermion fields are

$$[\phi] = 1, \;\; [\psi] = \frac{3}{2}, \tag{148}$$

the infinitesimal fermion ϵ_α must have the dimensionality $(mass)^{-1/2}$:

$$[\epsilon] = -\frac{1}{2}, \tag{149}$$

in contrast to an usual Weyl fermion that has dimension $(mass)^{3/2}$ (148). By simple dimensional counting, the infinitesimal transformation of the fermion field must therefore be proportional to the derivative of the boson field:

$$\delta\psi_\alpha = i(\sigma^\mu \epsilon^\dagger)_\alpha \, \partial_\mu \phi \text{ and } \delta\bar{\psi}^{\dot{\alpha}} = -i(\epsilon \, \sigma^\mu)^{\dot{\alpha}} \, \partial_\mu \phi^*. \tag{150}$$

Combining (146) and (150) and using the equations of motion, we see that the sum $\delta L_{scalar} + \delta L_{fermion}$ is a total divergence. This implies that the combined action, which is the space-time integral of the two free Lagrangians $L_{scalar} + L_{fermion}$, is invariant under this pair of transformations.

Does this transformation correspond to a supersymmetry transformation? To convince ourselves that this is the case, it is enough to start from a fermion ψ or from a boson ϕ, and to apply these transformations twice. We find the following chain:

$$\phi \to \psi \to \partial\phi, \;\; \psi \to \partial\phi \to \partial\psi, \tag{151}$$

which means that in both cases the combined effects of two successive supersymmetry transformations are equivalent to a space-time derivative ∂^μ, and hence to the momentum operator $P^\mu \sim i\partial^\mu$. Thus we recover the result of the previous section, namely $Q^2 \sim P$, and our transformations satisfy the supersymmetric algebra. This free Lagrangian model is actually the simplest Wess–Zumino model with a single chiral supermultiplet, without mass and without interactions.

If we wish to preserve supersymmetry *off-shell*, which will be essential once we include interactions, we cannot use the equations of motion to demonstrate supersymmetry. To overcome this problem, as discussed earlier, the action S must be modified by the addition of a term that contains an auxiliary field F:

$$S = \int d^4x \, (\mathcal{L}_{scalar} + \mathcal{L}_{fermion} + \mathcal{L}_{aux}), \tag{152}$$

$$\mathcal{L}_{aux} = F^* \, F, \tag{153}$$

In the *on-shell* case, the equation of motion for F would yield $F = F^* = 0$. However, its introduction modifies the supersymmetry transformations of the fields ψ and ϕ *off-shell*. Specifically, the transformation of the field ψ is affected by the scalar field F. To see this, we first observe that the dimension of the field F is of $(mass)^2$, so that its only possible transformation law is

$$\delta F = i\,\bar{\epsilon}^{\dot{\alpha}}\,(\overline{\sigma}^{\mu})^{\beta}_{\dot{\alpha}}\,\partial_{\mu}\psi_{\beta} \;\text{ and }\; \delta F^{*} = -i\,\partial_{\mu}\bar{\psi}^{\dot{\beta}}\,(\bar{\sigma}^{\mu})^{\alpha}_{\dot{\beta}}\,\epsilon_{\alpha}\,. \tag{154}$$

The variation of the term $\mathcal{L}_{aux}$ in S therefore gives

$$\delta\mathcal{L}_{aux} = i\,\bar{\epsilon}\,(\overline{\sigma}^{\mu})\,\partial_{\mu}\psi\,F^{*} - i\,\partial_{\mu}\bar{\psi}\,(\bar{\sigma}^{\mu})\,\epsilon\,F. \tag{155}$$

In the *on-shell* case, as we have already seen, the equation of motion for F would yield $F = F^* = 0$, and the variation (154) would also vanish, thanks to the equation of motion for ψ. To compensate the variation (155) in the *off-shell* case, we see that we require a supplementary term in the transformation law for ψ:

$$\delta\psi_{\alpha} = i(\sigma^{\mu}\bar{\epsilon})_{\alpha}\,\partial_{\mu}\phi + \epsilon_{\alpha}F \;\text{ et }\; \delta\bar{\psi}^{\dot{\alpha}} = -i(\epsilon\,\sigma^{\mu})^{\dot{\alpha}}\,\partial_{\mu}\phi^{*} + \bar{\epsilon}^{\dot{\alpha}}F^{*}. \tag{156}$$

Once again, the supplementary term vanishes when the on-shell condition $F = 0$ is applied. For simple dimensional reasons, the transformations of ϕ are not affected. It is easy to check that $\delta S = 0$ without using the equations of motion, and hence supersymmetry continues to be satisfied off-shell, thanks to the appearance of the auxiliary field F.

In fact, the auxiliary field plays an additional role. We must not forget that we have not observed supersymmetry in the range of energies explored so far. Hence, if supersymmetry exists at all in Nature, it must be broken in some way. The auxiliary field F (and the other auxiliary field D that we meet later) serve to break supersymmetry if their v.e.v.s are non-zero, as we will see in the last part of this section.

3.3.2 Interactions of the chiral multiplets

We now add to the theory interactions between the scalar and fermion fields that comprise chiral supermultiplets. The most general form of interaction that is at most quadratic in the fermion fields is

$$\mathcal{L}_{int} = -\frac{1}{2}W^{ij}(\phi,\phi^{*})\psi_{i}\psi_{j} + V(\phi,\ \phi^{*}) + c.c. \tag{157}$$

We do not demonstrate it in detail, but the quantity W^{ij} must be an analytic function of the fields ϕ_i, i.e., it does not depend on the ϕ_i^*, in order to ensure that the variation due to a supersymmetry transformation of the first term of $\mathcal{L}_{int}$ can be compensated by the variation of another term (basically because supersymmetry transforms ψ_i into ϕ_i and *vice versa*). For the same reason, W^{ij} must be completely symmetric. Hence W^{ij} must be of the form:

$$W^{ij} = \frac{\partial^2 W(\phi)}{\partial\phi_i\,\partial\phi_j}, \tag{158}$$

where the object W is called the *superpotential*. In order for the model to be renormalizable, the term in (157) that is bilinear in the fermion fields ψ_i can have at most a linear dependence on the scalar fields ϕ_i, implying that W can be at most cubic:

$$W = \frac{1}{2}m^{ij}\phi_i\phi_j + \frac{1}{6}y^{ijk}\phi_i\phi_j\phi_k \tag{159}$$

in the context of a renormalizable theory. Remarkably, apart from wave-function renormalization of the fields, there is no intrinsic renormalization of the superpotential parameters.

In general, the superpotential has dimension $(mass)^3$. The quadratic term in W (159) provides the (symmetric) mass matrix m^{ij} of the fermions, which is equal to the mass matrix of the scalar bosons,

by virtue of supersymmetry. The trilinear term in W provides the matrix of Yukawa couplings y^{ijk} betweeen a scalar and two fermions, and summarizes all the interactions that are not gauge interactions. As already noted, W is an analytical function of the complex fields ϕ_i, which has an importance that we discuss later.

The requirement that $\mathcal{L}_{int}$ be invariant under supersymmetry transformations also determines the form of the potential V. In presence of interactions, i.e., if the superpotential is non-zero, the auxiliary fields F^i introduced earlier (153) can be written in the form:

$$F_i = -\frac{\partial W(\phi)}{\partial \phi^i} = -W_i^*, \qquad F^{*i} = -\frac{\partial W(\phi)}{\partial \phi_i} = -W^i. \tag{160}$$

We may therefore write the Lagrangian without introducing explicitly the F fields, in which case the potential V of the theory is:

$$V = W_i^* W^i = F_i F^{*i}. \tag{161}$$

That is automatically non-negative, since it is a sum of modulus-squared terms. If we use the general form (159) of the superpotential, we have the general Lagrangian:

$$\mathcal{L} = -\partial^\mu \phi\, \partial_\mu \phi^* - i\psi^\dagger \bar{\sigma}^\mu\, \partial_\mu \psi - \frac{1}{2} m^{ij} \psi_i \psi_j - \frac{1}{2} m^*_{ij} \psi^{\dagger i} \psi^{\dagger j} - V - \frac{1}{2} y^{ijk} \phi_i \psi_j \psi_k - \frac{1}{2} y^*_{ijk} \phi^{*i} \psi^{\dagger j} \psi^{\dagger k}, \tag{162}$$

where V is given by (161), (160) and (159). It is easy to see from (159) that the boson and fermion masses are equal, as one would expect from supersymmetry.

3.3.3 Supersymmetric gauge theories

In addition to chiral fermions (quarks, leptons), the SM contains gauge fields of spin 1 (W and Z bosons, photons and gluons). In the section dedicated to the supersymmetry algebra, we saw that vector supermultiplets would provide the appropriate frameworks for such gauge fields. We now study the properties of such a supermultiplet, both with and without interactions [79]. We recall that a vector supermultiplet contains a massless gauge boson A_a^μ and a massless Weyl fermion, the gaugino λ_a, both in the adjoint representation of the gauge group. In order to go off-shell, one must introduce an auxiliary real scalar field D_a analogous to the auxiliary field F introduced for the chiral supermultiplet.

The form of the Lagrangian is completely determined by the condition of gauge invariance and of renormalizability:

$$\mathcal{L}_{gauge} = -\frac{1}{4} F^a_{\mu\nu} F^{a\mu\nu} - i\lambda^{a\dagger} \bar{\sigma}^\mu D_\mu \lambda^a + \frac{1}{2} D^a D^a, \tag{163}$$

where the gauge covariant derivative D_μ and $F^a_{\mu\nu}$ take the forms:

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - g f^{abc} A^b_\mu A^c_\nu, \tag{164}$$

$$D_\mu \lambda^a = \partial_\mu \lambda^a - g f^{abc} A^b_\mu, \tag{165}$$

as usual for a gauge theory. Remarkably, this Lagrangian is already supersymmetric, as can be checked using the following supersymmetry transformations for the fields of the vector supermultiplet:

$$\delta A^a_\mu = \frac{1}{\sqrt{2}} \left(\epsilon^\dagger \bar{\sigma}^\mu \lambda^a + \lambda^{a\dagger} \bar{\sigma}^\mu \epsilon \right), \tag{166}$$

$$\delta \lambda^a_\alpha = -\frac{i}{2\sqrt{2}} (\sigma^\mu \bar{\sigma}^\nu \epsilon)_\alpha F^a_{\mu\nu} + \frac{1}{\sqrt{2}} \epsilon_\alpha D^a, \tag{167}$$

$$\delta D^a = \frac{i}{\sqrt{2}} \left(\epsilon^\dagger \bar{\sigma}^\mu D_\mu \lambda^a - D_\mu \lambda^{a\dagger} \bar{\sigma}^\mu \epsilon \right). \tag{168}$$

In the absence of any interactions with chiral supermultiplets, the equation of motion for the auxiliary field D^a is simply $D^a = 0$, as seen directly from the Lagrangian (163), since it does not have a kinetic term and therefore does not propagate.

However, in the SM the gauge fields do interact with the chiral fermions. Hence, in our supersymmetric version we have to consider interactions between chiral supermultiplets and vector supermultiplets. As in the SM, the usual derivatives ∂^μ of the fermions must be replaced by gauge-covariant derivatives D^μ, and the same applies to their scalar supersymmetric partners. The supersymmetric transformation laws of the chiral supermultiplets must be changed to take into account the variations of these new terms. As a result, the equation of motion for D^a becomes:

$$D^a = -g(\phi^* T^a \phi), \tag{169}$$

where the T^a are the generators of the gauge group and g is its coupling constant, and the full scalar potential is

$$V = F_i F^{*i} + \frac{1}{2}\sum_a D^a D^a = W_i^* W^i + \frac{1}{2}\sum_a g^2(\phi^* T^a \phi)^2. \tag{170}$$

This potential is completely determined by the Yukawa couplings (*via* the F term) and by the gauge interactions (*via* the D term). The full scalar potential is automatically non-negative, which is important for the spontaneous breaking of the symmetry.

In a globally supersymmetric theory, spontaneous breaking may occur *via* a v.e.v. for the D term or the F term, either of which would give a positive contribution to the vacuum energy. However, it is difficult to construct models that are interesting for phenomenology, and most model-builders pursue the spontaneous breaking of local supersymmetry in the context of a supergravity theory, in which this positive contribution may be cancelled.

3.4 Low-energy supersymmetric models

In this section we apply the results obtained in the previous section, with the objective of supersymmetrizing the Standard Model while preserving its successful characteristics. The minimal supersymmetric extension of the SM is called the MSSM [85, 86]. We will present its particle content (including the nomenclature of the new particles), we will discuss how the electroweak symmetry may broken, and we will outline an effective framework for describing the breaking of supersymmetry. Later we will present typical predictions of the MSSM. Along the way, we will also mention possible variants of the MSSM, because Nature might very well have chosen a path more complex than this minimal model.

3.4.1 How many supersymmetries?

As well as mentioned already, the number of supersymmetric generators Q_α may be $\mathcal{N} \geq 1$. Supersymmetric theories with $\mathcal{N} \geq 2$ have some characteristic advantages, e.g., they have fewer divergences, which make them very interesting theoretically. Specifically, in the $\mathcal{N} = 2$ case there is only a finite number of divergent Feynman diagrams, and in the $\mathcal{N} = 4$ case there are none, i.e., any theory with $\mathcal{N} = 4$ supersymmetries is intrinsically finite, and it is easy to construct finite $\mathcal{N} = 2$.

Unfortunately, it is not possible to construct realistic models with $\mathcal{N} \geq 2$, because they do not allow the violation of parity that is observed in the weak interactions. This is because a supermultiplet of a theory with $\mathcal{N} \geq 2$ supersymmetries necessarily incorporates both left- and right-handed fermions in the same supermultiplet: applying a supersymmetry charge Q changes the helicity by 1/2, so applying two charges relates states with helicity $\pm 1/2$, implying that they are in the same representation of the gauge group, and hence have the same interactions. This contradicts experimental observations, which tell us, for example, that the left-handed electron (which forms part of a doublet in the SM) does not have the same interaction with W bosons as the right-handed electron (which is a singlet with zero electroweak isospin that does not feel the $SU(2)$ weak interaction). Models with $\mathcal{N} \geq 2$ cannot describe the physics of the SM particles observed at low energy.

3.4.2 The particle content in the MSSM

The supermultiplets in the minimal $\mathcal{N} = 1$ case are

- the chiral supermultiplet that includes a fermion of spin 1/2 and a boson of spin 0,
- the vector supermultiplet that includes a boson of spin 1 and one fermion of spin 1/2.

Could we link the particles of the SM in such multiplets, i.e., could we associate quarks and leptons with the bosons W, Z, the photon, and so on? The answer is no, because this would raise problems for the conservation of their quantum numbers. Specifically, the gauge bosons and the fermions do not have the same transformation properties under the SM gauge group, since they possess different quantum numbers, e.g., quarks are triplets of the colour group whereas gauge bosons are either octets (the gluons) or singlets (the other gauge bosons), and leptons carry lepton numbers whereas gauge bosons do not. Simple $\mathcal{N} = 1$ supersymmetry does not modify these quantum numbers, so we cannot associate any gauge boson with a known fermion or *vice versa*. Therefore, we have to postulate unseen supersymmetric partners for all the known particles. Table 3 lists, for every SM particle, the name, spin and notation for its spartner.

Table 3: Particle content of the MSSM

Particle	Spartner	Spin
quarks q	squarks $\tilde{q}$	0
$\rightarrow$ top t	stop $\tilde{t}$	
$\rightarrow$ bottom b	sbottom $\tilde{b}$	
...		
leptons l	sleptons $\tilde{l}$	0
$\rightarrow$ electron e	selectron $\tilde{e}$	
$\rightarrow$ muon μ	smuon $\tilde{\mu}$	
$\rightarrow$ tau τ	stau $\tilde{\tau}$	
$\rightarrow$ neutrinos ν_ℓ	sneutrinos $\tilde{\nu}_\ell$	
gauge bosons	gauginos	1/2
$\rightarrow$ photon γ	photino $\tilde{\gamma}$	
$\rightarrow$ boson Z	Zino $\tilde{Z}$	
$\rightarrow$ boson B	Bino $\tilde{B}$	
$\rightarrow$ boson W	Wino $\tilde{W}$	
$\rightarrow$ gluon g	gluino $\tilde{g}$	
Higgs bosons $H_i^{\pm,0}$	higgsinos $\tilde{H}_i^{\pm,0}$	1/2

Before going on to the following sections, we make a few observations. First, we note that the spartners of SM fermions and gauge bosons are of lower spin. *A priori*, one could have considered associating the fermions of the SM with spartners of spin 1, and the gauge bosons with spartners of spin 3/2. However, to introduce a particle of spin 1 would require introducing a new gauge interaction, and hence a non-minimal model. Also, introducing particles of spin > 1 would make the theory non-renormalizable, i.e., it would no longer be possible to absorb the divergences in perturbation theory in a finite number of physical quantities [12].

Secondly, we recall that in the SM the right-handed fermions have different interactions from the left-handed fermions, e.g., being singlets of $SU(2)$ instead of doublets. In supersymmetry, the left- and right-handed must belong to different supermultiplets, and have distinct spartners, e.g., $q_L \rightarrow \tilde{q}_L$ and

[12] Supergravity does allow a restricted number $\mathcal{N} \leq 8$ of spin-3/2 gravitino partners of the spin-2 graviton to be introduced, but they do not carry conventional gauge interactions.

$q_R \to \tilde{q}_R$. These two squarks are quite different, and we use the chirality index L or R to identify them, even though the concept of handedness does not make physical sense for a scalar particle, whose only helicity is $\lambda = 0$. In general, the $\tilde{f}_L$ and $\tilde{f}_R$ mix, and the physical mass eigenstates are combinations of them. In constructing the Yukawa interactions of the MSSM, it is often convenient to work with superfields that comprise conjugates of the $\tilde{f}_R$ and their scalar spartners: these are left-handed chiral supermultiplets denoted by F^c.

Thirdly, we note that, besides the new spartners, at least two doublets of Higgs bosons are required. To understand why, we recall that, in the study of supersymmetric theories, we introduced the notion of the superpotential. This governs all the possible Yukawa interactions of the matter particles with the Higgs fields. In the SM, if we use a Higgs field h to give masses to the quarks of type 'down', *via* Yukawa couplings $q\bar{d}h$, we could use the complex conjugate field h^* to give masses to quarks of type 'up', *via* couplings $q\bar{u}h^*$. However, we recall that in a supersymmetric theory the superpotential is an analytic function of the superfields that cannot depend on their complex conjugates. Therefore, we must use separate Higgs supermultiplets (denoted by capital letters) with opposite hypercharge quantum numbers, and interactions of the forms QD^cH_d and QU^cH_u. Charged leptons may acquire masses through interactions of the form LE^cH_d. We also note that pairs of Higgs superfields are needed in order to cancel the triangle anomalies that would be generated by higgsino fermion loops.

Fourthly, we note that in general the $\tilde{\gamma}, \tilde{Z}, \tilde{W}$ and $\tilde{H}$ mix, and the experimentally observable mass eigenstates are combinations of these gauginos and higgsinos that are generally named neutralinos $\tilde{N}^0_{1,2,3,4}$, which have zero electrical charge, and charginos $\tilde{C}^{\pm}_{1,2}$ [13], which are electrically charged and mix the $\tilde{W}^{\pm}$ and the $\tilde{H}^{\pm}$.

3.4.3 Interactions in the MSSM

The MSSM is the minimal supersymmetric extension of the Standard Model [85,86]. The quarks and the leptons are put together in chiral superfields with their superpartners that have the same charges under $SU(3)_C$, $SU(2)_L$ y $U(1)_Y$. The gauge bosons are placed with their fermionic superpartners in vector superfields. The superpotential of the MSSM is

$$\mathcal{W} = \mathcal{Y}_u Q U^c H_u + \mathcal{Y}_d Q D^c H_d + \mathcal{Y}_e L E^c H_d + \mu H_u H_d, \tag{171}$$

where we recall that the Q and L are the superfields containing the left-handed quarks and leptons, respectively, and the U^c, D^c and E^c are the superfields containing the left-handed antiquarks and antileptons, which are the charge conjugates of the right-handed quarks and leptons. Note that, for clarity, we have suppressed the $SU(2)$ indexes. The $\mathcal{Y}$ are 3×3 Yukawa matrices in flavour space, and do not have dimensions. After electroweak symmetry breaking, they give the masses to the quarks and leptons as well as the CKM angles and phases. As already mentioned, two Higgs doublets, H_u y H_d, are needed because of the analytical form of the superpotential.

The $\mu H_u H_d$ term is permitted by the symmetries of the MSSM and is required in order to have a suitable vacuum after electroweak symmetry breaking. The quantity μ has the dimension of a mass, and phenomenology requires it to be of the order of a TeV. The origin of μ is a puzzle: it might be associated to the scale of supersymmetry breaking.

The superpotential (171) determines all the non-gauge interactions of the MSSM, thanks to the formula (157), and the form of the effective potential of the theory is given by formula (170).

The next-to-minimal supersymmetric extension of the Standard Model (NMSSM) [93] is the simplest extension of the MSSM. In this model, the particle content is modified by the addition of a new singlet chiral supermultiplet S, with some additional superpotential terms:

$$\mathcal{W}_{NMSSM} = \frac{1}{6} k S^3 + \frac{1}{2} \mu_S S^2 + \lambda S H_u H_d + \mathcal{W}_{MSSM}. \tag{172}$$

[13] These are often denoted by $\tilde{\chi}^0_{1,2,3,4}$ and $\tilde{\chi}^{\pm}_{1,2}$, respectively.

The principal interest of the NMSSM is to propose a solution to the μ problem. Specifically, if the scalar part of S has a non-zero vacuum espectation value $\langle S \rangle$, the last term in (172) gives an effective μ term: $\mu_{eff} = \lambda \langle S \rangle$. Assuming that a soft supersymmetry-breaking scalar mass for S also appears in $\mathcal{L}_{soft}$, its v.e.v. is naturally of the order of $m_{soft} \sim \mathcal{O}(1)$ TeV, the typical mass scale of the other scalars and gauginos. Thus the effective value of μ is of the order of 1 TeV, rather than being a parameter whose magnitude is independent of the scale of supersymmetry breaking.

Phenomenologically the NMSSM differs from the MSSM because it allows the lightest Higgs boson to become heavier. In addition, the fermionic partner of S can mix with the four neutralinos of the MSSM. Thus the experimental signatures of the NMSSM may differ significantly from those of the MSSM.

3.4.4 Soft supersymmetry breaking

We have discussed so far the supersymmetric aspects of the MSSM. However, we know that supersymmetry must be broken: the selectron weighs more than the electron, squarks weigh more than quarks, etc. Therefore, we must introduce into the model the breaking of supersymmetry. However, the mechanism and the effective scale of its breaking are still unknown. Hence we adopt the *ad hoc* strategy of parametrizing the breaking of supersymmetry in terms of effective soft [14] low-energy supersymmetry-breaking terms that are added to the Lagrangian [94]. For a general supersymmetric theory, the form of these soft supersymmetry-breaking terms $\mathcal{L}_{soft}$ in the Lagrangian is

$$\mathcal{L} \supset \mathcal{L}_{soft} = -\frac{1}{2}(M_\lambda \lambda^a \lambda^a + \ c.c) - m_{ij}^2 \phi_j^* \phi_i + (\frac{1}{2} b_{ij} \phi_i \phi_j + \frac{1}{6} a_{ijk} \phi_i \phi_j \phi_k + \ c.c). \tag{173}$$

This breaks supersymmetry explicitly, since only the the gauginos λ^a and the scalars ϕ_i have mass terms, and the trilinear terms with coefficients a_{ijk} are also not of supersymmetric form. In the case of the MSSM, $\mathcal{L}_{soft}$ takes the following general form in terms of the spartner fields of the MSSM:

$$\begin{aligned} -\mathcal{L}_{soft} &= \frac{1}{2}(M_3 \tilde{g}\tilde{g} + M_2 \tilde{W}\tilde{W} + M_1 \tilde{B}\tilde{B} + \ c.c) \\ &+ \tilde{Q}^\dagger m_Q^2 \tilde{Q} + \bar{\tilde{U}}^\dagger m_{\bar{U}}^2 \bar{\tilde{U}} + \bar{\tilde{D}}^\dagger m_D^2 \bar{\tilde{D}} + \bar{\tilde{L}}^\dagger m_L^2 \bar{\tilde{L}} + \bar{\tilde{E}}^\dagger m_{\bar{E}}^2 \bar{\tilde{E}} \\ &+ (\bar{\tilde{U}}^\dagger a_U \tilde{Q} H_u - \bar{\tilde{D}}^\dagger a_D \tilde{Q} H_d - \bar{\tilde{E}}^\dagger a_E \tilde{L} H_d + \ c.c) \\ &+ m_{H_u}^2 H_u^* H_u + m_{H_d}^2 H_d^* H_d + (b H_u H_d + \ c.c). \end{aligned} \tag{174}$$

The masses M_3, M_2 and M_1 of the gauginos are complex in general, which introduces 6 parameters. The quantities m_Q, m_L and $m_{\bar{u}}$, are the mass matrices of the squarks and sleptons, which are hermitian 3×3 matrices in family space, adding 45 more unknown parameters. The couplings a_U, a_D, ..., are also complex 3×3 matrices, characterized by 54 parameters. In addition, the quadratic couplings of the Higgs bosons introduce 4 more parameters, so that the whole $\mathcal{L}_{soft}$ contains a total of 109 unknown parameters, including many that violate CP!

Supersymmetry itself is a very powerful principle whose implementation introduces only one new parameter (μ) in the MSSM. However, in our present state of ignorance, the breaking of supersymmetry introduces many new parameters. On the other hand, the number of soft parameters can be reduced by postulating symmetries or making supplementary hypotheses. Measuring the parameters of soft supersymmetry breaking would allow us to go beyond the phenomenological parametrization (174), and open the way to testing models of the high-energy dynamics that breaks supersymmetry.

[14] Here, the adjective 'soft' means that they do not introduce quadratic divergences.

3.4.5 Electroweak symmetry breaking and supersymmetric Higgs bosons

As we have already seen, the Higgs sector of the MSSM contains two complex doublets:

$$H_u = \begin{pmatrix} H_u^0 \\ H_u^- \end{pmatrix}, \; H_d = \begin{pmatrix} H_d^+ \\ H_d^0 \end{pmatrix}. \tag{175}$$

Electroweak symmetry breaking is a little bit more complicated than its analogue in the Standard Model. At tree level, we can write the effective scalar potential (after simplifications whose details we do not reproduce):

$$\begin{aligned} V &= (|\mu|^2 + m_{H_u}^2)|H_u^0|^2 + (|\mu|^2 + m_{H_d}^2)|H_d^0|^2 - b(H_u^0 H_d^0 + c.c) \\ &\quad + \frac{1}{8}(g_2^2 + g_1^2)(|H_u^0|^2 - |H_d^0|^2)^2. \end{aligned} \tag{176}$$

The terms proportional to $|\mu|^2$ originate from the F terms in the supersymmetric effective potential, and the terms proportional to the gauge couplings $(g_1,\ g_2)$ originate from the D terms. The other terms originate from $\mathcal{L}_{soft}$ (without mentioning the other scalars that do not play any role here). Spontaneous electroweak symmetry breaking can arise with this form of potential if the b parameter satisfies:

$$b^2 > (|\mu|^2 + m_{H_u}^2)(|\mu|^2 + m_{H_d}^2), \tag{177}$$

In addition, we want the potential to be bounded from below. Thus

$$2b < 2|\mu|^2 + m_{H_u}^2 + m_{H_d}^2 \tag{178}$$

at tree level [15]. After electroweak symmetry breaking, both the fields H_u^0 and H_d^0 must develop v.e.v.'s, in order to give masses to all the quarks and leptons:

$$< H_u^0 >= v_u, \; < H_d^0 >= v_d. \tag{179}$$

Comparing with the Standard Model, we have

$$v^2 = v_u^2 + v_d^2 = \frac{2m_Z^2}{(g_2^2 + g_1^2)}. \tag{180}$$

Conventionally, one defines also the $\tan\beta$ parameter:

$$\tan\beta = \frac{v_u}{v_d} : \; 0 < \beta < \frac{\pi}{2}. \tag{181}$$

At the minimun of the potential

$$\frac{\partial V}{\partial H_u^0} = \frac{\partial V}{\partial H_d^0} = 0, \tag{182}$$

giving the two relations

$$\begin{aligned} |\mu|^2 + m_{H_u}^2 &= b\tan\beta - \frac{m_Z}{2}\cos^2\beta, \\ |\mu|^2 + m_{H_d}^2 &= b\cot\beta + \frac{m_Z}{2}\cos^2\beta. \end{aligned} \tag{183}$$

These expressions are important because they relate a measurable quantity, m_Z, to the soft parameters. We note that some amount of fine-tuning would be required if the soft parameters were much larger than m_Z. We note also that the vacuum conditions (183) do not depend on the phase of μ.

[15] As we shall see shortly, radiative corrections to the effective potential play important roles.

The two complex Higgs doublets of the MSSM have a total of 8 degrees of freedom. However, the Higgs mechanism for electroweak breaking uses 3 degrees of freedom to give longitudinal polarization states, and hence masses, to the two W bosons and to the Z boson. Therefore, five physical Higgs bosons remain in the spectrum. Of these, two are neutral Higgs bosons that are even under the CP transformation, called $h^0$and H^0. In addition, there is one neutral Higgs boson that is odd under CP, called A^0. The final two Higgs bosons are charged, the $H^\pm$.

At tree level, the masses of the supersymetric Higgs bosons are:

$$m^2_{h^0,H^0} = \frac{1}{2}\left(m^2_{A^0} + m^2_Z \mp \sqrt{(m^2_{A^0} + m^2_Z)^2 - 4m^2_{A^0} m^2_Z \cos^2 2\beta}\right), \tag{184}$$

$$m^2_{A^0} = \frac{2b}{\sin 2\beta}, \tag{185}$$

$$m^2_{H^\pm} = m^2_{A^0} + m^2_W, \tag{186}$$

and the mass of the h^0 is bounded from above by:

$$m_{h^0} < |\cos 2\beta| m_Z. \tag{187}$$

This upper limit on m_{h^0} may be traced to the fact that the quartic Higgs coupling λ is fixed in the MSSM, being equal to the square of the electroweak gauge coupling (up to numerical factors). This means that λ and hence m_{h^0} cannot be very large.

However, the above relations are valid only at tree level, and the masses of Higgs scalars have one-loop radiative corrections that are not negligible [88]. The most important corrections for m_h are those due to the top quark and squark:

$$\Delta m^2_h = \frac{3m^4_t}{4\pi^2 v^2} \ln\left(\frac{m_{\tilde{t}_1} m_{\tilde{t}_2}}{m^2_t}\right) + \frac{3m^4_t}{8\pi^2 v^2} \mathrm{f}(m^2_{\tilde{t}_1}, m^2_{\tilde{t}_2}, \mu, \tan\beta), \tag{188}$$

where $m_{\tilde{t}_{1,2}}$ are the physical masses of the stops (that are mixtures of $\tilde{t}_R$ and $\tilde{t}_L$), and $\mathrm{f}(m^2_{\tilde{t}_1}, m^2_{\tilde{t}_2}, \mu, \tan\beta)$ is a non-logarithmic function that can be found in [10]. The correction Δm^2_h depends quartically on the mass of the top, making it more important than the one-loop corrections due to other quarks, leptons, and gauge multiplets. After including this correction, the mass of the lightest Higgs boson may be as large as

$$m_h \lesssim 130 \text{ GeV}, \tag{189}$$

for masses of sparticles about a TeV. This is seen in Fig. 21, which shows m_h as a function of m_{A^0} for different values of $\tan\beta$. As noted, the range (189) for the mass of the lightest supersymetric Higgs boson is in perfect agreement with the indications provided by the electroweak data, as discussed in Lecture 1! This is just one of many attractive features of supersymmetry that we review here.

3.4.6 R parity and dark matter

We introduced above the superpotential (174) of the MSSM, which includes only the Yukawa interactions of the SM. However, gauge invariance, Lorentz invariance, and analyticity in the SM fields would allow us to introduce in the superpotential other terms that do not have any correspondence with the SM, and do not preserve either baryon number and/or lepton number [16]. These terms are

$$\mathcal{W}_{RPV} = \lambda_{ijk} L_i L_j E_k + \lambda'_{ijk} L_i Q_j D^c_k + \lambda''_{ijk} U^c_i D^c_j D^c_k + \mu'_i L_i H_u, \tag{190}$$

[16]The conservation of B and L in the SM is an accidental symmetry of its renormalizable interactions that is *a priori* not obligatory. As we see later in the context of Grand Unified Theories, the SM, non-renormalizable terms that violate L or B may be added to the SM Lagrangian. In the MSSM, such L- and B-violating may appear at the renormalizable level.

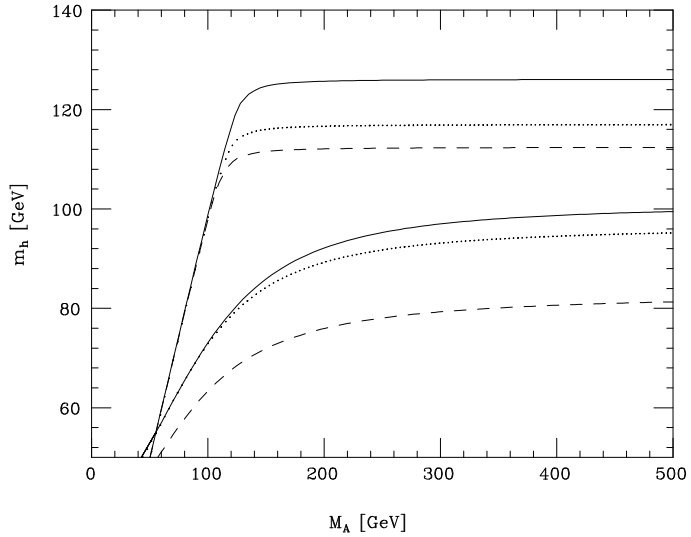


Fig. 21: The mass of the lightest supersymmetric Higgs boson as a function of m_{A^0} for different values of $\tan\beta$

where λ, λ' and λ'' are arbitrary dimensionless coupling constants, and the μ_i' are parameters with the dimension of a mass.

These parameters are subject to strong phenomenological restrictions. For example, a combination of the second and third terms would induce rapid disintegration of the proton *via* squark exchange, whereas the proton is very stable, with a lifetime exceeding $\sim 10^{33}$ years. This implies that the product of such terms must be strongly suppressed [95]:

$$|\lambda'\lambda''| \; < \; \mathcal{O}(10^{-9}). \tag{191}$$

One way to avoid all such terms is to add to the MSSM a new symmetry called R-parity, given by the following combination baryon number, lepton number, and spin S:

$$R = (-1)^{3(B-L)+2S}. \tag{192}$$

This is a multiplicatively-conserved quantum number in the SM, since all the SM particles and Higgs bosons have even R parity: $R = +1$. On the other hand, all the sparticles have odd R parity ($R = -1$).

Conservation of R parity would have important phenomenological consequences:

• The sparticles are produced in even numbers (usually two at time), for example: $\bar{p}\,p \to \tilde{q}\,\tilde{g}\,X$, $e^+\,e^- \to \tilde{\mu}^+\,\tilde{\mu}^-$.

• Each sparticle decays into another sparticle (or into an odd number of them), for example: $\tilde{q} \to q\,\tilde{g}$, $\tilde{\mu} \to \mu\,\tilde{\gamma}$.

• The lightest sparticle (LSP) must be stable, since it has $R = -1$. If it is electrically neutral, it can interact only weakly with ordinary matter, and may be a good candidate for the non-baryonic dark matter that is required by cosmology [34].

The dark matter particles should have neither electric charge nor strong interactions, otherwise they would be visible or detectable, e.g., through their binding to ordinary matter to form what would look like anomalous heavy nuclei, which have never been seen. We therefore expect any dark matter particle to have only weak interactions, in which case, if it was produced at a collider such as the LHC, it would carry energy–momentum away invisibly. Accordingly, most LHC searches for supersymmetry focus on events with missing transverse momentum, though searches for signatures of R-violating models are also considered.

The existence of a stable, weakly-interacting LSP is a very important prediction of the MSSM, but its nature and its total contribution to the density of dark matter depend on the parameters of the MSSM. One weakly-interacting candidate was the lightest sneutrino, but this has already been excluded by direct searches at LEP and by experiments searching directly for dark matter. The remaining candidate particles are the lightest neutralino χ of spin 1/2, and the gravitino of spin 3/2. As we discuss later, there are chances to detect a neutralino LSP at the LHC in events with missing energy, or directly as astrophysical dark matter. On the other hand, the interactions of the gravitino are so weak that it could not be detected as astrophysical dark matter, and could only be detected indirectly in collider experiments.

3.5 Phenomenology of supersymmetry

As we have seen, the soft supersymmetry-breaking sector of the MSSM has over a hundred parameters. This renders very difficult the interpretation of experimental constraints and (hopefully) the extraction of the experimental values of these parameters. A simplifying hypothesis is to assume *universality* at a certain scale before renormalization, leading us to the constrained MSSM (CMSSM):

• The gaugino masses are assumed to be equal at some input GUT or supergravity scale: $M_3 = M_2 = M_1 = m_{1/2}$;

• The scalar masses of squarks and sleptons are assumed to be universal at the same scale: $m_Q^2 = m_{U^c}^2 = ... = m_0^2$, as are the soft supersymmetry-breaking contributions to the Higgs masses $m_{H_u}^2 = m_{H_d}^2 = m_0^2$;

• The trilinear couplings are related by a universal coefficient A_0 to the corresponding Yukawa couplings: $a_u = A_0 y_u$, $a_d = A_0 y_d$, $a_e = A_0 y_e$.

Simplifying the MSSM to the CMSSM reduces the number of parameters from over one hundred to only 4: $m_{1/2}, m_0, A_0, \tan\beta$ and the sign of μ [the magnitude of μ is fixed by the electroweak vacuum conditions: see (183]. The CMSSM hypothesis is very practical from a phenomenological point of view, though questionable from a purely theoretical point of view. The CMSSM and the simplification of $\mathcal{L}_{soft}$ are inspired by simple supergravity models where the breaking of supersymmetry is mediated by gravity, though minimal supergravity models actually impose two additional constraints. On the other hand, generic string models often lead to different patterns of soft supersymmetry breaking.

Dropping universality for squarks or sleptons with the same quantum numbers but in different generations would lead to problems with flavour-changing neutral interactions, and Grand Unified Theories relate the soft supersymmetry-breaking masses of squarks and sleptons with different quantum numbers. However, there is no strong theoretical or phenomenological reason to postulate universality for the soft supersymmetry-breaking contributions to the Higgs masses. One may relax this assumption for the Higgs scalar masses-squared m_H^2 by assuming the same single-parameter non-universal Higgs mass parameter (the NUHM1), or by allowing the non-universal Higgs mass parameters to be different (the NUHM2).

3.6 Renormalization of the soft supersymmetry-breaking parameters

In our ignorance of the underlying mechanism of supersymmetry breaking, it is usually assumed that this occurs at some large mass scale far above a TeV, perhaps around the grand unification or Planck scale. The soft supersymmetry-breaking parameters therefore undergo significant renormalization between this input scale and the electroweak scale. Although quadratic divergences are absent from a softly-broken supersymmetric theory, it still has logarithmic divergences that may be treated using the renormalization group (RG).

At leading order in the RG, which resums the leading one-loop logarithms, the renormalizations of the soft gaugino masses M_a are the same as for the corresponding gauge couplings:

$$Q\frac{dM_a}{dQ} \;=\; \beta_a M_a, \tag{193}$$

where β_a is the standard one-loop renormalization coefficient including supersymmetric particles that is discussed in more detail in the next Lecture. As a result of (193), to leading order

$$M_a(Q) \;=\; \frac{\alpha_a(Q)}{\alpha_{GUT}} m_{1/2} \tag{194}$$

if the gauge couplings α_a and the gaugino masses are assumed to unify at the same large mass scale M_{GUT}. As a consequence of (194), one expects the gluino to be heavier than the wino: $m_{\tilde{g}}/m_{\tilde{W}} = \alpha_3/\alpha_2$ at leading order.

The soft supersymmetry-breaking scalar masses-squared m_0^2 acquire renormalizations related to the gaugino masses *via* the gauge couplings, and to the scalar masses and trilinear parameters A_λ *via* the Yukawa couplings:

$$\frac{Q dm_0^2}{dQ} \;=\; \frac{1}{16\pi^2}\left[-g_a^2 M_a^2 + \lambda^2(m_0^2 + A_\lambda^2)\right]. \tag{195}$$

The latter effect is significant for the stop squark, one of the Higgs multiplets, and possibly the other third-generation sfermions if $\tan\beta$ is large. For the other sfermions, at leading order one has

$$m_0^2(Q) \;=\; m_0^2 + C m_{1/2}^2, \tag{196}$$

where the coefficient C depends on the gauge quantum numbers of the corresponding sfermion. Consequently, one expects the squarks to be heavier than the sleptons. Specifically, in the CMSSM one finds at the electroweak scale that

$$\text{squarks} : m_{\tilde{q}}^2 \;\sim\; m_0^2 + 6m_{1/2}^2, \tag{197}$$

$$\text{left-handed sleptons: } m_{\tilde{\ell}_L}^2 \;\sim\; m_0^2 + 0.5 m_{1/2}^2, \tag{198}$$

$$\text{right-handed sleptons: } m_{\tilde{\ell}_R}^2 \;\sim\; m_0^2 + 0.15 m_{1/2}^2. \tag{199}$$

The difference between the left and right slepton masses may have implications for cosmology, as we discuss later. A small difference is also expected between the masses of the left and right squarks, but this is relatively less significant numerically.

The CKM mixing between quarks is related in the SM to off-diagonal entries in the Yukawa coupling matrix, and shows up in leading-order charged-current interactions and flavour-changing neutral current (FCNC) interactions induced at the loop level. One would expect additional FCNCs to be induced by similar loop diagrams involving squarks, which would propagate through the RGEs (195) and induce flavour-violating terms in the sfermion mass matrices. However, experiment imposes important upper limits on such additional supersymmetric flavour effects. As already discussed, these would be suppressed (though non-zero) if the soft supersymmetry-breaking scalar masses of all sfermions with the same quantum numbers were the same before renormalization. The hypothesis of Minimal Flavour Violation (MFV) is that flavour mixing of squarks and sleptons is induced only by the CKM mixing in the quark sector and the corresponding MNS mixing in the lepton sector: see the next Lecture. The MFV hypothesis requires also that the soft supersymmetry-breaking trilinear parameters A be universal for sfermions with the same quantum numbers: $A_\lambda = A_0\lambda$. However, the MFV hypothesis does permit the appearance of 6 additional phases beyond those in the CKM model for quarks: 3 phases for the different gaugino mass parameters, and 3 phases for the different A_0 coefficients [96].

Results of typical numerical calculations of these renormalization effects in the CMSSM are shown in Fig. 22. An important effect illustrated there is that the RGEs may drive $m_{H_u}^2$ negative at some low renormalization scale Q_N, thanks to the top quark Yukawa coupling appearing in (195) [17]. A negative value of $m_{H_u}^2$ would trigger electroweak symmetry breaking at a scale $\sim Q_N$. Since the negative value of

[17] The effect of the Yukawa coupling is to *increase* m_0^2 as Q increases, i.e., to *decrease* m_0^2 as Q decreases.

$m^2_{H_u}$ is due to the logarithmic renormalization by the top quark Yukawa coupling, electroweak symmetry breaking appears at a scale exponentially smaller than the input GUT or Planck scale:

$$\frac{m_W}{M_{GUT,P}} \; = \; \exp\left(-\frac{\mathcal{O}(1)}{\alpha_t}\right) : \; \alpha_t \; \equiv \; \frac{\lambda_t^2}{4\pi}. \tag{200}$$

In this way, it is possible for the electroweak scale to be generated naturally at a scale ~ 100 GeV if the top quark is heavy: $m_t \sim 60$ to 100 GeV, a realization that long predated the discovery of just such a heavy top quark.

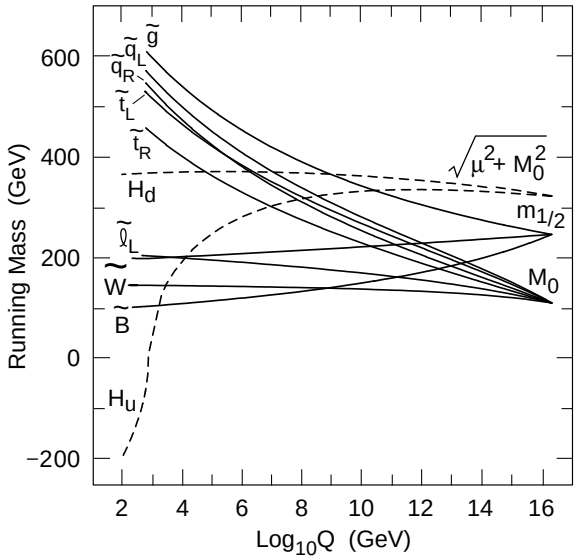


Fig. 22: Calculations of the renormalization of soft supersymmetry-breaking sparticle masses, assuming universal scalar and gaugino masses $m_0, m_{1/2}$ at the GUT scale. Note that strongly-interacting sparticles have larger physical masses at low scales, and the $m^2_{H_u}$ is driven negative, triggering electroweak symmetry breaking.

3.6.1 Sparticle masses and mixing

There are aspects of sparticle masses and mixing that are important for phenomenology, as we now discuss.

Sfermions: As we have seen, each flavour of charged lepton or quark has both left- and right-handed components $f_{L,R}$, and these have separate spin-0 boson superpartners $\tilde{f}_{L,R}$. These have different isospins $I = \frac{1}{2},\ 0$, but may mix as soon as the electroweak gauge symmetry is broken. Thus, for each flavour we should consider a 2×2 mixing matrix for the $\tilde{f}_{L,R}$, which takes the following general form:

$$M^2_{\tilde{f}} \equiv \begin{pmatrix} m^2_{\tilde{f}_{LL}} & m^2_{\tilde{f}_{LR}} \\ \\ m^2_{\tilde{f}_{LR}} & m^2_{\tilde{f}_{RR}} \end{pmatrix}. \tag{201}$$

The diagonal terms may be written in the form

$$m^2_{\tilde{f}_{LL,RR}} = m^2_{\tilde{f}_{L,R}} + m^{D^2}_{\tilde{f}_{L,R}} + m^2_f, \tag{202}$$

where m_f is the mass of the corresponding fermion, $\tilde{m}^2_{\tilde{f}_{L,R}}$ is the soft supersymmetry-breaking mass discussed in the previous section, and $m^{D^2}_{\tilde{f}_{L,R}}$ is a contribution due to the quartic D terms in the effective

potential:

$$m^{D^2}_{\tilde{f}_{L,R}} = m_Z^2 \ \cos 2\beta \ \ (I_3 + \sin^2\theta_W Q_{em}), \tag{203}$$

where the term $\propto I_3$ is non-zero only for the $\tilde{f}_L$. Finally, the off-diagonal mixing term takes the general form

$$m^2_{\tilde{f}_{L,R}} = m_f \left(A_f + \mu^{\tan\beta}_{\cot\beta} \right) \ \text{ for } \ f =^{e,\mu,\tau,d,s,b}_{u,c,t} . \tag{204}$$

It is clear that $\tilde{f}_{L,R}$ mixing is likely to be important for the $\tilde{t}$, and it may also be important for the $\tilde{b}_{L,R}$ and $\tilde{\tau}_{L,R}$ if $\tan\beta$ is large.

We also see from (202) that the diagonal entries for the $\tilde{t}_{L,R}$ would be different from those of the $\tilde{u}_{L,R}$ and $\tilde{c}_{L,R}$, even if their soft supersymmetry-breaking masses were universal, because of the m_f^2 contribution. In fact, we also expect non-universal renormalization of $m^2_{\tilde{t}_{LL,RR}}$ (and also $m^2_{\tilde{b}_{LL,RR}}$ and $m^2_{\tilde{\tau}_{LL,RR}}$ if $\tan\beta$ is large), because of Yukawa effects analogous to those discussed previously for the renormalization of the soft Higgs masses. For these reasons, the $\tilde{t}_{L,R}$ are not usually assumed to be degenerate with the other squark flavours.

Charginos: These are the supersymmetric partners of the $W^\pm$ and $H^\pm$, which mix through a 2×2 matrix

$$-\frac{1}{2} (\tilde{W}^-, \tilde{H}^-) \ M_C \ \begin{pmatrix} \tilde{W}^+ \\ \tilde{H}^+ \end{pmatrix} \ + \ \text{herm.conj.} \tag{205}$$

where

$$M_C \equiv \begin{pmatrix} M_2 & \sqrt{2} m_W \sin\beta \\ \sqrt{2} m_W \cos\beta & \mu \end{pmatrix} . \tag{206}$$

Here M_2 is the unmixed $SU(2)$ gaugino mass and μ is the Higgs mixing parameter introduced previously.

Neutralinos: These are characterized by a 4×4 mass mixing matrix [34], which takes the following form in the $(\tilde{W}^3, \tilde{B}, \tilde{H}^0_2, \tilde{H}^0_1)$ basis :

$$m_N = \begin{pmatrix} M_2 & 0 & \frac{-g_2 v_2}{\sqrt{2}} & \frac{g_2 v_1}{\sqrt{2}} \\ 0 & M_1 & \frac{g' v_2}{\sqrt{2}} & \frac{-g' v_1}{\sqrt{2}} \\ \frac{-g_2 v_2}{\sqrt{2}} & \frac{g' v_2}{\sqrt{2}} & 0 & \mu \\ \frac{g_2 v_1}{\sqrt{2}} & \frac{-g' v_1}{\sqrt{2}} & \mu & 0 \end{pmatrix} \tag{207}$$

Note that this has a structure similar to M_C (206), but with its entries replaced by 2×2 submatrices. As has already been mentioned, one often assumes that the $SU(2)$ and $U(1)$ gaugino masses $M_{1,2}$ are universal at the GUT or supergravity scale, so that

$$M_1 \simeq M_2 \ \frac{\alpha_1}{\alpha_2}, \tag{208}$$

so the relevant parameters of (207) are generally taken to be $M_2 = (\alpha_2/\alpha_{GUT}) m_{1/2}$, μ and $\tan\beta$.

In the limit $M_2 \to 0$, the lightest neutralino χ would be approximately a photino, and it would be approximately a higgsino in the limit $\mu \to 0$. However, these idealized limits are excluded by unsuccessful LEP and other searches for neutralinos and charginos. Possibilities that persist are that χ be approximately a Bino, $\tilde{B}$, or that it has a substantial higgsino component.

3.7 Constraints on the MSSM

Most of the current constraints on possible physics beyond the SM are negative and, specifically, no sparticle has ever been detected. The concordance with the SM predictions means that, in general, one can only set lower limits on the possible masses of supersymmetric particles. However, there are two observational indications of physics beyond the SM that may, in the supersymmetric context, be used for setting *upper* limits of the masses of the supersymmetric particles. As discussed earlier, these two hints for new physics are the anomalous magnetic moment of the muon, $g_\mu - 2$, which seems to disagree with the prediction of the SM (at least if this is calculated using low-energy e^+e^- data as an input), and the density of cold dark matter Ω_{CDM}. However, these discrepancies may be explained either with supersymmetry or with other possible extensions of the SM, so their interpretations require special care. Nevertheless, these may be regarded as additional phenomenological motivations for supersymmetry, in addition to the more theoretical motivations described in the beginning of this section, such as the naturalness of the hierarchy of mass scales in physics, grand unification, string theory, etc. Therefore, in addition to considering the more direct searches for supersymmetry, it is also natural to ask what $g_\mu - 2$ and Ω_{CDM} may imply for the parameters of supersymmetric models. Figure 23 compiles the impacts of various constraints on supersymmetry, assuming that the soft supersymmetry-breaking contributions $m_{1/2}, m_0$ to the different scalars and gauginos are each universal at the GUT scale (the scenario called the CMSSM), and that the lightest sparticle is the lightest neutralino χ.

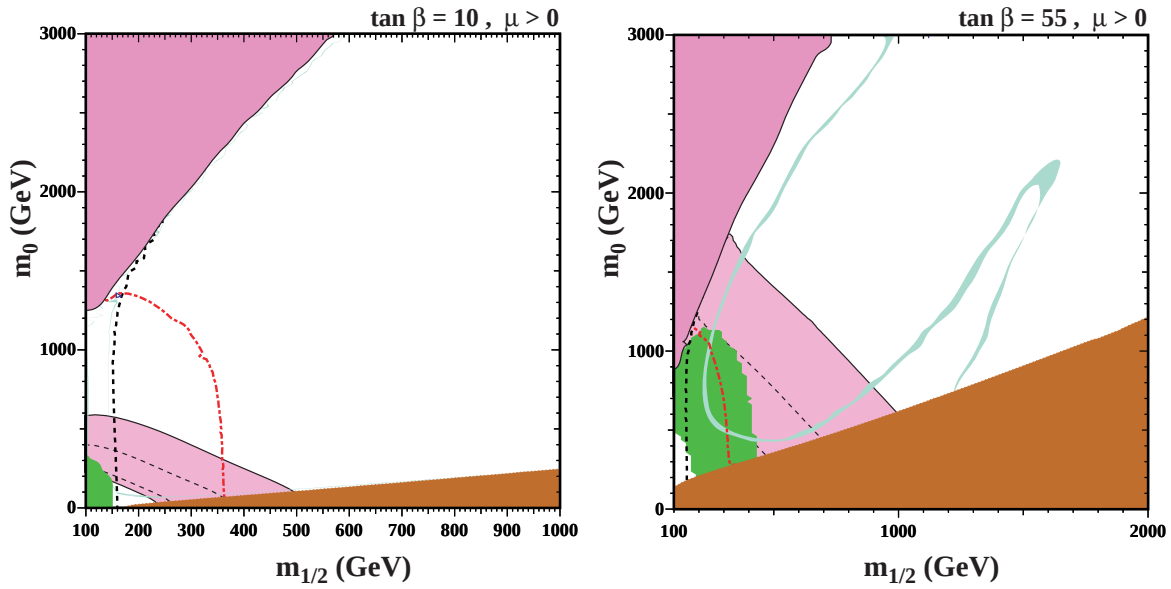


Fig. 23: The CMSSM $(m_{1/2}, m_0)$ planes for (a) $\tan\beta = 10$ and (b) $\tan\beta = 55$, assuming $\mu > 0$, $A_0 = 0$, $m_t = 173.1$ GeV and $m_b(m_b)^{\overline{MS}}_{SM} = 4.25$ GeV. The near-vertical (red) dot-dashed lines are the contours for $m_h = 114$ GeV, and the near-vertical (black) dashed line is the contour $m_{\chi^\pm} = 104$ GeV. Also shown by the dot-dashed curve in the lower left is the region excluded by the LEP bound $m_{\tilde{e}} > 99$ GeV. The medium (dark green) shaded region is excluded by $b \to s\gamma$, and the light (turquoise) shaded area is the cosmologically preferred region. In the dark (brick red) shaded region, the LSP is the charged $\tilde{\tau}_1$. The region allowed by the measurement of $g_\mu - 2$ at the 2-σ level, assuming the e^+e^- calculation of the Standard Model contribution, is shaded (pink) and bounded by solid black lines, with dashed lines indicating the 1-σ ranges (updated from [98]).

Experiments at LEP and the Tevatron collider, in particular, have made direct searches for supersymmetry using the missing-energy-momentum signature. LEP established lower limits ~ 100 GeV on the masses of many charged sparticles without strong interactions, such as sleptons and charginos. The Tevatron collider has established the best lower limits on the masses of squarks and gluinos, ~ 400 GeV. In view of the greater renormalization of the squark and gluino masses than for charginos and sleptons,

see (194) and (199), these two sets of limits are quite complementary.

Another important constraint is provided by the LEP lower limit on the Higgs mass: $m_H > 114.4$ GeV [20]. This holds in the Standard Model, for the lightest Higgs boson h in the general MSSM for $\tan\beta \lesssim 8$, and almost always in the CMSSM for all $\tan\beta$, at least as long as CP is conserved [18]. Since m_h is sensitive to sparticle masses, particularly $m_{\tilde{t}}$ *via* the loop corrections (188), the Higgs limit also imposes important constraints on the soft supersymmetry-breaking CMSSM parameters, principally $m_{1/2}$ [98], as seen in Fig. 23.

Important constraints are imposed on the CMSSM parameter space by flavour physics, specifically the agreement with data of the SM prediction for the decay $b \to s\gamma$, as well as the upper limit on the decay $B_s \to \mu^+\mu^-$, which is important at large $\tan\beta$ in particular.

We see in Fig. 23 that narrow strips of the $(m_{1/2}, m_0)$ planes are compatible [98] with the range of the astrophysical cold dark matter density favoured by WMAP and other experiments. However, these strips vary with $\tan\beta$ and A_0. In fact, foliation by these WMAP strips covers large fractions of the $(m_{1/2}, m_0)$ plane as $\tan\beta$ and A_0 are varied. Away from these narrow strips, the relic neutralino density exceeds the WMAP range over most of the $(m_{1/2}, m_0)$ planes shown in Fig. 23. In its left panel, the relic density is reduced into the WMAP range only in the shaded strip at $m_0 \sim 100$ GeV that extends to $m_{1/2} \sim 900$ GeV. This reduction is brought about by co-annihilations between the LSP χ (which is mainly a Bino) and sleptons that are only slightly heavier, most notably the lighter stau and the right selectron and smuon, which are significantly lighter than the left sleptons, as discussed earlier. In the right panel of Fig. 23 for $\tan\beta = 50$, this co-annihilation strip moves to larger m_0. Also, it is extended to larger $m_{1/2}$, as a result of a reduction in the relic density due to rapid $\chi - \chi$ annihilations though direct-channel heavy Higgs (H, A) states. In addition to these visible WMAP regions, there is in principle another allowed strip at very large values of m_0, called the focus-point region, where the LSP becomes relatively light and acquires a substantial higgsino component, favouring annihilation *via* W^+W^- final states.

Finally, also shown in the two panels of Fig. 23 are the regions favoured by the supersymmetric interpretation of the discrepancy (120) between the experimental measurement of $g_\mu - 2$ and the value calculated in the SM using low-energy e^+e^- data [98]. The favoured regions are displayed as bands corresponding to $\pm 2\sigma$. We see that they can be used to set *upper* limits on the sparticle masses! In particular, $g_\mu - 2$ disfavours the focus-point region, where m_0 is so large that the supersymmetric contribution to $g_\mu - 2$ is negligible, and also the region at large $\tan\beta$ and large $m_{1/2}$ where the neutralinos may annihilate rapidly though direct-channel heavy-Higgs states.

3.8 Frequentist analysis of the supersymmetric parameter space

In a recent paper [99] the likely range of parameters of the CMSSM and NUHM1 has been estimated using a frequentist approach, by building a χ^2 likelihood function with contributions from the various relevant observables, including precision electroweak physics, $g_\mu - 2$, the lower limit on the lightest Higgs boson mass (taking into taking into account the theoretical uncertainty in the `FeynHiggs` calculation of M_h [100]), the experimental measurement of BR($b \to s\gamma$)(which agrees with the SM), the experimental upper limit on BR($B_s \to \mu^+\mu^-$), and Ω_{CDM}. This frequentist analysis used a Markov chain Monte Carlo technique to sample thoroughly the $(m_0, m_{1/2})$ plane up to masses of several TeV, including the focus-point and rapid-annihilation regions, for a wide range of values of A_0 and $\tan\beta$.

We display in Fig. 24 the $\Delta\chi^2$ functions in the $(m_0, m_{1/2})$ planes for the CMSSM (left plot) and for the NUHM1 (right plot). The parameters of the best-fit CMSSM point are $m_0 = 60$ GeV, $m_{1/2} = 310$ GeV, $A_0 = 130$ GeV, $\tan\beta = 11$, and $\mu = 400$ GeV (corresponding nominally to $M_h = 114.2$ GeV and an overall $\chi^2 = 20.6$ for 19 d.o.f. with a probability of 36%), which are very

[18] The lower bound on the lightest MSSM Higgs boson may be relaxed significantly if CP violation feeds into the MSSM Higgs sector [97].

close to the ones previously reported in Ref. [101]. The corresponding parameters of the best-fit NUHM1 point are $m_0 = 150$ GeV, $m_{1/2} = 270$ GeV, $A_0 = -1300$ GeV, $\tan\beta = 11$, and $m^2_{h_1} = m^2_{h_2} = -1.2 \times 10^6$ GeV^2 or, equivalently, $\mu = 1140$ GeV, yielding $\chi^2 = 18.4$ (corresponding to a similar fit probability to the CMSSM) and $M_h = 120.7$ GeV. The similarities between the best-fit values of m_0, $m_{1/2}$ and $\tan\beta$ in the CMSSM and the NUHM1 suggest that the model frameworks used are reasonably stable: if they had been very different, one might well have wondered what would be the effect of introducing additional parameters, as in the NUHM2 with two non-universality parameters in the Higgs sector.

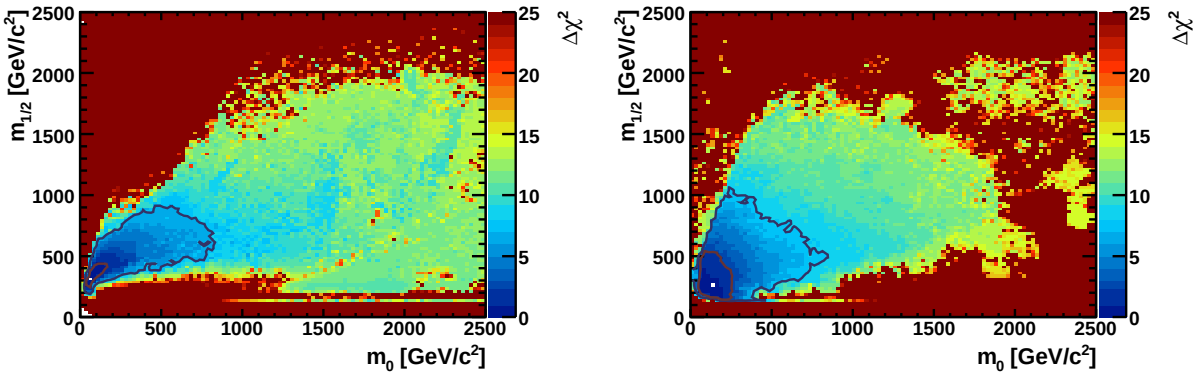


Fig. 24: The $\Delta\chi^2$ functions in the $(m_0, m_{1/2})$ planes for the CMSSM (left plot) and for the NUHM1 (right plot), as found in frequentist analyses of the parameter spaces. We see that the co-annihilation regions at low m_0 and $m_{1/2}$ are favoured in both cases [101].

These best-fit points are both in the co-annihilation region of the $(m_0, m_{1/2})$ plane, as can be seen in Fig. 24. The C.L. contours extend to slightly larger values of m_0 in the CMSSM, while they extend to slightly larger values of $m_{1/2}$ in the NUHM1, as was already shown in Ref. [101] for the 68% and 95% C.L. contours. However, the qualitative features of the $\Delta\chi^2$ contours are quite similar in the two models, indicating that the preference for small m_0 and $m_{1/2}$ are quite stable and do not depend on details of the Higgs sector. We recall that it was found in Ref. [101] that the focus-point region was disfavoured at beyond the 95% C.L. in both the CMSSM and the NUHM1. We see in Fig. 24 that this region is disfavoured at the level $\Delta\chi^2 \sim 8$ in the CMSSM and > 9 in the NUHM1.

The favoured values of the particle masses in both models are such that there are good prospects for detecting supersymmetric particles in CMS [28] and ATLAS [29] even in the early phase of the LHC running with reduced centre-of-mass energy and limited luminosity, as seen in Fig. 25. The best-fit points and most of the 68% confidence level regions are within the region of the $(m_0, m_{1/2})$ plane that could be explored with 100/pb of data at 14 TeV in the centre of mass, and hence perhaps with 200/fb of data at 10 TeV [19]. Almost all the 95% confidence level regions would be accessible to the LHC with 1/fb of data at 14 TeV. As seen in Fig. 25, in substantial parts of these regions there are good prospects for detecting $\tilde{q} \to q\ell^+\ell^-\chi$ decays, which are potentially useful for measuring sparticle mass parameters, and the lightest supersymmetric Higgs boson may also be detectable in $\tilde{q}$ decays.

The best-fit spectra in the CMSSM and NUHM1 are shown in Fig. 26: they are relatively similar, though the heavier Higgs bosons, the gluinos, and the squarks may be somewhat heavier in the CMSSM, whereas the heavier charginos and neutralinos may be heavier in the NUHM1 [101]. There are considerable uncertainties in these spectra, as seen in Fig. 27 [99]. However, in general there are strong

[19] The comparisons are made with experimental simulations for $\tan\beta = 10$ and $A_0 = 0$, whereas the frequentist analysis sampled all values of $\tan\beta$ and A_0. As it happens, the preferred values of $\tan\beta$ in both the CMSSM and the NUHM1 are quite close to 10: the value of A_0 is relatively unimportant for the experimental analysis.

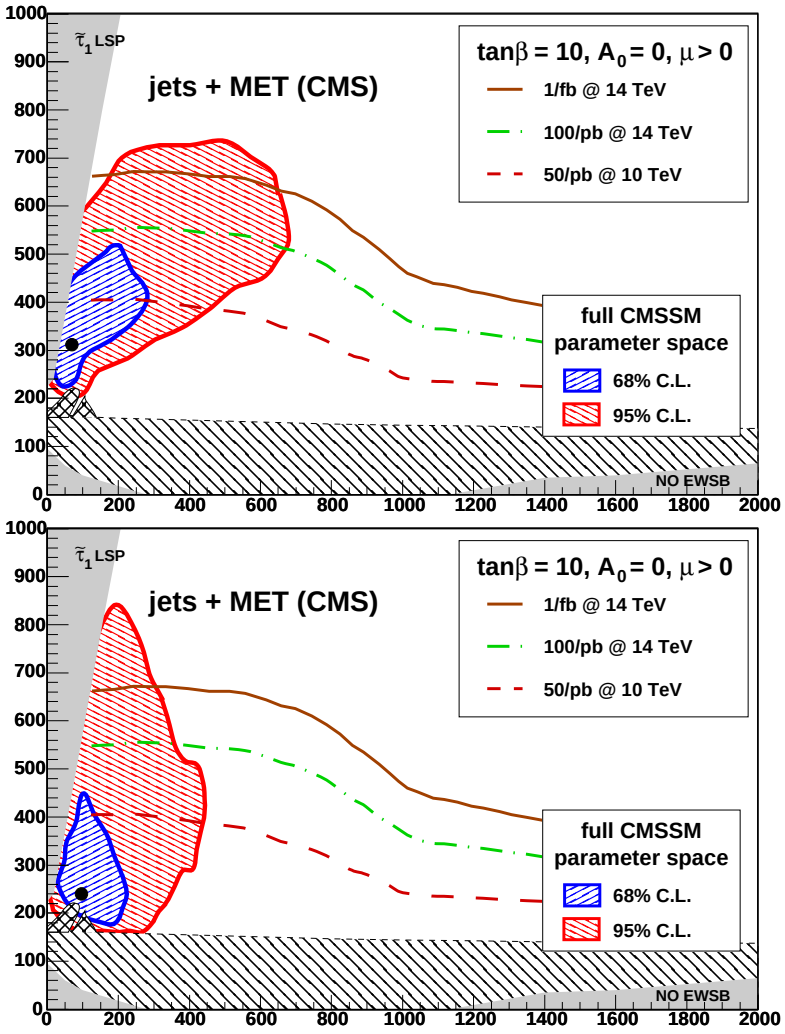


Fig. 25: The $(m_0, m_{1/2})$ planes in the CMSSM (upper) and the NUHM1 (lower) for $\tan\beta = 10$ and $A_0 = 0$. The dark shaded areas at low m_0 and high $m_{1/2}$ are excluded due to a scalar tau LSP, the light shaded areas at low $m_{1/2}$ do not exhibit electroweak symmetry breaking. The nearly horizontal line at $m_{1/2} \approx 160$ GeV in the lower panel has $m_{\tilde{\chi}_1^\pm} = 103$ GeV, and the area below is excluded by LEP searches. Just above this contour at low m_0 in the lower panel is the region that is excluded by trilepton searches at the Tevatron. Shown in each plot is the best-fit point [101], indicated by a star, and the 68 (95)% C.L. contours from the fit as dark grey/blue (light grey/red) overlays, scanned over all $\tan\beta$ and A_0 values. The plots also show some $5\,\sigma$ discovery contours for CMS [28] with 1 fb^{-1} at 14 TeV, 100 pb^{-1} at 14 TeV and 50 pb^{-1} at 10 TeV centre-of-mass energy [101].

correlations between the different sparticle masses, as exemplified in Fig. 28, though the correlation is weaker, e.g., for the lighter stau and the LSP in the NUHM1 [20].

Finally, a result from this frequentist analysis that also concerns LHC physics, but away from the high-energy frontier. We see in Fig. 29 that the branching ratio for $B_s \to \mu^+\mu^-$ may well exceed considerably its value in the SM, particularly at large $\tan\beta$. This is true to some extent in the CMSSM, and even more so in the NUHM1. Particularly in the latter case, this decay might perhaps be accessible to the LHCb experiment during initial LHC running. Therefore, there may be important competition for ATLAS and CMS in their quest to discover supersymmetry!

[20] This reflects the possible appearance of rapid direct-channel annihilations also at low $m_{1/2}$ and low $\tan\beta$, allowing an escape from the co-annihilation region where $m_\chi \sim m_{\tilde{\tau}_1}$.

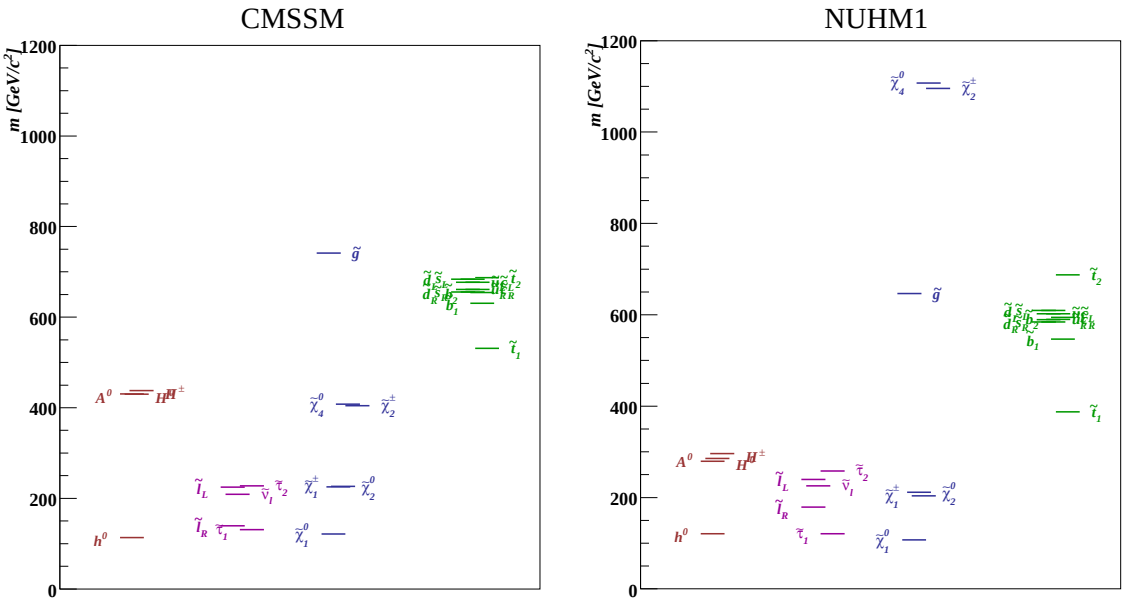


Fig. 26: The spectra at the best-fit points: left — in the CMSSM with $m_{1/2} = 311$ GeV, $m_0 = 63$ GeV, $A_0 = 243$ GeV, $\tan\beta = 11.0$, and right — in the NUHM1 with $m_{1/2} = 265$ GeV, $m_0 = 143$ GeV, $A_0 = -1235$ GeV, $\tan\beta = 10.4$, and $\mu = 1110$ GeV [101].

4 Further beyond: GUTs, string theory and extra dimensions

4.1 Grand unification

Gauge theories, particularly non-Abelian Yang–Mills theories, are the only suitable framework for describing interactions in particle physics. In the SM, there are three different gauge groups $SU(3)_C$, $SU(2)_L$, and $U(1)_Y$, and correspondingly there are three different couplings. It is logical to look for a single, more powerful non-Abelian grand unified gauge group with a single coupling g_{GUT} that would enable us to unify the three couplings, and might provide interesting relations between the other different SM parameters such as Yukawa couplings and hence fermion masses [21]. As a first approximation, we assume that the effects of the gravitational interaction are negligible, which is generally true if the grand unification scale M_{GUT} is significantly smaller that the Planck mass. As we see later, it turns out that typical estimations, based on extrapolation to very high energies of the known physics of the SM [102], give a grand unification scale of the order of 10^{16} GeV, which is about a thousand times smaller than the Planck scale $M_{Pl} = \mathcal{O}(10^{19})$ GeV.

Postulating a single group to describe all the interactions of particle physics also implies new relations between the matter particles themselves, as well as new gauge bosons. Specifically, if the symmetry changes then the representations, and hence the organization of the particles into multiplets, also change. There are some hints for this in low-energy physics, such as charge quantization and the correlation of fractional electrical charges with colour charges, and the cancellation of anomalies between the leptons and the quarks that also lead us to anticipate an organization simpler than the SM.

Clearly, one must recover the Standard Model at low energy, implying that in these Grand Unified Theories (GUTs) one must also study the breaking of the GUT group $G \to SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$.

This section begins with a presentation of the renormalization-group evolution equations of the three SM gauge couplings and studies their possible unification at some GUT scale. Subsequently, some specific examples of GUTs are discussed, notably the prototype based on the group $SU(5)$, which makes

[21] In this section, we denote the couplings by g_1 for the $U(1)$ subgroup, g_2 for $SU(2)$, and g_3 for $SU(3)$, which have the appropriate normalizations for grand unification [see later].

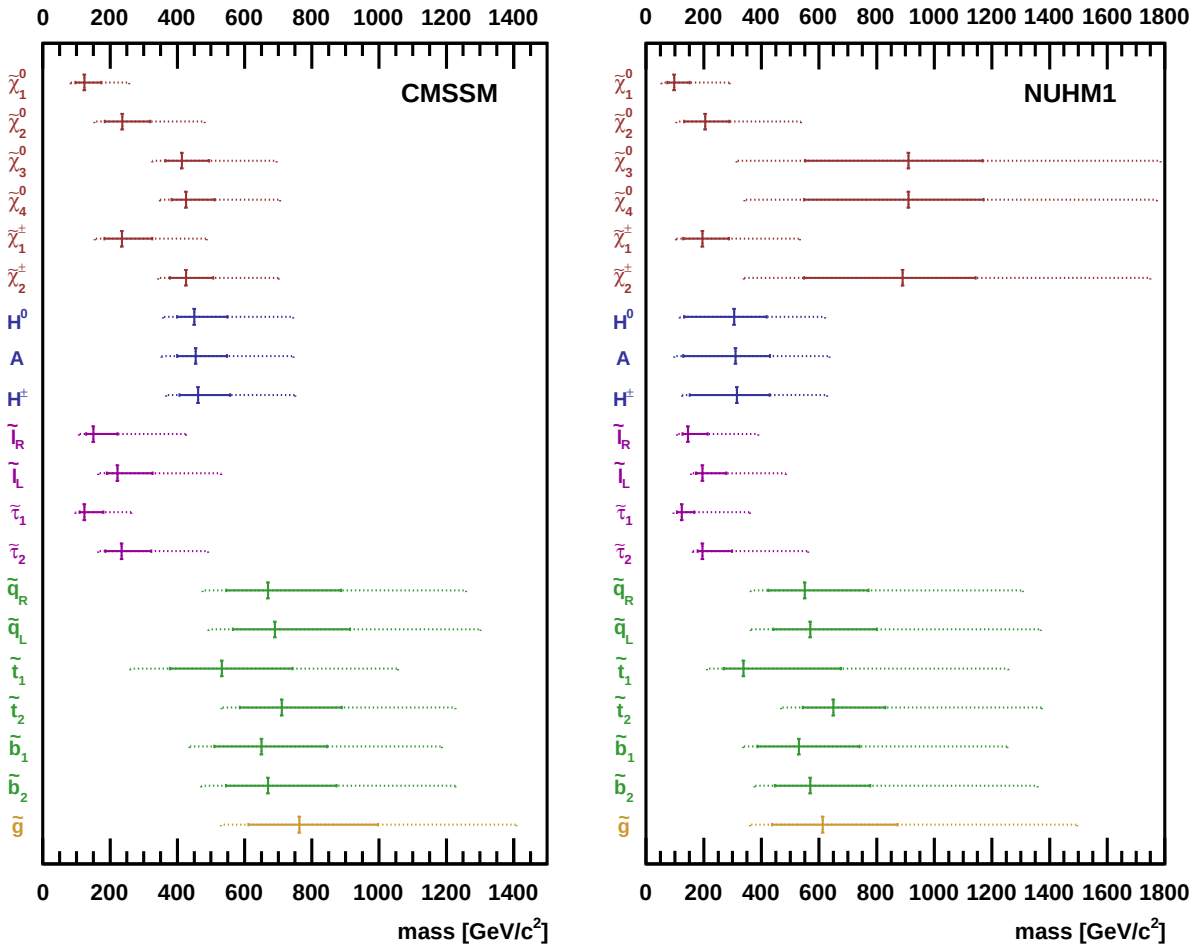


Fig. 27: Spectra in the CMSSM (left) and the NUHM1 (right). The vertical solid lines indicate the best-fit values, the horizontal solid lines are the 68% C.L. ranges, and the horizontal dashed lines are the 95% C.L. ranges for the indicated mass parameters [99].

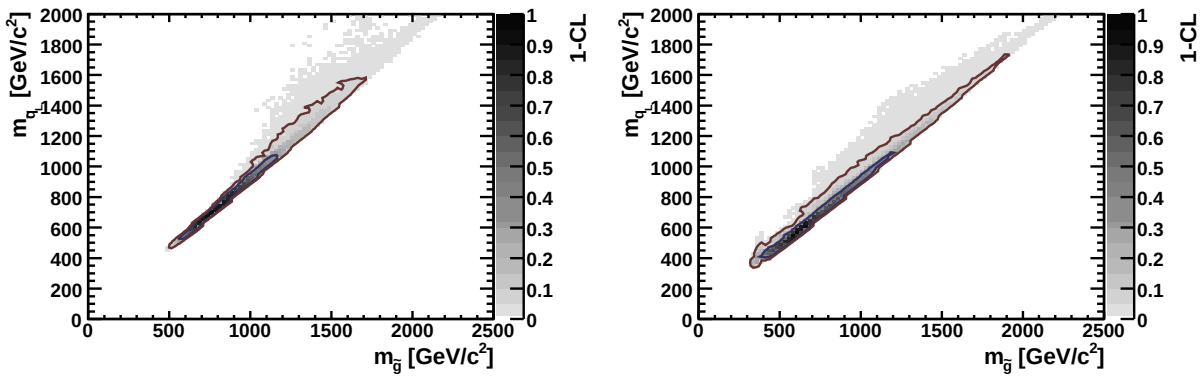


Fig. 28: The correlations between the gluino mass, $m_{\tilde{g}}$, and the masses of the the left-handed partners of the five light squark flavours, $m_{\tilde{q}_L}$, are shown in the CMSSM (left panel) and in the NUHM1 (right panel) [99].

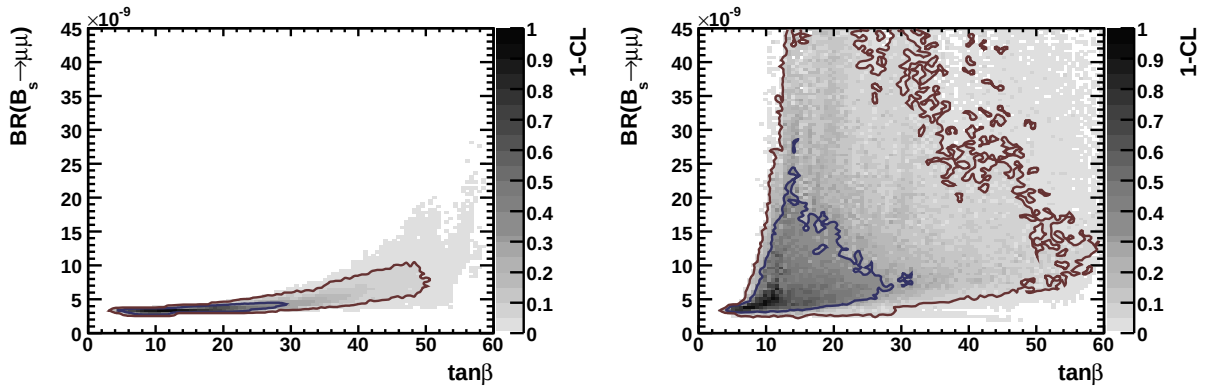


Fig. 29: The correlation between the branching ratio for $B_s \to \mu^+\mu^-$ and $\tan\beta$ in the CMSSM (left panel) and in the NUHM1 (right panel) [99].

possible a simple discussion of many properties of GUTs. This is followed by a short discussion of typical predictions of these models, such as the decay of the proton and the relations between the masses of the quarks and leptons. We finish by discussing some of the advantages, problems, and perspectives of GUT models.

4.1.1 The evolution equations for gauge couplings

The first apparent obstacle to the philosophy of grand unification is the fact that the strong coupling strength $\alpha_3 = g_3^2/4\pi$ is much stronger than the electroweak couplings at present-day energies: $\alpha_3 \gg \alpha_2, \alpha_1$. However, the strong coupling is asymptotically free [9]:

$$\alpha_3(Q) \simeq \frac{12\pi}{(33 - 2N_q)\ln(Q^2/\Lambda_3^2)} + \ldots, \tag{209}$$

where N_q is the number of quarks, $\Lambda_3 \simeq$ few hundred MeV is an intrinsic scale of the strong interactions, and the dots in (209) represent higher-loop corrections to the leading one-loop behaviour shown. The other SM gauge couplings also exhibit logarithmic violations analogous to (209). For example, the fine-structure constant $\alpha_{em} = 1/137.035999084(51)$ is renormalized to effective value of $\alpha_{em}(m_Z) \sim 1/128$ at the Z mass scale. The renormalization-group evolution for the $SU(2)$ gauge coupling corresponding to (209) is

$$\alpha_2(Q) \simeq \frac{12\pi}{(22 - 2N_q - N_{H/2})\ln(Q^2/\Lambda_2^2)} + \ldots, \tag{210}$$

where we have assumed equal numbers of quarks and leptons, and N_H is the number of Higgs doublets. Taking the inverses of (209) and (210), and then taking their difference, we find

$$\frac{1}{\alpha_3(Q)} - \frac{1}{\alpha_2(Q)} = \left(\frac{11 + N_{H/2}}{12\pi}\right)\ln\left(\frac{Q^2}{m_X^2}\right) + \ldots. \tag{211}$$

Note that we have absorbed the scales Λ_3 and Λ_2 into a single grand unification scale M_X where $\alpha_3 = \alpha_2$.

Evaluating (211) when $Q = \mathcal{O}(M_W)$, where $\alpha_3 \gg \alpha_2 = 0(\alpha_{em})$, we derive the characteristic feature [102]

$$\frac{m_{GUT}}{m_W} = \exp\left(\mathcal{O}\left(\frac{1}{\alpha_{em}}\right)\right), \tag{212}$$

i.e., the grand unification scale is exponentially large. As we see in more detail later, in most GUTs there are new interactions mediated by bosons weighing $\mathcal{O}(m_X)$ that cause protons to decay with a lifetime αm_X^4. In order for the proton lifetime to exceed the experimental limit, we need $m_X \gtrsim 10^{14}$ GeV and hence $\alpha_{em} \lesssim 1/120$ in (212) [103]. On the other hand, if the neglect of gravity is to be consistent, we need $m_X \lesssim 10^{19}$ GeV and hence $\alpha_{em} \gtrsim 1/170$ in (212) [103]. The fact that the measured value of the fine-structure constant α_{em} lies in this allowed range may be another hint favouring the GUT philosophy.

Further empirical evidence for grand unification is provided by the prediction it makes for the neutral electroweak mixing angle [102]. Calculating the renormalization of the electroweak couplings, one finds

$$\sin^2\theta_W = \frac{\alpha_{em}(m_W)}{\alpha_2(m_W)} \simeq \frac{3}{8}\left[1 - \frac{\alpha_{em}}{4\pi}\frac{110}{9}\ln\frac{m_X^2}{m_W^2}\right], \tag{213}$$

which can be evaluated to yield $\sin^2\theta_W \sim 0.210$ to 0.220, if there are only SM particles with masses $\lesssim m_X$ [102]. This is to be compared with the experimental value $\sin^2\theta_W = 0.23120 \pm 0.00015$ in the $\overline{\rm MS}$ renormalization scheme. Considering that $\sin^2\theta_W$ could *a priori* have had any value between 0 and 1, this is an impressive qualitative success. The small discrepancy can be removed by adding some extra particles, such as the supersymmetric particles in the MSSM.

To see this explicitly, we may write

$$\sin^2\theta(m_Z) \;=\; \frac{g'^2}{g_2^2 + g'^2} \;=\; \frac{3}{5}\frac{g_1^2(m_Z)}{g_2^2(m_Z) + \frac{3}{5}g_1^2(m_Z)}, \tag{214}$$

where g_1 is defined in such a way that its quadratic Casimir coefficient, summed over all the particles in a single generation, is the same as for g_2 and g_3, which is the appropriate normalization within a GUT. Using the one-loop RGEs, we can then write

$$\sin^2\theta(m_Z) \;=\; \frac{1}{1+8x}\left[3x + \frac{\alpha_{em}(m_Z)}{\alpha_3(m_Z)}\right] \;=\; \frac{1}{5}\left(\frac{b_2 - b_3}{b_1 - b_2}\right), \tag{215}$$

where the b_i are the one-loop coefficients in the RGEs for the different SM couplings. Their values in the SM (on the left) and the MSSM (on the right) are:

$$\frac{4}{3}N_G - 11 \;\leftarrow\; b_3 \;\rightarrow\; 2N_G - 9 \;=\; -3 \tag{216}$$

$$\frac{1}{6}N_H + \frac{4}{3}N_G - \frac{22}{3} \;\leftarrow\; b_2 \;\rightarrow\; \frac{1}{2}N_H + 2N_G - 6 \;=\; +1 \tag{217}$$

$$\frac{1}{10}N_H + \frac{4}{3}N_G \;\leftarrow\; b_1 \;\rightarrow\; \frac{3}{10}N_H + 2N_G \;=\; \frac{33}{5} \tag{218}$$

$$\frac{23}{218} \;=\; 0.1055 \;\leftarrow\; x \;\rightarrow\; \frac{1}{7}. \tag{219}$$

Experimentally, using $\alpha_{em}(m_Z) = 1/128, \alpha_3 = 0.119 \pm 0.003, \sin^2\theta_W(m_Z) = 0.2315$, we find

$$x \;=\; \frac{1}{6.92 \pm 0.07}, \tag{220}$$

in striking agreement with the MSSM prediction in (219)!

Another qualitative success is the prediction of the b quark mass [104, 105]. In many GUTs, such as the minimal $SU(5)$ model, discussed shortly, the b quark and the τ lepton have equal Yukawa couplings when renormalized at the GUT sale. The renormalization group then tells us that

$$\frac{m_b}{m_\tau} \simeq \left[\ln\left(\frac{m_b^2}{m_X^2}\right)\right]^{\frac{12}{33-2N_q}}. \tag{221}$$

Using $m_\tau = 1.78$ GeV, we predict that $m_b \simeq 5$ GeV, in agreement with experiment. Happily, this prediction remains successful if the effects of supersymmetric particles are included in the renormalization-group calculations [106].

To examine the GUT predictions for $\sin^2\theta_W$ etc. in more detail, one needs to study the renormalization-group equations beyond the leading one-loop order. Through two loops, one finds that

$$Q\ \frac{\partial \alpha_i(Q)}{\partial Q} = -\frac{1}{2\pi}\ \left(b_i + \frac{b_{ij}}{4\pi}\ \alpha_j(Q)\right)\ \ [\alpha_i(Q)]^2\,, \tag{222}$$

where the b_i receive the one-loop contributions

$$b_i = \begin{pmatrix} 0 \\ -\frac{22}{3} \\ -11 \end{pmatrix} + N_g \begin{pmatrix} \frac{4}{3} \\ \frac{4}{3} \\ \frac{4}{3} \end{pmatrix} + N_H \begin{pmatrix} \frac{1}{10} \\ \frac{1}{6} \\ 0 \end{pmatrix} \tag{223}$$

from gauge bosons, N_g matter generations and N_H Higgs doublets, respectively, and at two loops

$$b_{ij} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\frac{136}{3} & 0 \\ 0 & 0 & -102 \end{pmatrix} + N_g \begin{pmatrix} \frac{19}{15} & \frac{3}{5} & \frac{44}{15} \\ \frac{1}{5} & \frac{49}{3} & 4 \\ \frac{4}{30} & \frac{3}{2} & \frac{76}{3} \end{pmatrix} + N_H \begin{pmatrix} \frac{9}{50} & \frac{9}{10} & 0 \\ \frac{3}{10} & \frac{13}{6} & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{224}$$

It is important to note that these coefficients are all independent of any specific GUT model, depending only on the light particles contributing to the renormalization.

Including supersymmetric particles as in the MSSM, one finds [107]

$$b_i = \begin{pmatrix} 0 \\ -6 \\ -9 \end{pmatrix} + N_g \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} + N_H \begin{pmatrix} \frac{3}{10} \\ frac12 \\ 0 \end{pmatrix}, \tag{225}$$

and

$$b_{ij} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -24 & 0 \\ 0 & 0 & -54 \end{pmatrix} + N_g \begin{pmatrix} \frac{38}{15} & \frac{6}{5} & \frac{88}{15} \\ \frac{2}{5} & 14 & 8 \\ \frac{11}{5} & 3 & \frac{68}{3} \end{pmatrix} + N_H \begin{pmatrix} \frac{9}{50} & \frac{9}{10} & 0 \\ \frac{3}{10} & \frac{7}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{226}$$

again independent of any specific supersymmetric GUT.

One can use these two-loop equations to make detailed calculations of $\sin^2\theta_W$ in different GUTs. These confirm that non-supersymmetric models are not consistent with the determinations of the gauge couplings from LEP and elsewhere [108]. Previously, we argued that these models predicted a wrong value for $\sin^2\theta_W$, given the experimental value of α_3. In Fig. 19(a) we see the converse, namely that extrapolating the experimental determinations of the α_i using the non-supersymmetric renormalization-group equations (223), (224) does not lead to a common value of the gauge couplings at any renormalization scale. In contrast, we see in Fig. 19(b) that extrapolation using the supersymmetric renormalization-group equations (225), (226) **does** lead to possible unification at $M_{GUT} \sim 10^{16}$ GeV [89], if the spartners of the SM particles weigh ~ 1 TeV.

Turning this success around, and assuming $\alpha_3 = \alpha_2 = \alpha_1$ at M_{GUT} with no threshold corrections at this scale, one may estimate that [109]

$$\sin^2\theta_W(M_Z)\Big|_{\overline{\rm MS}} = 0.2029 + \frac{7\alpha_{em}}{15\alpha_3} + \frac{\alpha_{em}}{20\pi}\left[-3\ln\left(\frac{m_t}{m_Z}\right) + \frac{28}{3}\ln\left(\frac{m_{\tilde{g}}}{m_Z}\right) - \frac{32}{3}\ln\left(\frac{m_{\tilde{W}}}{m_Z}\right) - \ln\left(\frac{m_A}{m_Z}\right) - 4\ln\left(\frac{\mu}{m_Z}\right) + \ldots\right]. \quad (227)$$

Setting all the sparticle masses to 1 TeV reproduces approximately the value of $\sin^2\theta_W$ observed experimentally. Can one invert this successful argument to estimate the supersymmetric particle mass scale? One can show [110] that the sparticle mass thresholds in (227) can be lumped into the parameter

$$T_{susy} \equiv |\mu|\left(\frac{m_W^2}{m_{\tilde{g}}}\right)^{14/19}\left(\frac{m_A^2}{\mu^2}\right)^{3/38}\left(\frac{m_{\tilde{W}}^2}{\mu^2}\right)^{2/19}\prod_{i=1}^{3}\left(\frac{m_{\tilde{\ell}_{Li}}^3 m_{\tilde{q}_i}^7}{m_{\tilde{\ell}_{R_i}}^2 m_{\tilde{u}_i}^5 m_{\tilde{d}_i}^3}\right)^{1/19}. \quad (228)$$

If one assumes sparticle mass universality at the GUT scale, then [110]

$$T_{susy} \simeq |\mu|\left(\frac{\alpha_2}{\alpha_3}\right)^{3/2} \simeq \frac{\mu}{7}, \quad (229)$$

approximately. The measured value of $\sin^2\theta_W$ is consistent with $T_{susy} \sim 100$ GeV to 1 TeV, roughly as expected from the hierarchy argument. However, the uncertainties are such that one cannot use this consistency to constrain T_{susy} very tightly [111]. In particular, even if one accepts the universality hypothesis, there could be important model-dependent threshold corrections around the GUT scale [109, 112].

4.1.2 Specific GUTs

What groups may be used to construct a GUT [113]?

First, suitable groups must be sufficiently large to include the SM. The latter is of rank four, i.e., there are four simultaneously-diagonalizable symmetry generators [22]: $SU(3)_C$ have two, $SU(2)_L$ one, and $U(1)_Y$ one also. It is striking that all of the diagonal generators are traceless: this is trivial for the non-Abelian groups $SU(3)_C$ and $SU(2)_L$, but non-trival for $U(1)_Y$, and a possible hint that it should be embedded in a non-Abelian GUT group. Therefore, we must first find in the Cartan classification of Lie groups a group of rank higher than or equal to four. Secondly, a GUT group must possess complex representations, in order that the matter particles and their antiparticles (described by complex conjugate spinors) could be in inequivalent representations. Thirdly, we should also keep track of the hypercharges $Y = Q - T_3$. One of the major puzzles of the SM is why

$$\sum_{q,\ell} Q_i = 3Q_u + 3Q_d + Q_e = 0. \quad (230)$$

In the SM, the hypercharge assignments are *a priori* independent of the $SU(3) \times SU(2)_L$ assignments, although constrained by the fact that quantum consistency requires the resulting triangle anomalies to cancel. In a simple GUT group, the relation (230) is automatic: whenever Q is a generator of a simple gauge group, $\sum_R Q = 0$ for particles in any representation R, cf., the values of I_3 in any representation of $SU(2)$.

There are only two groups of rank 4 that have complex representations and hence are suitable *a priori* for GUTs, namely $SU(5)$ and $SU(3) \otimes SU(3)$. However, $SU(3) \otimes SU(3)$ does not allow

[22] Each one is associated with a quantum number, a 'charge', that may be used to label particle states.

simultaneously the leptons to have an integer electric charge and the quarks to have a fractional electric charge. Moreover, if one tried to use $SU(3) \times SU(3)$, one would need to embed the electroweak gauge group in the second $SU(3)$ factor. This would be possible only if $\sum_q Q_q = 0 = \sum_\ell Q_\ell$, which is not the case for the known quarks and leptons. Therefore, attention has focused on $SU(5)$ [113] as the only possible rank-4 GUT group.

The group $SU(5)$ is the simplest GUT group capable of including the SM. Other possible GUT groups have higher rank, and groups that are commonly used are $SO(10)$, the only suitable simple group of rank 5 with complex representations, and the exceptional group E_6 of rank 6. As examples that may help understand the new physics that appears when the symmetry of the SM is enhanced, we are first going to study key aspects of the group $SU(5)$ and then, more briefly, some aspects of the group $SO(10)$.

The $SU(5)$ group

As in the SM, particles must be arranged in suitable representations of $SU(5)$. This group has a fundamental spinorial representation of dimension 5 and a 2-index antisymmetric spinorial representation of dimension 10. Together they are suitable for accommodating the fermions of a given generation, which consist of $3 \times 2 \times 2 = 12$ quarks + 2 charged leptons + 1 neutrino. To see how this may be done, we first decompose the smallest representations of $SU(5)$ in terms of representations of $SU(3) \otimes SU(2)$:

$$\bar{\mathbf{5}} = (\bar{\mathbf{3}}, \mathbf{1}) + (\mathbf{1}, \mathbf{2}), \tag{231}$$

$$\mathbf{10} = (\bar{\mathbf{3}}, \mathbf{1}) + (\mathbf{3}, \mathbf{2}) + (\mathbf{1}, \mathbf{1}). \tag{232}$$

For example, in (231) the representation $\bar{\mathbf{5}}$ of $SU(5)$ can accommodate a colour antitriplet that is also an $SU(2)$ singlet, and a colour singlet that is also an $SU(2)$ doublet. In addition, it is necessary that the sum of the charges in each of these two multiplets be zero. The only possible combination of first-generation fermions in the SM is:

$$\bar{\mathbf{5}} : (\psi_i)_L = \begin{pmatrix} \bar{d}_1 \\ \bar{d}_2 \\ \bar{d}_3 \\ e^- \\ -\nu_e \end{pmatrix}_L , \tag{233}$$

and the rest of the first-generation fermions may be accommodated uniquely, as follows:

$$\mathbf{10} : (\chi^{ij})_L = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & \bar{u}_3 & -\bar{u}_2 & u_1 & d_1 \\ -\bar{u}_3 & 0 & \bar{u}_1 & u_2 & d_2 \\ u_2 & -\bar{u}_1 & 0 & u_3 & d_3 \\ -u_1 & -u_2 & -u_3 & 0 & e^+ \\ -d_1 & -d_2 & -d_3 & -e^+ & 0 \end{pmatrix}_L , \tag{234}$$

where we neglect the eventual mixings between the fermions in different generations. We must repeat the previous classification of fermions in $\mathbf{10} + \bar{\mathbf{5}}$ representations for the other two generations: there is no explanation in $SU(5)$ for the presence of three generations [23].

After discussing the matter fermions, we now discuss the GUT gauge bosons. Groups of type $SU(N)$ have $N^2 - 1$ symmetry generators in an adjoint representation (e.g., $SU(3)_C$ has 8 gluons, $SU(2)$ has 2 W bosons, etc.), so that $SU(5)$ has 24 gauge bosons. Of these 24 gauge bosons, 12 correspond to the SM gluons, $W^\pm$, Z^0 and γ, and 12 are new. Decomposing this 24-dimensional adjoint representation into representations of $SU(3) \otimes SU(2) \otimes U(1)$, we find

$$\mathbf{24} = \underbrace{(\mathbf{3}, \mathbf{2}, \frac{5}{3}) \oplus (\bar{\mathbf{3}}, \mathbf{2}, -\frac{5}{3})}_{new\ bosons} \oplus \underbrace{(\mathbf{8}, \mathbf{1}, 0)}_{gluons\ G_a} \oplus \underbrace{(\mathbf{1}, \mathbf{3}, 0)}_{W_i} \oplus \underbrace{(\mathbf{1}, \mathbf{1}, 0)}_{B} , \tag{235}$$

[23] The pairing of $\bar{\mathbf{5}}$ and $\mathbf{10}$ representations is free of triangle anomalies.

where the third numbers in the parentheses are the hypercharges of the multiplets. The new bosons, called X and Y, have electric charges 4/3 and 2/3, respectively, carry leptoquark quantum numbers, are coloured and have isospin 1/2 [24]. In matrix notation,

$$A = \sum_{a=1}^{24} T_a A^a = \begin{pmatrix} G_i & G_i & G_i & \bar{X} & \bar{Y} \\ G_i & G_i & G_i & \bar{X} & \bar{Y} \\ G_i & G_i & G_i & \bar{X} & \bar{Y} \\ X & X & X & W_i & W_i \\ Y & Y & Y & W_i & W_i \end{pmatrix}, \tag{236}$$

where the T_a are the generators of $SU(5)$ represented by 5×5 matrices (the equivalents for $SU(5)$ of the Pauli matrices of $SU(2)$). The basis is chosen so that $SU(3)_C$ corresponds to the first three lines and columns, and $SU(2)_L$ to the last two lines. The top-left and bottom-right blocks therefore contain the gluons and W bosons, respectively, and the $U(1)$ boson B (not shown) corresponds to a traceless diagonal generator.

The remaining steps in constructing an $SU(5)$ GUT are the choices of representations for Higgs bosons, first to break $SU(5) \to SU(3) \times SU(2) \times U(1)$ and subsequently to break the electroweak $SU(2) \times U(1)_Y \to U(1)_{em}$. The simplest choice for the first stage is an adjoint $\mathbf{24}$ of Higgs bosons Φ with a v.e.v.

$$< 0|\Phi|0 >= \begin{pmatrix} 1 & 0 & 0 & \vdots & 0 & 0 \\ 0 & 1 & 0 & \vdots & 0 & 0 \\ 0 & 0 & 1 & \vdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \vdots & -\frac{3}{2} & 0 \\ 0 & 0 & 0 & \vdots & 0 & -\frac{3}{2} \end{pmatrix} \times \mathcal{O}(m_{GUT}). \tag{237}$$

It is easy to see that this v.e.v. preserves colour $SU(3)$, which reshuffles the first three rows and columns, weak $SU(2)$, which reshuffles the last two rows and columns, and the hypercharge $U(1)$, which is a diagonal generator. The subsequent breaking of $SU(2) \times U(1)_Y \to U(1)_{em}$ is most economically accomplished by a $\mathbf{5}$ representation of Higgs bosons H:

$$< 0|\phi|0 >= (0, 0, 0, 0, 1) \times 0(m_W). \tag{238}$$

It is clear that this v.e.v. has an $SU(4)$ symmetry which yields [104] the relation $m_b = m_\tau$ before renormalization that leads, after renormalization (221), to a successful prediction for m_b in terms of m_τ. However, the same trick does not work for the first two generations, indicating a need for epicycles in this simplest GUT model [114].

Making the minimal $SU(5)$ GUT supersymmetric, as motivated by the naturalness of the gauge hierarchy, is not difficult [94]. One must replace the above GUT multiplets by supermultiplets: $\bar{\mathbf{5}}$ $\bar{F}$ and $\mathbf{10}$ T for the matter particles, $\mathbf{24}$ Φ for the GUT Higgs fields that break $SU(5) \to SU(3) \times SU(2) \times U(1)$. The only complication is that one needs both $\mathbf{5}$ and $\bar{\mathbf{5}}$ Higgs representations H and $\bar{H}$ to break $SU(2) \times U(1)_Y \to U(1)_{em}$, just as two doublets were needed in the MSSM to cancel anomalies and give masses to all the matter fermions. The simplest possible form of the Higgs potential is specified by the superpotential [94]:

$$W = (\mu + \frac{3\lambda}{2} M) + \lambda \bar{H} \Phi H + f(\Phi) \tag{239}$$

[24] They have direct interactions with quarks and leptons, which we discuss in the next section.

where $\mu = \mathcal{O}(1)$ TeV and $M = \mathcal{O}(M_{GUT})$, and $f(\Phi)$ is chosen so that $\partial f/\partial\Phi = 0$ when

$$< 0|\Phi|0 >= M \begin{pmatrix} 1 & 0 & 0 & \vdots & 0 & 0 \\ 0 & 1 & 0 & \vdots & 0 & 0 \\ 0 & 0 & 1 & \vdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \vdots & -\frac{3}{2} & 0 \\ 0 & 0 & 0 & \vdots & 0 & -\frac{3}{2} \end{pmatrix}. \tag{240}$$

Inserting this into the second term of (239), one finds terms $\lambda M \bar{H}_3 H_3, \ -3/2\lambda M \bar{H}_2 H_2$ for the colour-triplet and weak-doublet components of $\bar{H}$ and H, respectively. Combined with the bizarre coefficient of the first term, these lead to terms

$$W \ni (\mu + \frac{5\lambda}{2} M)\bar{H}_3 H_3 + \mu \bar{H}_2 H_2. \tag{241}$$

Thus we have heavy Higgs triplets with masses $\mathcal{O}(M_{GUT})$ and light Higgs doublets with masses $\mathcal{O}(\mu)$. However, this requires fine tuning the coefficient of the first term in W (239) to about 1 part in 10^{13}! In the absence of supersymmetry, such fine tuning would be destroyed by quantum loop corrections [105].

A primary advantage of supersymmetry is that its no-renormalization theorems [80, 81] guarantee that this fine tuning is *natural*, in the sense that quantum corrections do not destroy it, unlike the situation without supersymmetry. On the other hand, supersymmetry alone does not explain the *origin* of the hierarchy. A second advantage of supersymmetry, as we saw earlier in this section, is that it would make possible a much more precise unification of the gauge couplings. However, a potential snag is that the exchanges of the supersymmetric partners of the heavy Higgs triplets $\bar{H}_3, H_3$ may cause rapid proton decay, as discussed later.

Another possible GUT group that is frequently studied is $SO(10)$ [113, 115]. It is a group of rank 5, that contains $SU(5) \otimes U(1)$. The principal advantage of $SO(10)$ over $SU(5)$ is that it possesses a fundamental spinorial representation of dimension 16 that can accommodate all the fermions of one generation, as well as a singlet right-handed neutrino, thanks to its decomposition in terms of $SU(5)$ representations [25]

$$\mathbf{16} = \mathbf{10} \oplus \bar{\mathbf{5}} \oplus \mathbf{1}. \tag{242}$$

The appearance of an $SU(5)$ singlet provides a natural framework for the physics of the neutrinos and the seesaw mechanism [26]. In $SO(10)$ the number of gauge bosons rises to 45, which includes 33 additional gauge bosons beyond the SM, and therefore many possible interactions, including additional options for proton decay. In addition, the breaking of $SO(10)$ is more complicated than that of $SU(5)$, because it is done in two steps. One may pass from $SO(10)$ to $SU(5) \otimes U(1)$ or $SU(4) \otimes SU(2)_L \otimes SU(2)_R$, and then to $SU(2) \otimes U(1)$. The Higgs sector is potentially quite extensive, and may include large multiplets of dimensions 10, 16, 45, 54, 120 and 126, depending on the model.

4.1.3 Baryon decay

Baryon instability is to be expected on general grounds, since there is no exact gauge symmetry to guarantee that baryon number B is conserved. Indeed, baryon decay is a generic prediction of GUTs, which we illustrate with the simplest $SU(5)$ model, that is anyway embedded in larger and more complicated

[25] The $SO(10)$ group is anomaly-free, so this decomposition explains finally the freedom from anomalies of $SU(5)$ and the SM.

[26] In $SU(5)$, singlet right-handed neutrinos could be added 'by hand', in which case they would have no gauge interactions. In the case of $SO(10)$, the gauge interactions of $SO(10)$ do not have any direct influence on accessible neutrino phenomenology, but may provide interesting restrictions on their Yukawa interactions.

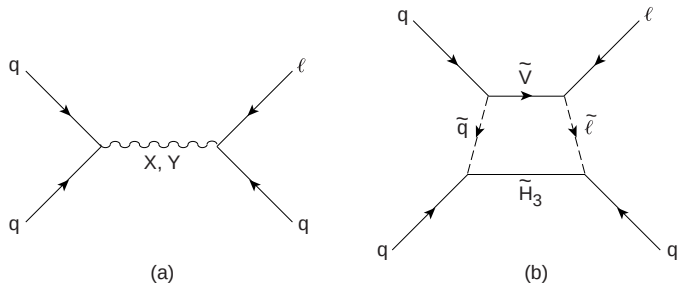


Fig. 30: Diagrams contributing to baryon decay (a) in minimal $SU(5)$ and (b) in minimal supersymmetric $SU(5)$

GUTs. We see in (236) that there are two species of gauge bosons in $SU(5)$, called X and Y, that couple the colour $SU(3)$ indices (1,2,3) to the electroweak $SU(2)$ indices (4,5). As we can see from the matter representations (234), these may enable two quarks or a quark and lepton to annihilate, as seen in Fig. 30(a). Combining these possibilities leads to an interaction with $\Delta B = \Delta L = 1$. The forms of effective four-fermion interactions mediated by the exchanges of massive Z and Y bosons, respectively, are [105]

$$\begin{aligned} \left(\epsilon_{ijk} u_{R_k} \gamma_\mu u_{L_j}\right) \ \frac{g_X^2}{8m_X^2} \ \left(2e_R\, \gamma^\mu\, d_{L_i} + e_L\, \gamma^\mu\, d_{R_i}\right) \ , \\ \left(\epsilon_{ijk} u_{R_k} \gamma_\mu d_{L_j}\right) \ \frac{g_Y^2}{8m_X^2} \ \left(\nu_L\, \gamma^\mu\, d_{R_i}\right) \ , \end{aligned} \tag{243}$$

up to generation mixing factors.

Since the gauge couplings $g_X = g_Y = g_{3,2,1}$ in an $SU(5)$ GUT, and $m_X \simeq m_Y$, we expect that

$$G_X \equiv \frac{g_X^2}{8m_X^2} \simeq G_Y \equiv \frac{g_Y^2}{8m_Y^2}. \tag{244}$$

It is clear from (243) that the baryon decay amplitude $A \propto G_X$, and hence the baryon $B \to \ell+$ meson decay rate

$$\Gamma_B = c G_X^2 m_p^5, \tag{245}$$

where the factor of m_p^5 comes from dimensional analysis, and c is a coefficient that depends on the GUT model and the non-perturbative properties of the baryon and meson.

The decay rate (245) corresponds to a proton lifetime

$$\tau_p = \frac{1}{c}\ \frac{m_X^4}{m_p^5}. \tag{246}$$

It is clear from (246) that the proton lifetime is very sensitive to m_X, which must therefore be calculated very precisely. In minimal $SU(5)$, the best estimate was

$$m_X \simeq (1 \text{ to } 2) \times 10^{15} \times \Lambda_{QCD} \tag{247}$$

where Λ_{QCD} is the characteristic QCD scale in the $\overline{\rm MS}$ prescription with four active flavours. Making an analysis of the generation mixing factors [116], one finds that the preferred proton (and bound neutron) decay modes in minimal $SU(5)$ are

$$\begin{aligned} p &\to e^+\pi^0 \ , \ \ e^+\omega \ , \ \ \bar\nu\pi^+ \ , \ \ \mu^+ K^0 \ , \ \ \ldots \\ n &\to e^+\pi^- \ , \ \ e^+\rho^- \ , \ \ \bar\nu\pi^0 \ , \ \ \ldots , \end{aligned} \tag{248}$$

and the best numerical estimate of the lifetime is

$$\tau(p \to e^+\pi^0) \simeq 2 \times 10^{31\pm1} \times \left(\frac{\Lambda_{QCD}}{400\ \mathrm{MeV}}\right)^4\ y\ . \tag{249}$$

This is in *prima facie* conflict with the latest experimental lower limit

$$\tau(p \to e^+\pi^0) > 8.2 \times 10^{33}\ y \tag{250}$$

from super-Kamiokande [117]. However, this failure of minimal $SU(5)$ is not as conclusive as the failure of its prediction for $\sin^2\theta_W$.

We saw earlier that supersymmetric GUTs, including $SU(5)$, fare better with $\sin^2\theta_W$. They also predict a larger GUT scale [107]:

$$m_X \simeq 2 \times 10^{16}\ \mathrm{GeV}, \tag{251}$$

so that $\tau(p \to e^+\pi^0)$ is considerably longer than the experimental lower limit. However, this is not the dominant proton decay mode in supersymmetric $SU(5)$ [118]. In this model, there are important $\Delta B = \Delta L = 1$ interactions mediated by the exchange of colour-triplet higgsinos $\tilde{H}_3$, dressed by gaugino exchange as seen in Fig. 30(b) [119], these give

$$G_X \to \mathcal{O}\left(\frac{\lambda^2 g^2}{16\pi^2}\right)\frac{1}{m_{\tilde{H}_3}\tilde{m}}, \tag{252}$$

where λ is a generic Yukawa coupling. Taking into account colour factors and the values of λ for more massive particles, it was found [118] that decays into neutrinos and strange particles should dominate:

$$p \to \bar{\nu}K^+\ ,\ \ n \to \bar{\nu}K^0\ ,\ \ \ldots \tag{253}$$

Because there is only one factor of a heavy mass $m_{\tilde{H}_3}$ in the denominator of (252), these decay modes are expected to dominate over $p \to e^+\pi^0$ etc. in minimal supersymmetric $SU(5)$. The current experimental limit is $\tau(p \to \bar{\nu}K^+) > 10^{33}y$ [120]. Calculating carefully the other factors in (252) [121], it seems that the modes (253) may be close to detectability in this model, possibly even too close for comfort, in which case a more complicated supersymmetric GUT might be needed.

There are non-minimal supersymmetric GUT models such as flipped $SU(5)$ [122] in which the $\tilde{H}_3$- exchange mechanism (252) is suppressed. In such models, $p \to e^+\pi^0$ may again be the preferred decay mode [123]. However, this is not necessarily the case, as colour-triplet Higgs boson exchange may also be important, in which case $p \to \mu^+K^0$ could be dominant [124], or there may be non-intuitive generation mixing in the couplings of the X and Y bosons, offering the possibility $p \to \mu^+\pi^0$ etc. Therefore, the continuing search for proton decay should be open-minded about the possible decay modes. The current experimental limits for these process are $\tau(p \to e^+\pi^0) > 10^{33}y$ [117], $\tau(p \to \mu^+K^0) > 10^{33}y$ [120], and $\tau(p \to \mu^+\pi^0) > 10^{33}y$ [117].

4.1.4 Neutrino masses and oscillations

The experimental upper limits on neutrino masses are far below the corresponding lepton masses [13]. From studies of the end-point of tritium β decay, we have

$$m_{\nu_e} \lesssim 2\ \mathrm{eV}, \tag{254}$$

to be compared with $m_e = 0.511$ MeV. Neglecting mixing effects, from studies of $\pi \to \mu\nu_\mu$ decays, we have

$$m_{\nu_\mu} < 190\ \mathrm{keV}, \tag{255}$$

to be compared with m_μ = 105 MeV, and from studies of $\tau \to$ pions + ν_τ, again neglecting mixing effects, we have

$$m_{\nu_\tau} < 18.2\ \mathrm{MeV}, \tag{256}$$

to be compared with m_τ = 1.78 GeV.

On the other hand, there is no good symmetry reason to expect the neutrino masses to vanish. We expect masses to vanish only if there is a corresponding exact gauge symmetry, cf., m_γ = 0 in QED with an unbroken $U(1)$ gauge symmetry.

However, although there is no candidate gauge symmetry to ensure $m_\nu = 0$, this is a prediction of the SM. We recall that the neutrino couplings to charged leptons take the form

$$J_\mu = \bar{e}\gamma_\mu(1-\gamma_5)\nu_e + \bar{\mu}\gamma_\mu(1-\gamma_5)\nu_\mu + \bar{\tau}\gamma_\mu(1-\gamma_5)\nu_\tau, \tag{257}$$

and that only left-handed neutrinos have ever been detected. In the cases of charged leptons and quarks, their masses arise in the SM from couplings between left- and right-handed components *via* a Higgs field:

$$g_{H\bar{f}f}\ H_{\Delta I=\frac{1}{2},\Delta L=0}\ \ \bar{f}_R f_L + h.c. \to m_f = g_{H\bar{f}f}\langle 0|H_{\Delta I=\frac{1}{2},\Delta L=0}|0\rangle. \tag{258}$$

Such a left–right coupling is conventionally called a Dirac mass. The following questions arise for neutrinos: if there is no ν_R, can one have $m_\nu \neq 0$? On the other hand, if there is a ν_R, why are the neutrino masses so small?

The answer to the first question is positive, because it is possible to generate neutrino masses *via* the Majorana mechanism that involves the ν_L alone. This is possible because an $(\overline{f_R})$ field is in fact left-handed: $(\overline{f_R}) = (f^c)_L = f_L^T C$, where the superscript T denotes a transpose, and C is a 2×2 conjugation matrix. We can therefore imagine replacing

$$(\overline{f_R}) f_L \to f_L^T\ C\ f_L, \tag{259}$$

which we denote by $f_L \cdot f_L$. In the cases of quarks and charged leptons, one cannot generate masses in this way, because $q_L \cdot q_L$ has ΔQ_{em}, Δ(colour) $\neq 0$ and $\ell_L \cdot \ell_L$ has $\Delta Q_{em} \neq 0$. However, the coupling $\nu_L \cdot \nu_L$ is not forbidden by such exact gauge symmetries, and would lead to a neutrino mass:

$$m^M\ \nu_L^T\ C\ \nu_L = m^M (\overline{\nu^c})_L \nu_L \equiv m^M\ \nu_L \cdot \nu_L. \tag{260}$$

Such a combination has non-zero net lepton number $\Delta L = 2$ and weak isospin $\Delta I = 1$. There is no corresponding Higgs field in the SM or in the minimal $SU(5)$ GUT, but there is no obvious reason to forbid one. If one were present, one could generate a Majorana neutrino mass *via* the renormalizable coupling

$$\tilde{g}_{H\bar{\nu}\nu}\ \ H_{\Delta I=1,\Delta L=L}\ \ \nu_L \cdot \nu_L \Rightarrow m^M = \tilde{g}_{H\bar{\nu}\nu}\langle 0|H_{\Delta I=1,\Delta L=2}|0\rangle. \tag{261}$$

However, one could also generate a Majorana mass without such an additional Higgs field, *via* a non-renormalizable coupling to the conventional $\Delta I = \frac{1}{2}$ SM Higgs field:

$$\frac{1}{M}\ \ \left(H_{\Delta I=\frac{1}{2}}\nu_L\right) \cdot \left(H_{\Delta I=\frac{1}{2}}\nu_L\right) \Rightarrow m^M = \frac{1}{M}\langle 0|H_{\Delta I=\frac{1}{2}}|0\rangle^2, \tag{262}$$

where M is some (presumably heavy mass scale: $M \gg m_W$).

The simplest possibility for generating a non-renormalizable interaction of the form (262) would be *via* the exchange of a heavy field N that is a singlet of $SU(3) \times SU(2) \times U(1)$ or $SU(5)$:

$$\frac{1}{M} \to \frac{\lambda^2}{M_N}, \tag{263}$$

where one postulates a renormalizable coupling $\lambda H_{\Delta I=1/2} \nu_L \cdot N$. As already mentioned, such a heavy singlet field appears automatically in extensions of the $SU(5)$ GUT, such as $SO(10)$, though it does not actually *require* the existence of any new GUT gauge bosons.

We now have all the elements we need for the see-saw mass matrix [125] favoured by GUT model-builders:

$$(\nu_L, N) \cdot \begin{pmatrix} m^M & m^D \\ m^D & M^M \end{pmatrix} \begin{pmatrix} \nu_L \\ N \end{pmatrix}, \tag{264}$$

where the $\nu_L \cdot \nu_L$ Majorana mass m^M might arise from a $\Delta I = 1$ Higgs with coupling $\tilde{g}_{H\bar{\nu}\nu}$, (261), the $\nu_L \cdot N$ Dirac mass m^D could arise from a conventional Yukawa coupling λ (263) and should be of the same order as a conventional quark or lepton mass, and M^M could *a priori* be $\mathcal{O}(M_{GUT})$ [27]. Diagonalizing (264) and assuming that $m^M = 0$ or that $\langle 0|H_{\Delta I=1}|0\rangle = \mathcal{O}(m_W^2/m_{GUT})$, as generically expected in GUTs, one obtains the mass eigenstates

$$\nu_L + 0\left(\frac{m_W}{m_X}\right) N \quad : \quad m = \mathcal{O}\left(\frac{m_W^2}{M_{GUT}}\right), \tag{265}$$

$$N + 0\left(\frac{m_W}{m_X}\right) \nu_L \quad : \quad M = \mathcal{O}(M_{GUT}). \tag{266}$$

We see that one mass eigenstate (265) is naturally much lighter than the electroweak scale, whereas the other (266) is naturally much heavier.

There is evidence for atmospheric neutrino oscillations [127] between ν_μ and ν_τ with $\Delta m_A^2 \sim (10^{-2}$ to $10^{-3})$ eV^2 and a large mixing angle: $\sin^2\theta_{23} \gtrsim 0.9$. In addition, there is evidence [128] for solar neutrino oscillations with $\Delta m_S^2 \simeq 10^{-5}$ eV^2 and $\sin^2\theta_{12} \sim 0.6$. We also know that the third neutrino mixing angle θ_{13} must be small, but it is an open experimental question just how small it may be. The pattern of MNS neutrino mixing seems very different from that of CKM quark mixing, perhaps reflecting special ingredients related to the see-saw mechanism. Other open questions include the magnitude of the CP-violating phase in the neutrino mixing matrix (analogous to the Kobayashi–Maskawa phase in quark mixing), and also the sequence of neutrino mass eigenstates.

CP-violating decays of heavy singlet neutrinos provide a simple mechanism for generating the baryon number of the Universe [129], by first providing a lepton asymmetry that is subsequently converted partially into a baryon asymmetry by non-perturbative electroweak interactions [15]. Essential ingredients in this scenario are the violation of lepton number *via* Majorana neutrino masses and CP violation [38]. The CP-violating phase observable in neutrino oscillations does not play a direct role in this scenario for baryogenesis [130], but its observation would nevertheless be of great conceptual importance.

4.2 Local supersymmetry and supergravity

Why study a local theory of supersymmetry [82,83]? One motivation is the analogy with gauge theories, in which bosonic symmetries are made local. Another is that local supersymmetry necessarily involves the introduction of gravity. Since both gravity and (surely!) supersymmetry exist, this seems an inevitable step. It also leads to the possibility of unifying all the particle interactions including gravity, which was one of our original motivations for supersymmetry. Moreover, it is interesting that local supersymmetry (supergravity) admits an elegant mechanism for supersymmetry breaking [131], analogous to the Higgs mechanism in gauge theories, which allows us to address more seriously the possible existence of a cosmological constant.

[27] It is often assumed that there are three singlet neutrinos N, but this need not be the case. If there were only two, one of the light neutrinos would be massless. On the other hand, there could be many more than three [126].

The basic building block in a supergravity theory [82, 83] is the graviton supermultiplet, which contains particles with helicities $(2, 3/2)$, the latter being the gravitino of spin $3/2$. Why is this required when one makes supersymmetry local?

We recall the basic global supersymmetry transformation laws (150, 151) for bosons and fermions. Consider now the combination of two such global supersymmetry transformations

$$[\delta_1, \delta_2] \; (\phi \text{ or } \psi) = -(\bar{\xi}_2 \gamma_\mu \xi_1) \; (i \, \partial_\mu) \; (\phi \text{ or } \psi) + \dots \tag{267}$$

The operator $(i \, \partial_\mu)$ corresponds to the momentum P_μ, and we see again that the combination of two global supersymmetry transformations is a translation. Consider now what happens when we consider local supersymmetry transformations characterized by a varying spinor $\xi(x)$. It is evident that the infinitesimal translation $\bar{\xi}_2 \gamma^\mu \xi_1$ in (267) is now x-dependent, and the previous global translation becomes a local coordinate transformation, as occurs in General Relativity.

How do we make the theory invariant under such local supersymmetry transformations? Consider again the simplest globally supersymmetric model containing a free spin-1/2 fermion and a free spin-0 boson (143), and make the local versions of the transformations (151), we can obtain

$$\delta \mathcal{L} = \partial_\mu (\cdots) + 2 \bar{\psi} \gamma_\mu \partial\!\!\!/ S (\partial^\mu \xi(x)) + \text{herm. conj.} \tag{268}$$

In contrast to the global case, the action $A = \int d^4x \mathcal{L}$ is not invariant, because of the second term in (268). To cancel it out and restore invariance, we need more fields.

We proceed by analogy with gauge theories. In order to make the kinetic term $(i \bar{\psi} \partial\!\!\!/ \psi)$ invariant under gauge transformations $\psi \to e^{i\epsilon(x)} \psi$, we need to cancel a variation

$$-\bar{\psi} \partial_\mu \psi \partial^\mu \epsilon(x), \tag{269}$$

which is done by introducing a coupling to a gauge boson

$$g \bar{\psi} \gamma_\mu \psi A^\mu(x) \; , \tag{270}$$

and the corresponding transformation

$$\delta A_\mu(x) = \frac{1}{g} \partial_\mu \epsilon(x). \tag{271}$$

In the supersymmetric case, we cancel the second term in (268) by a coupling

$$\kappa \bar{\psi} \gamma_\mu \partial\!\!\!/ S \psi^\mu(x) \tag{272}$$

to a spin-3/2 spinor $\psi^\mu(x)$, representing a gauge fermion or gravitino, with the corresponding transformation

$$\delta \psi^\mu = -\frac{2}{\kappa} \, \partial^\mu \xi(x), \tag{273}$$

where $\kappa \equiv 8\pi / m_P^2$.

For completeness, let us at least write down the Lagrangian for the graviton–gravitino supermultiplet

$$L = -\frac{1}{2\kappa^2} \sqrt{-g} R - \frac{1}{2} \, \epsilon^{\mu\nu\rho\sigma} \bar{\psi}_\mu \gamma_5 \gamma_\nu \mathcal{D}_\rho \psi_\sigma, \tag{274}$$

where g denotes the determinant of the metric tensor

$$g_{\mu\nu} = \epsilon_\mu^m \eta_{mn} \epsilon_\nu^\mu, \tag{275}$$

ϵ_μ^m is the vierbein and η_{mn} the Minkowski metric tensor, and $\mathcal{D}_\rho$ is a covariant derivative

$$\mathcal{D}_\rho \equiv \partial_\rho + \frac{1}{4} \, \omega_\rho^{mn} [\gamma_m, \gamma_n], \tag{276}$$

where ω_ρ^{mn} is the spin connection. This is the simplest possible generally-covariant model of a spin-3/2 field. It is remarkable that it is invariant under the local supersymmetry transformations

$$\begin{aligned} \delta\epsilon_\mu^m &= \frac{x}{2}\,\bar{\xi}(x)\gamma^m\psi_\mu(x), \\ \delta\omega_\mu^{mn} &= 0, \delta\psi_\mu = \frac{1}{x}\,\mathcal{D}_\mu\xi(x), \end{aligned} \tag{277}$$

just as the simplest possible $(1/2, 0)$ theory (143) was globally supersymmetric, and also the action of an adjoint spin-1/2 field in a gauge theory.

As already remarked, supergravity admits an elegant analogue of the Higgs mechanism of spontaneous symmetry breaking [131]. Just as one combines the two polarization states of a massless gauge field with the single state of a massless Goldstone boson to obtain the three polarization states of a massive gauge boson, one may combine the two polarization states of a massless gravitino ψ_μ with the two polarization states of a massless Goldstone fermion λ to obtain the four polarization states of a massive spin-3/2 particle $\tilde{G}$. This super-Higgs mechanism corresponds to a spontaneous breakdown of local supersymmetry, since the massless graviton G has a different mass from the gravitino $\tilde{G}$:

$$m_G = 0 \neq m_{\tilde{G}}. \tag{278}$$

This is the only known consistent way of breaking local supersymmetry, just as the Higgs mechanism is the only way to generate $m_W \neq 0$.

Moreover, this can be achieved while keeping zero vacuum energy (cosmological constant), at least at the tree level. The reason for this is the appearance in local supersymmetry (supergravity) of a third term in the effective potential (170), which has a *negative* sign [131]. There is no time in these lectures to discuss this exciting feature in detail: the interested reader is referred to the original literature and the simplest example [132]. In this particular case, $\Lambda = V = 0$ for *any* value of the gravitino mass, for which reason it was named no-scale supergravity [133].

Again, there is no time to discuss here details of the coupling of supergravity to matter [131]. However, it is useful to have in mind the general features of the theory in the limit where $\kappa \to 0$, but the gravitino mass $m_{\tilde{G}} \equiv m_{3/2}$ remains fixed. One generally has non-zero gaugino masses $m_{1/2} \propto m_{3/2}$, and their universality is quite generic. One also has non-zero scalar masses $m_0 \propto m_{3/2}$, but their universality is much more problematic, and even violated in generic string models. It was this failing that partly refuelled interest in gauge-mediated models. A generic supergravity theory also yields non-universal trilinear soft supersymmetry-breaking couplings $A_\lambda\lambda\phi^3 : A_\lambda \propto m_{3/2}$ and bilinear scalar couplings $B_\mu\mu\phi^2 : B_\mu \propto m_{3/2}$. Therefore, supergravity may generate the full menagerie of soft supersymmetry-breaking terms:

$$-\frac{1}{2}\sum_a m_{1/2_a}\,\tilde{V}_a\tilde{V}_a - \sum_i m_{0_i}^2|\phi_i|^2 - \left(\sum_\lambda A_\lambda\lambda\phi^3 + \text{h.c.}\right) - \left(\sum_\mu B_\mu\mu\phi^2 + \text{h.c.}\right). \tag{279}$$

In a minimal supergravity (mSUGRA) framework, the gaugino masses $m_{1/2}$, scalar masses m_0, and trilinear couplngs A are universal, as assumed in the CMSSM, but there are specific conditions: $B = A - 1$, and the gravitino mass is fixed: $m_{3/2} = m_0$. The former condition is more restrictive than in the CMSSM, and the latter condition implies that the gravitino is the LSP in significant regions of parameter space. Hence, the CMSSM and mSUGRA are distinct scenarios [134].

Since these soft supersymmetry-breaking parameters are generated at the supergravity scale near $m_P \sim 10^{19}$ GeV, the soft supersymmetry-breaking parameters are renormalized as discussed earlier. The analogous parameters in gauge-mediated models would also be renormalized, but to a different extent, because the mediation scale $\ll m_P$. This difference may provide a signature of such models, as discussed elsewhere [135, 136].

Also renormalized is the vacuum energy (cosmological constant), which is a potential embarassment. Loop corrections in a non-supersymmetric theory are quartically divergent, whereas those in a generic supergravity theory are only quadratically divergent, suggesting a contribution to the cosmological constant of order $m_{3/2}^2 m_P^2$, perhaps $O(10^{-32})m_P^4$! Particular models may have a one-loop quantum correction of order $m_{3/2}^4 = O(10^{-64})m_P^2$, but more magic (a new symmetry?) is needed to suppress the cosmological constant to the required level

$$\Lambda \lesssim 10^{-123} m_P^4. \tag{280}$$

This is one of the motivations for seeking a fundamental Theory of Everything including gravity.

Once upon a time, supergravity was considered a possible candidate for such a Theory of Everything, particularly the maximal $\mathcal{N} = 8$ supergravity in 4 dimensions. However, this candidature would need two elements that are still lacking: a proof that the theory is finite, or at least renormalizable, and a demonstration of how it could lead to a low-energy theory resembling the SM, e.g., *via* the formation of bound states: see Ref. [137] for a review of these issues. In the meantime, string theory [90] is the most plausible candidate for a Theory of Everything.

4.3 Towards a Theory of Everything

4.3.1 Problems in quantum gravity

One of the most important unfinished tasks for understanding the Universe and the fundamental interactions is the unification of the two great theories of the 20th century: general relativity and quantum mechanics. To write such a unified Theory of Everything is one of the major challenges for physicists in our century. The solution of the problem of the cosmological constant, for example, will have to find a place in the frame of such a Theory of Everything.

Gravity is a puzzle for conventional quantum theory, in particular because incontrollable, non-renormalizable infinities appear when one tries to calculate Feynman diagrams that contain loops with gravitons. These correction terms diverge increasingly rapidly as the order of the perturbative calculation increases, essentially because the coupling of gravity has negative mass dimensionality, being $\propto 1/M_P^2$, where $M_P \simeq 1.2 \times 10^{19}$ GeV.

There are also non-perturbative problems in the quantization of gravity, which first appeared in connection with black holes. We recall that a black hole is a non-perturbative solution of the equations of General Relativity, in which the curvature of space-time induced by gravitational forces becomes so strong that no particle can escape the event horizon. The existence of this horizon is linked to the existence of entropy S and a non-zero temperature T of the black hole. From the pioneering work of Bekenstein and Hawking [138] on black-hole thermodynamics, we know that the mass of a black hole is proportional to the surface area A of its horizon, which is related in turn to its entropy:

$$S = \frac{1}{4} A \, . \tag{281}$$

The appearance of non-zero entropy means that the quantum description of a black hole must involve mixed states. The intuition underlying this feature is that information can be lost through the event horizon. To see how this may happen, consider, for example, a pure quantum-mechanical pair state $|A, B\rangle \equiv \sum_i c_i |A_i\rangle |B_i\rangle$ prepared near the horizon, and what happens if one of the particles, say A, falls through the horizon while B escapes, as seen in Fig. 31. In this case, all the information about the component $|A_i\rangle$ of the wave function is lost, so that

$$\sum_i c_i |A_i B_i\rangle \to \sum_i |c_i|^2 |B_i\rangle\langle B_i| \tag{282}$$

and B emerges in a mixed state, as in Hawking's original treatment of the black-hole radiation that bears his name [138]. The problem is that conventional quantum mechanics does not permit the evolution of a pure initial state into a mixed final state.

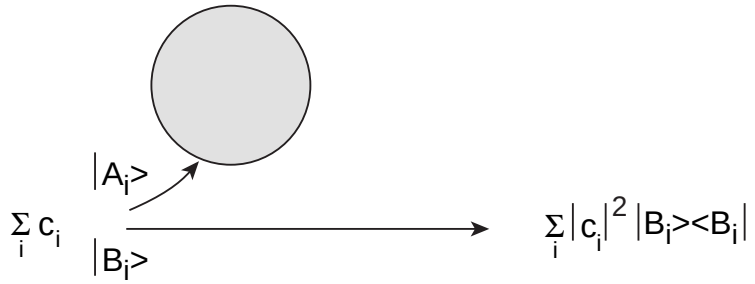


Fig. 31: If a pair of particles $|A\rangle\ |B\rangle$ is produced near the horizon of a black hole, and one of them ($|A\rangle$, say) falls in, the remaining particle $|B\rangle$ will appear to be in a mixed state, since the state of $|A\rangle$ is unobservable

For a discussion of these and other open problems in quantum black hole physics, see Ref. [139]. Many theorists consider that these problems point to a fundamental conflict between the proudest achievements of early-twentieth-century physics, namely quantum mechanics and General Relativity. One or the other should be modified, and perhaps both. Since quantum mechanics is sacred to field theorists, most particle physicists prefer to modify General Relativity by elevating it to string theory, as we now discuss.

4.3.2 Introduction to string theory

As was just mentioned, one of the major issues of quantum gravity is that it has an infinite number of infinities. These divergences can be traced to the absence of a short-distance cut-off in conventional field theories, where the particles are points. The problem is that one can in principle approach infinitely near a point particle, giving rise to interactions of infinite strength:

$$\int^{\Lambda\to\infty} d^4k \left(\frac{1}{k^2}\right) \leftrightarrow \int_{1/\Lambda\to 0} d^4x \left(\frac{1}{x^6}\right) \sim \Lambda^2 \to \infty. \tag{283}$$

Such divergences can be avoided or removed if one replaces point particles by extended objects. The simplest possibility is to extend in just one dimension, leading to a theory of strings. In such a theory, instead of point particles moving along one-dimensional world lines, one has strings moving over two-dimensional world sheets. Historically, closed loops of string have been the most popular, and the corresponding world sheet would be tubes. The 'wiring diagrams' generated by the Feynman rules of conventional point-like particle theories become 'plumbing circuits' generated by the junctions and connections of these tubes of closed string. One could imagine generalizing this idea to higher-dimensional extended objects such as membranes describing world volumes, etc., and we return later to this option.

Back in the early 1960s, there existed a quantum theory of the electromagnetic force (QED), but successful descriptions of the weak and strong forces were not yet known. At that time, theoretical efforts were concentrated on developing a theory that would determine the scattering (S) matrix, which describes on-mass-shell scattering amplitudes, which should possess certain properties abstracted from quantum field theory, such as unitarity and maximal analytic properties. These characteristics would ensure the requirements of causality and non-negative probabilities. A key idea in those years was maximal analyticity in the angular momentum plane, i.e., that the conventional partial-wave amplitudes $a_l(s)$ defined in the first instance for discrete angular momenta $l = 0, 1, ...$, can be extended uniquely to analytic functions of l, $a(l, s)$. These have isolated 'Regge' poles that move along Regge trajectories $l = \alpha(s)$ in the complex angular-momentum plane. The values of s for which l take suitable discrete values correspond to a physical hadron states. Experimental results indicated that the Regge trajectories are approximately linear, with a common slope α':

$$\alpha(s) = \alpha(0) + \alpha' s, \tag{284}$$

where $\alpha' \sim 1.0(\text{GeV})^{-2}$. These ideas were insufficient to determine the S matrix, and additional principles were invoked, such as the *bootstrap* idea, according to which the exchanges of hadrons in crossed channels provide forces that are responsible for forming hadronic bound states. In the narrow-resonance approximation, i.e., if resonance decay widths are negligible compared to their masses, the scattering amplitude can be expanded in an infinite series of s-channel poles, and this should give the same result as its expansion in an infinite series of t-channel poles due to exchanged particles. The narrow-resonance version of the bootstrap idea, which was called duality, had a precise formulation with a definite solution.

The decisive contribution to the solution was made by Veneziano in 1968 [140]: he gave an analytic formula that exhibited duality with linear Regge trajectories. Its structure was the sum of three Euler beta functions [141]:

$$T \;=\; A(s,t) + A(s,u) + A(t,u): \quad A(s,t) \;=\; \frac{\Gamma(-\alpha(s))\Gamma(-\alpha(t))}{\Gamma(-\alpha(s)-\alpha(t))}, \tag{285}$$

where α is a linear Regge trajectory, with $\alpha(s) = \alpha(0) + \alpha' s$ as described above. In the course of the next few years, several further breakthroughs were achieved. Virasoro [142] showed how to generalize the Veneziano formula to one with full symmetry in the three Mandelstam invariants s, t, u. Multiparticle generalizations of the Veneziano and Virasoro formulas were constructed and shown to factorize consistently on a finite spectrum of single-particle states at each energy level, which could be described by an infinite number of simple harmonic oscillators. This surprising result led to the first ideas of strings [143]: they could be interpreted as the scattering modes of a relativistic string: open strings in the Veneziano case and closed strings in the Virasoro case [28].

While looking for a way to incorporate baryons into the string framework, in 1971 Ramond [75] constructed a dual-resonance model generalization of the Dirac equation. The solutions of this equation gave the spectrum of a noninteracting fermionic string. In combination with work by Neveu and Schwarz [76], this led to a unified interacting theory of bosons and fermions, which was essentially a prototype for what later came to be known as superstring theory. The action of this theory has two-dimensional global supersymmetry on the world-sheet, described by infinitesimal fermionic transformations of the type discussed in the previous Lecture.

Initially, it was regarded as a disadvantage that this first incarnation of string theory was not able to accommodate the point-like partons seen inside hadrons at this time. In retrospect, this was the converse of the quantum-gravity motivation for string theory mentioned at the beginning of this section, which disfavours point-like structures. Then in 1973 along came QCD which incorporated these point-like scaling properties and provided a qualitative understanding of confinement that has now become quantitative with the advent of modern lattice calculations. Thus string theory languished as a candidate model of the strong interactions, though there is still hope that some as yet undiscovered variant of string theory might provide a useful alternative description of the strong interactions. In the mean time, interest was sparked in 1973 by the realization that string theory predicted the existence of a massless spin-2 state [144]. Could this be the graviton? It was known that in any consistent theory of a massless spin-2 particle its low-energy interactions would be identical with those of general relativity. Might string theory be a consistent high-energy completion of this theory, in which case it might be the longsought Theory of Everything?

As already mentioned, one of the primary reasons for studying extended objects in connection with quantum gravity is the softening of divergences associated with short-distance behaviour. Since the string propagates on a world sheet, the basic formalism is two-dimensional. Accordingly, string vibrations may be described in terms of left- and right-moving waves:

$$\phi(r,t) \to \phi_L(r-t),\ \phi_R(r+t). \tag{286}$$

[28] It still seems amazing that the mathematical formulae preceded the string interpretation [141].

If the string has no boundary, as for a closed string, the left- and right-movers are independent. When quantized, they may be described by a two-dimensional field theory. Compared to a four-dimensional theory, it is relatively easy to make a two-dimensional field theroy finite. In this case, it has conformal symmetry, which has an infinite-dimensional symmetry group in two dimensions. However, as you already know from gauge theories, one must be careful to ensure that this classical symmetry is not broken at the quantum level by anomalies. If the quantum string theory is to be consistent in a flat background space-time, the conformal anomaly fixes the number of left- and right-movers each to be equivalent to 26 free bosons if the theory has no supersymmetry, or 10 boson/fermion supermultiplets if the theory has $N = 1$ supersymmetry on the world sheet. There are other important quantum consistency conditions, and it was the demonstration by Green and Schwarz [145] that certain string theories are completely anomaly-free that opened the floodgates of theoretical interest in string theory as a potential Theory of Everything.

Among consistent string theories, one may enumerate the following. The *bosonic string* exists in 26 dimensions, but this is not even its worst problem! It contains no fermionic matter degrees of freedom, and the flat-space vacuum is intrinsically unstable. *Superstrings* exist in 10 dimensions, have fermionic matter and also a stable flat-space vacuum. On the other hand, the ten-dimensional theory is left-right symmetric, and the incorporation of parity violation in four dimensions is not trivial. The *heterotic string* was originally formulated in 10 dimensions, with parity violation already incorporated, since the left- and right movers were treated differently. This theory also has a stable vacuum, but still suffers from the disadvantage of having too many dimensions. *Four-dimensional heterotic strings* may be obtained either by compactifying the six surplus dimensions: $10 = 4 + 6$ compact dimensions with size $R \sim 1/m_P$, or by direct construction in four dimensions, replacing the missing dimensions by other internal degrees of freedom such as fermions or group manifolds or ...? In this way it was possible to incorporate a GUT-like gauge group [122] or even something resembling the Standard Model.

What are the general features of such string models? First, they predict there are no more than 10 dimensions, which agrees with the observed number of 4. Secondly, they suggest that the rank of the four-dimensional gauge group should not be very large, in agreement with the rank 4 of the Standard Model [29]. Thirdly, the simplest four-dimensional string models do not accommodate large matter representations [146], such as an **8** of SU(3) or a **3** of SU(2), again in agreement with the known representation structure of the Standard Model. Fourthly, simple string models predict fairly successfully the mass of the top quark, from the requirement that the theory make sense at all energies up to the Planck mass. Fifthly, string theory makes a fairly successful prediction for the gauge unification scale in terms of m_P. If the intrinsic string coupling g_s is weak, one predicts

$$M_{GUT} = O(g) \times \frac{m_P}{\sqrt{8\pi}} \simeq \text{few} \times 10^{17}\text{GeV}, \tag{287}$$

where g is the gauge coupling, which is $\mathcal{O}(20)$ higher than the value calculated on the basis of LEP measurement of the gauge couplings. Nevertheless, it would be nice to obtain closer agreement, and this provides the major motivation for considering strongly-coupled string theory, which corresponds to a large internal dimension $l > m_{GUT}^{-1}$, as we discuss next.

4.3.3 M theory

As was already said, the bosonic string model has many more disadvantages than other models. It has 26 dimensions, does not contain fermions, and has an unstable vacuum. Consequently, physicists focused on superstring models, of which five types exist:

- Type IIA, that reduces at low energy to a non-chiral $N = 2$ supergravity in $d = 10$ dimensions;
- Type IIB, that reduces at low energy to a chiral $N = 2$ supergravity in $d = 10$ dimensions;

[29] However, the number of gauge symmetries may be enhanced by non-perturbative effects.

- The heterotic $E(8) \times E(8)$ theory, that reduces at low energy to an $N = 1$ supergravity in $d = 10$, connected to a Yang–Mills gauge theory with an $E(8) \times E(8)$ gauge group;
- The heterotic theory $SO(32)$, that reduces at low energy to an $N = 1$ supergravity in $d = 10$, connected to a Yang–Mills gauge theory with an $SO(32)$ gauge group;
- Type I, that contains simultaneously opened and closed strings, and that reduces at low energy to an $N = 1$ supergravity in $d = 10$ connected to a Yang–Mills gauge theory with an $SO(32)$ gauge group.

These theories all look different. For example, the Type I theory is the only one that contains simultaneously open and closed strings, whereas the others contain only closed strings. In addition, the low-energy gauge structures of the five theories are different. It seems then, that we have five distinct theories that may describe gravity at the quantum level. How may we understand this? Is it possible that there is a link between the different theories?

Current developments involve going beyond strings to consider higher-dimensional extended objects, such as generalized membranes with various numbers of internal dimensions. These can be regarded as solitons (non-perturbative classical solutions) of string theory [147], with masses

$$m \propto \frac{1}{g_s}, \tag{288}$$

somewhat analogously to monopoles in gauge theory. It is evident from (288) that such membrane-solitons become light in the limit of strong string coupling: $g_s \to \infty$.

It was observed some time ago that there should be a strong-coupling/weak-coupling duality between elementary excitations and monopoles in supersymmetric gauge theories. These ideas were confirmed in a spectacular solution of $\mathcal{N} = 2$ supersymmetric gauge theory in four dimensions [148]. Similarly, it was shown that there are analogous dualities in string theory [149], whereby solitons in some strongly-coupled string theory are equivalent to light string states in some other weakly-coupled string theory. Indeed, it appears that all string theories are related by such dualities. A peculiarity of this discovery is that the string coupling strength g_s is related to an extra dimension in such a way that its size $R \to \infty$ as $g_s \to \infty$. This then leads to the idea of an underlying 11-dimensional framework called M theory [71] that reduces to the different string theories in different strong/weak-coupling linits, and reduces to eleven-dimensional supergravity in the low-energy limit (see Fig. 32).

A particular class of string solitons called D-branes offers a promising approach to the black hole information paradox mentioned previously. According to this picture, black holes are viewed as solitonic balls of string, and their entropy simply counts the number of internal string states. These are in principle countable, so string theory may provide an accounting system for the information contained in black holes. Within this framework, the previously paradoxical process (282) becomes

$$|A, B\rangle + |BH\rangle \to |B'\rangle + |BH'\rangle \tag{289}$$

and the final state is pure if the initial state was. The apparent entropy of the final state in (282) is now interpreted as entanglement with the state of the black hole. The 'lost' information is encoded in the black-hole state, and this information could in principle be extracted if we measured all properties of this ball of string [150].

In practice, we do not know how to recover this information from macroscopic black holes, so they appear to us as mixed states. What about microscopic black holes, namely fluctuations in the space-time background with $\Delta E = O(m_P)$, that last for a period $\Delta t = O(1/m_P)$ and have a size $\Delta x = O(1/m_P)$? Do these steal information from us, or do they give it back to us when they decay? Most people think there is no microscopic leakage of information in this way, but not all of us [151] are convinced. The neutral kaon system is among the most sensitive experimental areas for testing this speculative possibility.

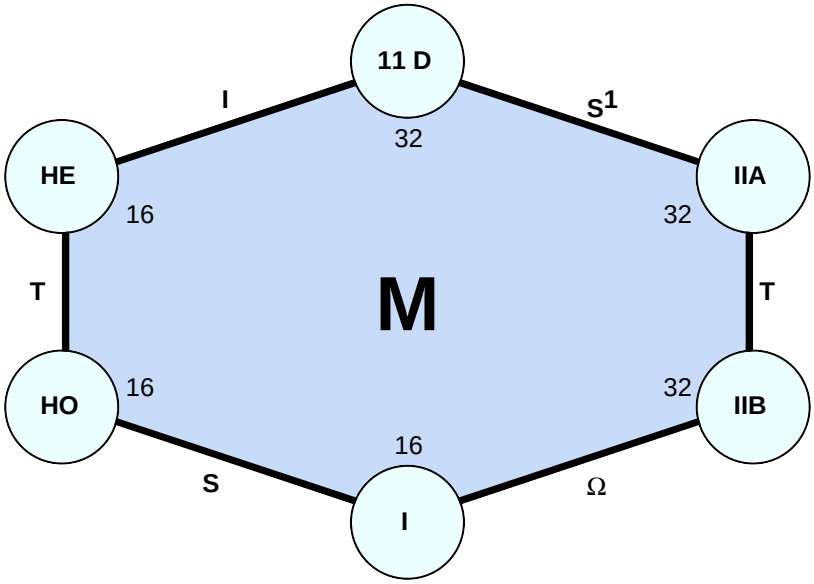


Fig. 32: The different limits of the M theory are joined by different duality relations. The numbers 16 and 32 are the numbers of spinor components in the theory.

How large might the extra dimension be in M theory? Remember that the naïve string unification scale (287) is about 20 times larger than m_{GUT} as inferred from LEP data. If one wants to maintain consistency of LEP data with supersymmetric GUTs, it seems that the extra dimension may be relatively large, with size $L_{11} \gg 1/m_{GUT} \simeq 1/10^{16}$ GeV $\gg 1/m_P$ [152]. This may be traced to the fact that the gravitational interaction strength, although growing rapidly as a power of energy

$$\sigma_G \sim E^2/m_P^4, \tag{290}$$

is still much smaller than the gauge coupling strength at $E = m_{GUT}$. However, if an extra space-time dimension appears at an energy $E < m_{GUT}$, the gravitational interaction strength grows faster, as indicated in Fig. 33. Unification with gravity around 10^{16} GeV then becomes possible, *if* the gauge couplings do not also acquire a similar higher-dimensional kick. Thus we are led to the startling capacitor-plate framework for fundamental physics shown in Fig. 34.

Each capacitor plate is *a priori* ten-dimensional, and the bulk space between them is *a priori* eleven-dimensional. Six dimensions are compactified on a scale $L_6 \sim 1/m_{GUT}$, leaving a theory which is effectively five-dimensional in the bulk and four-dimensional on the walls. Conventional gauge interactions and observable matter particles are hypothesized to live on one capacitor plate, and there are other hidden gauge interactions and matter particles living on the other plate. The fifth dimension has a characteristic size which is estimated to be $\mathcal{O}(10^{12}$ to 10^{13} GeV$)^{-1}$. Physics at smaller energies (large distances) looks effectively four-dimensional, whereas gravitational physics at larger energies (smaller distances) looks five-dimensional, and the strength of the gravitational coupling rises rapidly to unify with the gauge couplings. Supersymmetry breaking is expected to originate on the hidden capacitor plate in this scenario, and to be transmitted to the observable wall by gravitational-strength interactions in the bulk.

The phenomenological richness of this speculative M-theory approach is only beginning to be explored, and it remains to be seen whether it offers a realistic phenomenological description. However, it does embody all the available theoretical wisdom as well as offering the prospect of unifying all the observable gauge interactions with gravity at a single effective scale $\sim m_{GUT}$, including the interactions of the Standard Model. As such, it constitutes our best contemporary guess about the Theory of Everything within and beyond the Standard Model.

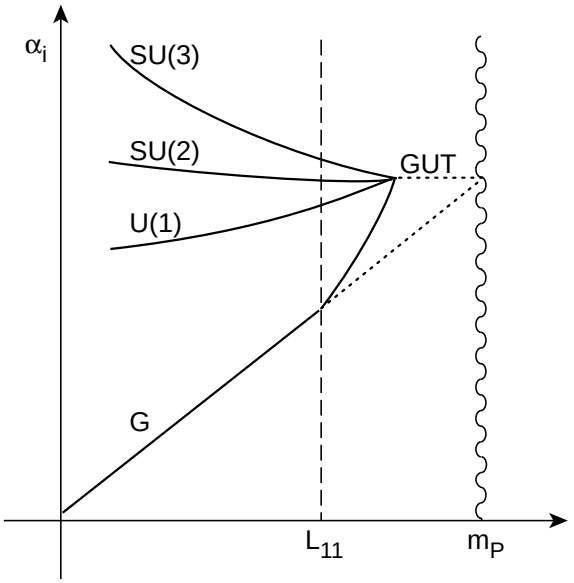


Fig. 33: Sketch of the possible evolution of the gauge couplings and the gravitational coupling G: if there is a large fifth dimension with size $\gg m_{GUT}^{-1}$, G may be unified with the gauge couplings at the GUT scale [152]

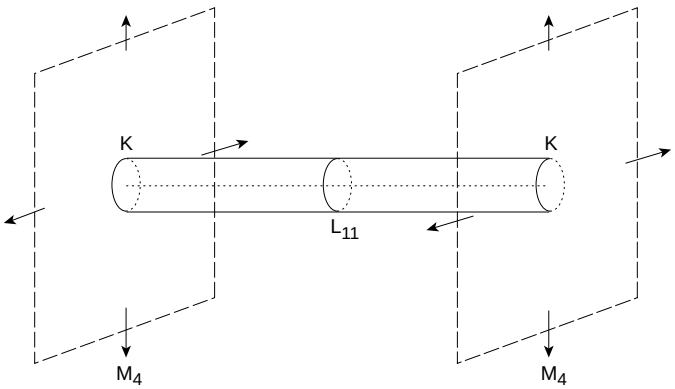


Fig. 34: The capacitor-plate scenario favoured in eleven-dimensional M theory. The eleventh dimension has a size $L_{11} \gg M_{GUT}^{-1}$, whereas dimensions $5, ..., 10$ are compactified on a small manifold K with characteristic size $\sim M_{GUT}^{-1}$. The remaining four dimensions form (approximately) a flat Minkowski space M_4 [152].

4.4 Extra dimensions

We have seen that string theories suggest that there may be extra unseen dimensions of space, but this speculation did not originate with string theorists. The idea of extra dimensions was first developed by Kaluza [69] and Klein [70]. They noticed that gravitational and electromagnetic interactions, being so alike in many ways, could be descendants of a common ancestor. Indeed, if we formulate a theory with extra spatial dimensions, it is possible to unify gravity and electromagnetism. In the same way, non-Abelian gauge fields can be unified with Einstein's gravity in more complicated models with extra dimensions. Thus, the first reason why extra dimensions were studied was to unify the gravitational and gauge interactions. These initial discussions concerned gravitation at the classical level. If you want to quantize gravity, you would be well advised to look at the best available candidate, namely string or M-theory, which, as we have seen, can be formulated consistently in a space with six or seven extra dimensions. From this point of view, the quantization of gravitational interactions becomes a second reason for extra dimensions.

In all the scenarios considered above, the extra dimensions were very small, close to the Planck size or perhaps somewhat larger, but undetectable in conceivable experiments.

However, it was suggested by Antoniadis [153] that an extra dimension might be a good way to break supersymmetry, in which case its size would be $\sim 1/$ TeV, in which case it might have some observable manifestations at the LHC.

Another suggestion, discussed in Lecture 2, was the possibility that boundary conditions in an extra dimension might be used to break the electroweak gauge symmetry. In this case also, the size of the extra dimension should be $\sim 1/$ TeV, and potentially detectable at the LHC [66–68].

Arkani-Hamed, Dimopoulos and Dvali (ADD) [154] went even further, observing that the Higgs mass hierarchy problem might be addressed in models with large extra dimensions, if they were of a millimetre or micron in size. Because the extra dimensions are so large in the ADD framework, their effects might be measurable even in low-energy table-top experiments. These models can be embedded in string theory framework, as discussed in Ref. [155]. The main ingredients of the simplest ADD scenario are [156]:

- The particles of the SM live on a 3-brane, while gravity spreads to all 4+N dimensions;
- There is a new fundamental scale of gravity in extra dimensions, M_*, which together with the ultraviolet completion scale of the SM is around a few TeV or so, thus eliminating the Higgs mass hierarchy problem;
- N extra dimensions are compactified.

If we define in this context the 4-dimensional Planck mass

$$M_{Pl}^2 = M_*^{2+N}(2\pi L)^N, \tag{291}$$

and postulate that the quantum gravity scale $M_* \sim$ TeV, we can estimate the size of the extra dimensions to be

$$L \sim 10^{-17+30/N}\text{cm}\ . \tag{292}$$

For one extra dimension, $N = 1$, we obtain $L \sim 10^{13}$ cm, which is excluded within the ADD framework, because gravity would have become higher-dimensional at distances $\sim 10^{13}$ cm. On the other hand, for $N = 2$ we get $L \sim 10^{-2}$ cm. This case is very interesting, because it predicts a modification of the 4-dimensional laws of gravity at submillimeter distances — which has become the subject of active experimental studies [156]. For larger N, the value of L should decrease but, even for $N = 6$, L is very large compared to $1/M_P$.

Randall and Sundrum (RS) went much further still [157], showing that a model with an *infinite* warped extra dimension could provide an attractive way to reformulate the hierarchy problem. In this scenario, 4-dimensional gravity on a brane is obtained through the phenomenon of localization of gravity. The brane is embedded in a 5-dimension bulk space with negative cosmological constant. In this case we find a relation between the 4-dimensional Planck mass and M_*

$$M_{Pl}^2 = M_*^3(2L). \tag{293}$$

This is similar to the relation between the fundamental scale M_*, the size L of the extra dimension, and the Planck mass M_P in the ADD model with one extra dimension (291). This similarity is based on the fact that in both theories the effective size of the extra dimension that is felt by the zero-mode graviton is finite and $\sim L$.

So, are extra dimensions very small, small, large or infinite, and how do we tell? There are several ways to search for extra dimensions in experiments at the TeV scale at the LHC.

Typical examples in theories with TeV-scale extra dimensions are the appearance of Kaluza–Klein excitations, corresponding to particle wave functions that wrap themselves around the extra dimension.

These show up as resonances that can appear in cross sections at specific energies related to the compactification scale. These Kaluza–Klein excitations occur in 'towers' that can be understood by analogy with a quantum-mechanical particle in a potential well. Its energy is quantized due to the boundary conditions at the walls of the well. In our case, the supplementary dimension plays the role of the wall of the well.

In models with very large extra dimensions, there are many Kaluza–Klein excitations of the graviton, which may be detectable *via* missing-energy events.

Another speculative possibility is the creation of a microscopic black hole [158]. Any concentration of energy or mass m will be transformed into a black hole if it is squeezed below its Schwarzschild radius: G/m. The larger the mass, the easier it can be squeezed below its Schwarzschild radius. Moreover, as we have seen, extra dimensions can increase the value of G. Hence, if there are a few extra dimensions of sufficient size, it is conceivable that collisions in the LHC might squeeze a pair of partons below their combined Schwarzschild radius, and hence create a microscopic black hole. These should evaporate rapidly, since Hawking radiation implies that the black hole loses energy at a rate inversely proportional to its mass. Studies performed by the CMS [28] and ATLAS [29] collaborations have demonstrated that such Hawking radiation would be visible in the LHC *via* energetic jets, leptons and photons, as well as missing energy carried away by neutrinos. See Fig. 35 for some results for simulated black hole production at the LHC [159].

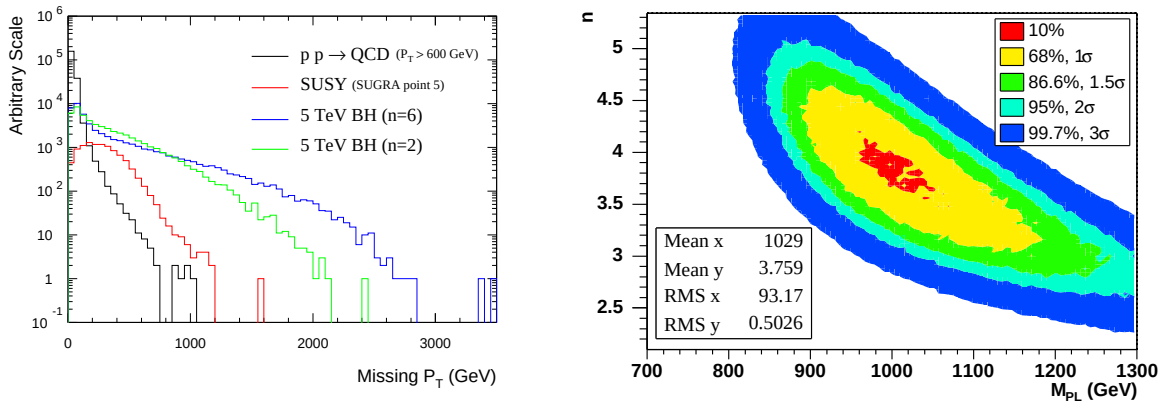


Fig. 35: Left: a comparison of the missing transverse momentum spectra in the SM, in a typical supersymmetric model, and in two black hole scenarios, and right: the results of a fit to the number of extra dimensions n and the higher-dimensional Planck mass M_{PL} on the basis of simulated black hole production at the LHC, taken from Ref. [159].

4.5 And now for something completely different?

In 1982, Prime Minister Thatcher of the United Kingdom visited CERN: I was placed in the receiving line, and introduced as a theoretical physicist. "So what do theoretical physicists *do*?" she boomed. I replied that "We think of things for the experimentalists to look for, and we hope they find something different". Mrs Thatcher was not sure about this, and asked "Wouldn't it be better if they found what you had predicted?" My response was that "In that case, we would not be learning anything new." In the same spirit, let us hope that new experiments, particularly at the LHC, will soon reveal new physics beyond the Standard Model. Perhaps it will look something like the possibilities discussed in these Lectures, but let us hope that it will take us beyond the beyonds imagined by theorists.

References

[1] P. Q. Hung and C. Quigg, Science **210**, 1205 (1980).

[2] S. Weinberg, Int. J. Mod. Phys. A **23**, 1627 (2008).

[3] C. Quigg, arXiv:0905.3187 [hep-ph].

[4] F. J. Hasert *et al.* [Gargamelle Collaboration], Phys. Lett. B **46**, 121 (1973); Phys. Lett. B **46**, 138 (1973).

[5] G. Arnison *et al.* [UA1 Collaboration], Phys. Lett. B **122**, 103 (1983).

[6] M. Banner *et al.* [UA2 Collaboration], Phys. Lett. B **122**, 476 (1983).

[7] P. Bagnaia *et al.* [UA2 Collaboration], Phys. Lett. B **129**, 130 (1983).

[8] C. Rubbia, Rev. Mod. Phys. **57**, 699 (1985).

[9] J. R. Ellis, *Beyond the Standard Model for Hill Walkers*, arXiv:hep-ph/9812235.

[10] J. R. Ellis, *Supersymmetry for Alp Hikers*, arXiv:hep-ph/0203114.

[11] J. Welzel, D. Gherson and J. R. Ellis, *Nouvelles Physiques des Particules*, arXiv:hep-ph/0506163.

[12] C. Quigg, *Gauge Theories of the Strong, Weak, and Electromagnetic Interactions* (Addison-Wesley, Reading, MA, 1997).

[13] C. Amsler *et al.* [Particle Data Group], Phys. Lett. B **667**, 1 (2008).

[14] D. D. Ryutov, Plasma Phys. Control. Fusion **49**, B429 (2007).

[15] G. 't Hooft, Phys. Rev. Lett. **37**, 8 (1976); Phys. Rev. D **14**, 3432 (1976) [Erratum-ibid. D **18**, 2199 (1978)].

[16] ALEPH, CDF, D0, DELPHI, L3, OPAL and SLD Collaborations, LEP and Tevatron Electroweak Working Groups, SLD Electroweak and Heavy Flavour Groups, arXiv:0911.2604.

[17] H. Flacher, M. Goebel, J. Haller, A. Hocker, K. Moenig and J. Stelzer, Eur. Phys. J. C **60**, 543 (2009) [arXiv:0811.0009 [hep-ph]].

[18] J. R. Ellis and G. L. Fogli, Phys. Lett. B **231**, 189 (1989).

[19] J. R. Ellis, G. L. Fogli and E. Lisi, Phys. Lett. B **274**, 456 (1992).

[20] S. Schael *et al.* [ALEPH, DELPHI, L3, OPAL Collaborations and LEP Working Group for Higgs Boson Searches], Eur. Phys. J. C **47** (2006) 547 [arXiv:hep-ex/0602042].

[21] B. Kayser, in the *Proceedings of 32nd SLAC Summer Institute on Particle Physics (SSI 2004): Nature's Greatest Puzzles*, Menlo Park, CA, 2004, pp. L004 [arXiv:hep-ph/0506165].

[22] R. N. Mohapatra *et al.*, Rep. Prog. Phys. **70**, 1757 (2007) [arXiv:hep-ph/0510213].

[23] B. W. Lee, C. Quigg and H. B. Thacker, Phys. Rev. Lett. **38**, 883 (1977); Phys. Rev. D **16**, 1519 (1977).

[24] Tevatron New Phenomena & Higgs Working Group, arXiv:0911.3930;
`http://tevnphwg.fnal.gov/results/SM_Higgs_Fall_09/`.

[25] J. Ellis, J. R. Espinosa, G. F. Giudice, A. Hoecker and A. Riotto, Phys. Lett. B **679**, 369 (2009) [arXiv:0906.0954 [hep-ph]].

[26] J. R. Ellis and D. Ross, Phys. Lett. B **506**, 331 (2001) [arXiv:hep-ph/0012067].

[27] A. Djouadi, Phys. Rep. **457**, 1 (2008) [arXiv:hep-ph/0503172].

[28] G. L. Bayatian *et al.* [CMS Collaboration], J. Phys. G **34**, 995 (2007).

[29] G. Aad *et al.* [The ATLAS Collaboration], arXiv:0901.0512 [hep-ex].

[30] A. G. Riess *et al.* [Supernova Search Team Collaboration], Astron. J. **116**, 1009 (1998) [arXiv:astro-ph/9805201];
S. Perlmutter *et al.* [Supernova Cosmology Project Collaboration], Astrophys. J. **517**, 565 (1999) [arXiv:astro-ph/9812133].

[31] A. Dobado and A. L. Maroto, Astrophys. Space Sci. **320**, 167 (2009) [arXiv:0802.1873 [astro-ph]].

[32] A. Harvey, Eur. J. Phys. **30**, 877 (2009).

[33] J. Dunkley *et al.* [WMAP Collaboration], Astrophys. J. Suppl. **180**, 306 (2009) [arXiv:0803.0586 [astro-ph]];
E. Komatsu *et al.* [WMAP Collaboration], Astrophys. J. Suppl. **180**, 330 (2009) [arXiv:0803.0547 [astro-ph]].

[34] J. Ellis, J.S. Hagelin, D.V. Nanopoulos, K.A. Olive and M. Srednicki, Nucl. Phys. B **238**, 453 (1984);
see also H. Goldberg, Phys. Rev. Lett. **50**, 1419 (1983).

[35] D. N. Spergel *et al.* [WMAP Collaboration], Astrophys. J. Suppl. **170**, 377 (2007) [arXiv:astro-ph/0603449].

[36] A. D. Sakharov, Pisma Zh. Eksp. Teor. Fiz. **5**, 32 (1967) [JETP Lett. **5**, 24 (1967)].

[37] J. M. Cline, arXiv:hep-ph/0609145.

[38] A. Pilaftsis, J. Phys. Conf. Ser. **171**, 012017 (2009) [arXiv:0904.1182 [hep-ph]].

[39] P. A. M. Dirac, Proc. Roy. Soc. Lond. A **133**, 60 (1931).

[40] P. A. M. Dirac, Phys. Rev. **74**, 817 (1948).

[41] P. A. M. Dirac, in *Proceedings Orbis Scientiae: New Pathways in High-Energy Physics*, Coral Gables, 1976, Vol. I, A. Perlmutter (ed.) (Plenum, New York, 1976), pp. 1–14.

[42] B. Cabrera, Phys. Rev. Lett. **48**, 1378 (1982).

[43] J. H. Schwarz and N. Seiberg, Rev. Mod. Phys. **71**, S112 (1999) [arXiv:hep-th/9803179].

[44] A. Ashtekar, Nuovo Cim. **122B**, 135 (2007) [arXiv:gr-qc/0702030].

[45] K. D. Lane, arXiv:hep-ph/9401324.

[46] J. F. Gunion, H. E. Haber, G. Kane, and S. Dawson, *The Higgs Hunter's Guide* (Perseus Publishing, New York,1990).

[47] M. S. Chanowitz, Phys. Rev. D **66**, 073002 (2002) [arXiv:hep-ph/0207123].

[48] G. P. Zeller *et al.* [NuTeV Collaboration], Phys. Rev. Lett. **88**, 091802 (2002) [Erratum-ibid. **90**, 239902 (2003)] [arXiv:hep-ex/0110059].

[49] R. Barbieri and A. Strumia, arXiv:hep-ph/0007265.

[50] For a review of little Higgs models, see M. Schmaltz and D. Tucker-Smith, Annu. Rev. Nucl. Part. Sci. **55**, 229 (2005) [arXiv:hep-ph/0502182].

[51] N. Arkani-Hamed, A. G. Cohen and H. Georgi, Phys. Lett. B **513**, 232 (2001) [arXiv:hep-ph/0105239].

[52] H. C. Cheng and I. Low, JHEP **0408**, 061 (2004) [arXiv:hep-ph/0405243].

[53] M. Perelstein, Prog. Part. Nucl. Phys. **58**, 247 (2007) [arXiv:hep-ph/0512128].

[54] E. Farhi and L. Susskind, Phys. Rept. **74**, 277 (1981).

[55] C. T. Hill and E. H. Simmons, Phys. Rep. **381**, 235 (2003) [Erratum-ibid. **390**, 553 (2004)] [arXiv:hep-ph/0203079].

[56] S. Weinberg, Phys. Rev. D **19**, 1277 (1979).

[57] L. Susskind, Phys. Rev. D **20**, 2619 (1979).

[58] A. Martin, arXiv:0812.1841 [hep-ph].

[59] J. Bardeen, L. N. Cooper and J. R. Schrieffer, Phys. Rev. **108**, 1175 (1957).

[60] T. Appelquist, M. Piai and R. Shrock, Phys. Rev. D **69**, 015002 (2004) [arXiv:hep-ph/0308061].

[61] S. Dimopoulos and J. R. Ellis, Nucl. Phys. B **182**, 505 (1982).

[62] J. R. Ellis, M. K. Gaillard, D. V. Nanopoulos and P. Sikivie, Nucl. Phys. B **182**, 529 (1981).

[63] J. R. Ellis, G. L. Fogli and E. Lisi, Phys. Lett. B **343**, 282 (1995).

[64] M. T. Frandsen, arXiv:0710.4333 [hep-ph].

[65] R. Contino, C. Grojean, M. Moretti, F. Piccinini and R. Rattazzi, in preparation.

[66] S. K. Rai, Int. J. Mod. Phys. A **23**, 823 (2008) [arXiv:hep-ph/0510339].

[67] R. Barbieri, G. Marandella and M. Papucci, Phys. Rev. D **66**, 095003 (2002) [arXiv:hep-ph/0205280].

[68] J. F. Gunion and B. Grzadkowski, arXiv:hep-ph/0004058.

[69] Th. Kaluza, Sitzungsber. Preuss. Akad. Wiss. Phys. Math. Klasse 996 (1921); Reprinted with an English translation in *Modern Kaluza–Klein Theories*, eds. T. Appelquist, A. Chodos and P.G.O. Freund (Addison-Wesley, Menlo Park, 1987).

[70] O. Klein, Z. Phys. **37**, 895 (1926); Reprinted with an English translation in *Modern Kaluza–Klein Theories*, eds. T. Appelquist, A. Chodos and P.G.O. Freund (Addison-Wesley, Menlo Park, 1987).

[71] For a review, see: Miao Li, hep-th/9811019.

[72] H. C. Cheng, arXiv:0710.3407 [hep-ph].

[73] S. R. Coleman and J. Mandula, Phys. Rev. **159**, 1251 (1967).

[74] Yu. A. Golfand and E. P. Likhtman, JETP Lett. **13**, 323 (1971) [Pisma Zh. Eksp. Teor. Fiz. **13**, 452 (1971)].

[75] P. Ramond, Phys. Rev. D **3**, 2415 (1971).

[76] A. Neveu and J. H. Schwarz, Nucl. Phys. B **31**, 86 (1971).

[77] D. V. Volkov and V. P. Akulov, Phys. Lett. B **46**, 109 (1973).

[78] J. Wess and B. Zumino, Phys. Lett. B **49**, 52 (1974); Nucl. Phys. B **70**, 39 (1974).

[79] J. Wess and B. Zumino, Nucl. Phys. B **78**, 1 (1974).

[80] J. Iliopoulos and B. Zumino, Nucl. Phys. **B76**, 310 (1974).

[81] S. Ferrara, J. Iliopoulos and B. Zumino, Nucl. Phys. **77**, 413 (1974).

[82] D. Z. Freedman, P. van Nieuwenhuizen and S. Ferrara, Phys. Rev. **D13**, 3214 (1976).

[83] S. Deser and B. Zumino, Phys. Lett. **62B**, 335 (1976).

[84] P. Fayet and S. Ferrara, Phys. Rep. **32**, 249 (1977).

[85] H. P. Nilles, Phys. Rep. **110**, 1 (1984).

[86] H. E. Haber and G. L. Kane, Phys. Rep **117**, 75 (1985).

[87] S. P. Martin, *A Supersymmetry Primer*, arXiv:hep-ph/9709356.

[88] Y. Okada, M. Yamaguchi and T. Yanagida, Prog. Theor. Phys. **85**, 1 (1991);
J. R. Ellis, G. Ridolfi and F. Zwirner, Phys. Lett. B **257**, 83 (1991);
H. E. Haber and R. Hempfling, Phys. Rev. Lett. **66**, 1815 (1991).

[89] J. Ellis, S. Kelley and D.V. Nanopoulos, Phys. Lett. **260**, 131 (1991);
U. Amaldi, W. de Boer and H. Furstenau, Phys. Lett. **B260**, 447 (1991);
P. Langacker and M. Luo, Phys. Rev. **D44**, 817 (1991);
C. Giunti, C. W. Kim and U. W. Lee, Mod. Phys. Lett. A **6**, 1745 (1991).

[90] M. B. Green, J. H. Schwarz and E. Witten, *Superstring Theory* (Cambridge Univ. Press, 1987).

[91] H. N. Brown *et al.* [Muon g-2 Collaboration], Phys. Rev. Lett. **86**, 2227 (2001) [arXiv:hep-ex/0102017].

[92] M. Davier, A. Hoecker, B. Malaescu, C. Z. Yuan and Z. Zhang, arXiv:0908.4300 [hep-ph].

[93] J. R. Ellis, J. F. Gunion, H. E. Haber, L. Roszkowski and F. Zwirner, Phys. Rev. D **39**, 844 (1989).

[94] S. Dimopoulos and H. Georgi, Nucl. Phys. **B193**, 150 (1981).

[95] R. Barbieri *et al.*, arXiv:hep-ph/0406039.

[96] J. R. Ellis, J. S. Lee and A. Pilaftsis, Phys. Rev. D **76**, 115011 (2007) [arXiv:0708.2079 [hep-ph]].

[97] M. Carena, J. R. Ellis, A. Pilaftsis and C. E. Wagner, Nucl. Phys. B **586**, 92 (2000) [arXiv:hep-ph/0003180], Phys. Lett. B **495**, 155 (2000) [arXiv:hep-ph/0009212]; and references therein.

[98] J. R. Ellis, K. A. Olive, Y. Santoso and V. C. Spanos, Phys. Lett. B **565**, 176 (2003) [arXiv:hep-

ph/0303043].

[99] O. Buchmueller *et al.*, arXiv:0907.5568 [hep-ph].

[100] T. Hahn, S. Heinemeyer, W. Hollik, H. Rzehak and G. Weiglein, Comput. Phys. Commun. **180**, 1426 (2009).

[101] O. Buchmueller *et al.*, JHEP **0809**, 117 (2008) [arXiv:0808.4128 [hep-ph]].

[102] H. Georgi, H. Quinn and S. Weinberg, Phys. Rev. Lett. **33**, 451 (1974).

[103] J. Ellis and D.V. Nanopoulos, Nature **292**, 436 (1981).

[104] M. Chanowitz, J. Ellis and M. K. Gaillard, Nucl. Phys. **B128**, 506 (1977).

[105] A. J. Buras, J. Ellis, M. K. Gaillard and D. V. Nanopoulos, Nucl. Phys. **B135**, 66 (1978).

[106] D. V. Nanopoulos and D. A. Ross, Phys. Lett. **118B**, 99 (1982).

[107] S. Dimopoulos and H. Georgi [94];
S. Dimopoulos, S. Raby and F. Wilczek, Phys. Rev. **D24**, 1681 (1981);
L. Ibàñez and G. G. Ross, Phys. Lett. **105B**, 439 (1981).

[108] J. Ellis, S. Kelley and D. V. Nanopoulos, Phys. Lett. **B249**, 441 (1990).

[109] J. Ellis, S. Kelley and D. V. Nanopoulos, Nucl. Phys. **B373**, 55 (1992).

[110] P. Langacker and N. Polonsky, Phys. Rev. **D47**, 4028 (1993).

[111] F. Anselmo, L. Cifarelli, A. Peterman and A. Zichichi, Nuovo Cimento **104A**, 1817 (1991);
F. Anselmo, L. Cifarelli, A. Peterman and A. Zichichi, Nuovo Cimento **105A**, 1210 (1992).

[112] R. Barbieri and L. J. Hall, Phys. Rev. Lett. **68**, 752 (1992);
J. Hisano, T. Moroi, K. Tobe and T. Yanagida, Phys. Lett. **B342**, 138 (1995).

[113] H. Georgi and S. L. Glashow, Phys. Rev. Lett. **32**, 438 (1974).

[114] J. Ellis and M. K. Gaillard, Phys. Lett. **88B**, 315 (1979).

[115] H. Fritzsch and P. Minkowski, Ann. Phys. (N.Y.) **93**, 193 (1975).

[116] J. Ellis, M. K. Gaillard and D. V. Nanopoulos, Phys. Lett. **91B**, 67 (1980).

[117] H. Nishino *et al.* [Super-Kamiokande Collaboration], Phys. Rev. Lett. **102**, 141801 (2009) [arXiv:0903.0676 [hep-ex]].

[118] J. Ellis, D. V. Nanopoulos and S. Rudaz, Nucl. Phys. **B202**, 43 (1982);
S. Dimopoulos, S. Raby and F. Wilczek, Phys. Lett. **112B**, 133 (1982).

[119] S. Weinberg, Phys. Rev. **D26**, 287 (1982),
N. Sakai and T. Yanagida, Nucl. Phys. **B197**, 533 (1982).

[120] K. Kobayashi *et al.* [Super-Kamiokande Collaboration], Phys. Rev. D **72**, 052007 (2005) [arXiv:hep-ex/0502026].

[121] J. Ellis, D. V. Nanopoulos and S. Rudaz, Nucl. Phys. **B202**, 43 (1982).

[122] I. Antoniadis, J. Ellis, J. S. Hagelin and D. V. Nanopoulos, Phys. Lett. **B194**, 231 (1987) and **B231**, 65 (1989).

[123] J. Ellis, J. S. Hagelin, S. Kelley and D. V. Nanopoulos, Nucl. Phys. **B311**, 1 (1988).

[124] B. A. Campbell, J. Ellis and S. Rudaz, Phys. Lett. **141B**, 229 (1984).

[125] T. Yanagida, *Proc. Workshop on Unified Theories and Baryon Number in the Universe*, Tsukuba, Japan, 1979 (KEK, Japan, 1979, report KEK-79-18);
R. Slansky, Talk at the *Sanibel Symposium*, Palm Coast, FL, USA, 1979, Caltech preprint CALT-68-709 (1979).

[126] J. R. Ellis and O. Lebedev, Phys. Lett. B **653**, 411 (2007) [arXiv:0707.3419 [hep-ph]].

[127] Y. Fukuda *et al.* [Super-Kamiokande Collaboration], Phys. Rev. Lett. **81**, 1562 (1998).

[128] Q. R. Ahmad *et al.* [SNO Collaboration], Phys. Rev. Lett. **89**, 011301 (2002) [arXiv:nucl-ex/0204008].

[129] M. Fukugita and T. Yanagida, Phys. Lett. B **174**, 45 (1986).

[130] J. R. Ellis and M. Raidal, Nucl. Phys. B **643**, 229 (2002) [arXiv:hep-ph/0206174].

[131] J. Polonyi, Hungary Central Inst. Res., KFKI-77-93;
E. Cremmer, B. Julia, J. Scherk, S. Ferrara, L. Girardello and P. Van Nieuwenhuizen, Nucl. Phys. **B147**, 105 (1979).

[132] E. Cremmer, S. Ferrara, C. Kounnas and D.V. Nanopoulos, Phys. Lett. **133B**, 61 (1983).

[133] J. Ellis, A.B. Lahanas, D.V. Nanopoulos and K.A. Tamvakis, Phys. Lett. **134B**, 429 (1984).

[134] J. R. Ellis, K. A. Olive, Y. Santoso and V. C. Spanos, Phys. Rev. D **70**, 055005 (2004) [arXiv:hep-ph/0405110].

[135] A. Strumia, Phys. Lett. **B409**, 213 (1997).

[136] J. R. Ellis, K. A. Olive and P. Sandick, Phys. Lett. B **642**, 389 (2006) [arXiv:hep-ph/0607002]; JHEP **0706**, 079 (2007) [arXiv:0704.3446 [hep-ph]]; JHEP **0808**, 013 (2008) [arXiv:0801.1651 [hep-ph]].

[137] J. Alexandre, J. Ellis and N. E. Mavromatos, arXiv:0901.2532 [hep-th].

[138] J. Bekenstein, Phys. Rev. **D12**, 3077 (1975);
S. Hawking, Commun. Math. Phys. **43**, 199 (1975).

[139] A. Strominger, arXiv:0906.1313 [hep-th].

[140] G. Veneziano, Nuovo Cimento **57A**, 190 (1968) and Phys. Rep. **C9**, 199 (1974).

[141] J. H. Schwarz, arXiv:0708.1917 [hep-th].

[142] M. A. Virasoro, Phys. Rev. **177**, 2309 (1969).

[143] Y. Nambu, *Proc. Int. Conf. on Symmetries and Quark Models*, Wayne State University, Detroit, MI, USA, 1969 (Gordon and Breach, New York, 1970), p. 269;
P. Goddard, J. Goldstone, C.Rebbi and C. Thorn, Nucl. Phys. **B181**, 502 (1981).

[144] J. Scherk and J. H. Schwarz, Nucl. Phys. B **81**, 118 (1974).

[145] M.B. Green and J.H. Schwarz, Phys. Lett. **149B**, 117 (1984) and **151B**, 21 (1985).

[146] H. Dreiner, J. Lopez, D. V. Nanopoulos and D. B. Reiss, Phys. Lett. **B216**, 283 (1989).

[147] J. Polchinski, Phys. Rev. Lett. **75**, 4724 (1995) [arXiv:hep-th/9510017].

[148] N. Seiberg and E. Witten, Nucl. Phys. B **426**, 19 (1994) [Erratum-ibid. B **430**, 485 (1994)] [arXiv:hep-th/9407087].

[149] C. M. Hull and P. K. Townsend, Nucl. Phys. B **438**, 109 (1995) [arXiv:hep-th/9410167].

[150] For a recent take on this, see S. B. Giddings, arXiv:0911.3395 [hep-th].

[151] J. Ellis, N. E. Mavromatos and D. V. Nanopoulos, Mod. Phys. Lett. **A10**, 425 (1995) and references therein.

[152] P. Horava and E. Witten, Nucl. Phys. **B460**, 506 (1996) and Nucl. Phys. **B475**, 94 (1996);
P. Horava, Phys. Rev. **D54**, 7561 (1996).

[153] I. Antoniadis, Phys. Lett. B **246**, 377 (1990).

[154] N. Arkani-Hamed, S. Dimopoulos and G. R. Dvali, Phys. Lett. B **429**, 263 (1998) [arXiv:hep-ph/9803315]; N. Arkani-Hamed, S. Dimopoulos and G. R. Dvali, Phys. Rev. D **59**, 086004 (1999) [arXiv:hep-ph/9807344].

[155] I. Antoniadis, N. Arkani-Hamed, S. Dimopoulos and G. R. Dvali, Phys. Lett. B **436**, 257 (1998) [arXiv:hep-ph/9804398].

[156] G. Gabadadze, arXiv:hep-ph/0308112.

[157] L. Randall and R. Sundrum, Phys. Rev. Lett. **83**, 4690 (1999) [arXiv:hep-th/9906064].

[158] S. B. Giddings and S. D. Thomas, Phys. Rev. D **65**, 056010 (2002) [arXiv:hep-ph/0106219];
S. Dimopoulos and G. L. Landsberg, Phys. Rev. Lett. **87**, 161602 (2001) [arXiv:hep-ph/0106295].

[159] C. M. Harris, M. J. Palmer, M. A. Parker, P. Richardson, A. Sabetfakhri and B. R. Webber, JHEP **0505**, 053 (2005) [arXiv:hep-ph/0411022].

Neutrino physics

P. Hernández
IFIC, Universidad de València and CSIC, E-46071 Valencia, Spain

Abstract
The topics discussed in this lecture include: general properties of neutrinos in the SM, the theory of neutrino masses and mixings (Dirac and Majorana), neutrino oscillations both in vacuum and in matter, an overview of the experimental evidence for neutrino masses and of the prospects in neutrino oscillation physics. We also briefly review the relevance of neutrinos in leptogenesis and in beyond-the-Standard-Model physics.

1 Neutrinos in the Standard Model

LEP era established the validity of the Standard Model (SM) with an accuracy below the per cent level. The SM is based on the gauge group $SU(3) \times SU(2) \times U_Y(1)$ that is spontaneously broken to the subgroup $SU(3)_{color} \times U(1)_{em}$. All the fermions of the SM fall into irreducible representations of this group with the quantum numbers summarized in Table 1 [1].

Neutrinos are the most elusive particles of this table. They do not carry electromagnetic or colour charge, but only the weak charge under the spontaneously broken subgroup. For this reason they are extremely weakly interacting, since their interactions are mediated by massive gauge bosons.

The history of neutrinos goes back to W. Pauli who postulated the existence of the electron neutrino in an attempt to restore energy–momentum conservation in β decay, but he did so with great regret: *I have done a terrible thing, I have postulated a particle that cannot be detected*. Fortunately Pauli was wrong, not only have neutrinos been detected but they have been extremely useful in establishing the two most striking features of Table 1: the left–handedness of the weak interactions (the left–right asymmetry of the table) and the family structure (the three–fold repetition of the same representations).

In the SM only the left-handed fields carry the $SU(2)$ charge, where by left-handed we denote the negative chirality component (i.e., eigenstate of γ_5 with eigenvalue minus one) of the fermion field [1]:

$$\Psi = \Psi_R + \Psi_L = \underbrace{\left(\frac{1+\gamma_5}{2}\right)}_{P_R}\Psi + \underbrace{\left(\frac{1-\gamma_5}{2}\right)}_{P_L}\Psi \,. \tag{1}$$

For relativistic fermions (i.e., massless), it is easy to see that the chiral projectors are equivalent to the projectors on helicity components:

$$P_{R,L} = \frac{1}{2}\left(1 \pm \frac{\mathbf{s}\cdot\mathbf{p}}{|p|}\right) + O\left(\frac{m_i}{E}\right), \tag{2}$$

where the helicity operator $\mathbf{\Sigma} = \frac{\mathbf{s}\cdot\mathbf{p}}{|p|}$ measures the component of the spin in the direction of the spatial momentum. Therefore for massless fermions only the left-handed states (with the spin pointing in the opposite direction to the momentum) carry $SU(2)$ charge. This is not inconsistent with Lorenzt invariance, since for a fermion travelling at the speed of light, the helicity is the same in any reference frame. In other words, the helicity operator commutes with the Hamiltonian for a massless fermion and is thus a good quantum number.

The discrete symmetry under CPT (charge conjugation, parity, and time reversal), which is a basic building block of any Lorenzt invariant and unitary field theory, requires that for any left-handed fermion, there exists a right-handed antiparticle, with opposite charge, but the right-handed particle state may not

Table 1: Irreducible fermionic representations in the Standard Model: $(I_{SU(3)}, I_{SU(2)})_Y$

$(\mathbf{1},\mathbf{2})_{-\frac{1}{2}}$	$(\mathbf{3},\mathbf{2})_{-\frac{1}{6}}$	$(\mathbf{1},\mathbf{1})_{-1}$	$(\mathbf{3},\mathbf{1})_{-\frac{2}{3}}$	$(\mathbf{3},\mathbf{1})_{-\frac{1}{3}}$
$\begin{pmatrix}\nu_e\\ e\end{pmatrix}_L$	$\begin{pmatrix}u^i\\ d^i\end{pmatrix}_L$	e_R	u^i_R	d^i_R
$\begin{pmatrix}\nu_\mu\\ \mu\end{pmatrix}_L$	$\begin{pmatrix}c^i\\ s^i\end{pmatrix}_L$	μ_R	c^i_R	s^i_R
$\begin{pmatrix}\nu_\tau\\ \tau\end{pmatrix}_L$	$\begin{pmatrix}t^i\\ b^i\end{pmatrix}_L$	τ_R	t^i_R	b^i_R

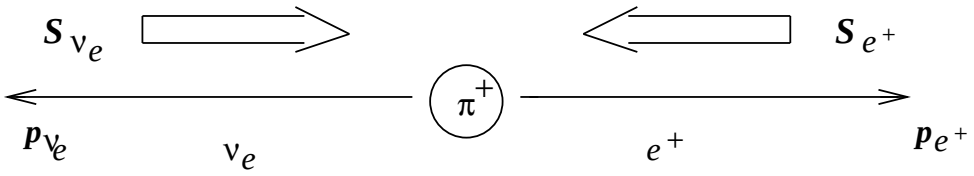


Fig. 1: Kinematics of pion decay

exist. This is precisely what happens with neutrinos in the SM. Since only the left-handed states carry charge and their masses were compatible with zero when the SM was established, they were postulated to be Weyl fermions: i.e., a left-handed particle and a right-handed antiparticle.

Under parity, a left-handed particle state transforms into a right-handed particle state, thus the left-handedness of the weak interactions implies a maximal violation of parity, which is nowhere more obvious than in the neutrino sector, where the reflection of a SM neutrino in a mirror is nothing.

The weak current is therefore $V-A$ since it only couples to the left fields: $\bar{\Psi}_L\gamma_\mu\Psi_L = \bar{\Psi}\gamma_\mu(1-\gamma_5)/2\Psi$. This structure is clearly seen in the kinematics of weak decays involving neutrinos, such as the classic example of pion decay to $e\nu_e$ or $\mu\nu_\mu$. In the limit of vanishing electron or muon mass, this decay is forbidden, because the spin of the initial state is zero and thus it is impossible to conserve simultaneously momentum and angular momentum if the two recoiling particles must have opposite helicities, as shown in Fig. 1. Thus the ratio of the decay rates to electrons and muons, in spite of the larger phase space in the former, is strongly suppressed by the factor $\left(\frac{m_e}{m_\mu}\right)^2 \sim 2\times 10^{-5}$.

Another profound consequence of the chiral nature of the weak interaction is anomaly cancellation. The chiral coupling of fermions to gauge fields leads generically to inconsistent gauge theories due to chiral anomalies: if any of the diagrams depicted in Fig. 2 is non-vanishing, the weak current is conserved at tree level but not at one loop, implying a catastrophic breaking of gauge invariance. Anomaly cancellation is the requirement that all these diagrams vanish, which imposes strong constraints on the hypercharge assignments of the fermions in the SM, which are *miraculously* satisfied:

$$\overbrace{\sum_{i=\text{quarks}} Y_i^L - Y_i^R}^{GGB} = \overbrace{\sum_{i=\text{doublets}} Y_i^L}^{WWB} = \overbrace{\sum_i Y_i^L - Y_i^R}^{Bgg} = \overbrace{\sum_i (Y_i^L)^3 - (Y_i^R)^3}^{B^3} = 0, \tag{3}$$

where $Y_i^{L/R}$ are the hypercharges of the left/right components of the fermionic field i, and the triangle

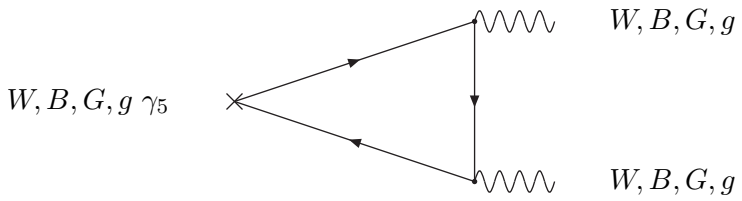


Fig. 2: Triangle diagrams that can give rise to anomalies. W, B, G are the gauge bosons associated to the $SU(2), U_Y(1), SU(3)$ gauge groups, respectively, and g is the graviton

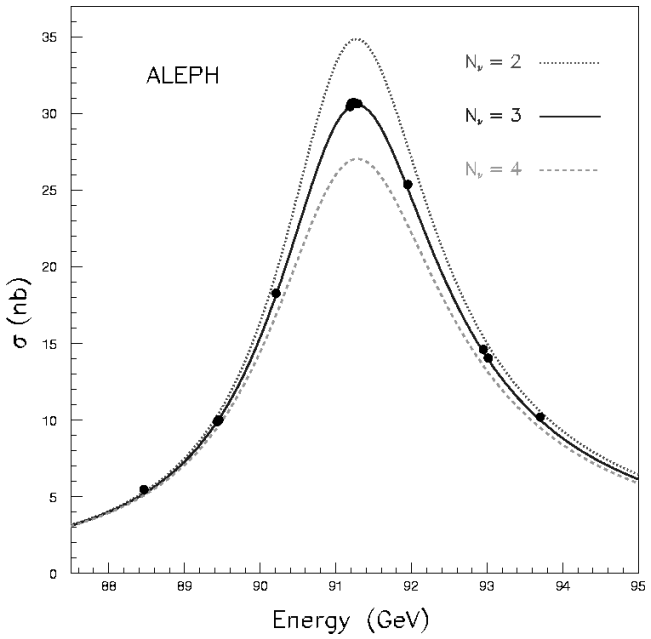


Fig. 3: Z^0 resonance from the ALEPH experiment at LEP. Data are compared to the case of $N_\nu = 2, 3$ and 4

diagram corresponding to each of the sums is indicated above the bracket.

Concerning the family structure, we know, thanks to neutrinos, that there are exactly three families in the SM. An extra SM family with quarks and charged leptons so heavy that they remain unobserved, would also have massless neutrinos that would have been produced in Z^0 decay, modifying its width, which has been measured at LEP with an impressive precision, as shown in Fig. 3. This measurement excludes any number of standard neutrino families different from three [2]:

$$N_\nu = 2.984 \pm 0.008. \tag{4}$$

2 Neutrino masses and mixings

When the SM was invented, there were only upper limits on the neutrino masses so these were conjectured to be zero. The direct limit on neutrino masses comes from the precise measurement of the end-point of the lepton energy spectrum in weak decays, which gets modified if neutrinos are massive. In particular the most stringent limit is obtained from tritium β-decay for the electron neutrino:

$$H^3 \rightarrow^3 He + e^- + \bar{\nu}_e. \tag{5}$$

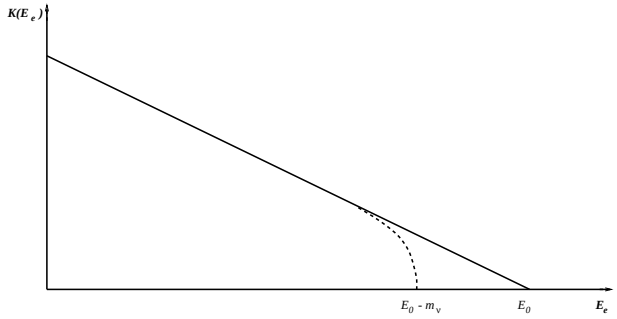


Fig. 4: Effect of a neutrino mass in the end-point of the lepton energy spectrum in β decay

Figure 4 shows the effect of a neutrino mass in the end-point electron energy spectrum in this decay. The functional form of this curve is $K(E_e) \propto \sqrt{(E_0 - E_e)((E_0 - E_e)^2 - m_\nu^2)^{1/2}}$. The best limit has been obtained by the Mainz and Troitsk experiments [3]:

$$m_{\nu_e} < 2.2 \text{ eV (Mainz)}, \quad m_{\nu_e} < 2.1 \text{ eV (Troitsk)}\,, \tag{6}$$

both at 95% CL. The direct limits on the other two neutrino masses are much weaker. The best limit on the ν_μ mass ($m_{\nu_\mu} < 170$ keV [4]) was obtained from the end-point spectrum of the decay $\pi^+ \to \mu^+ \nu_\mu$, while that on the ν_τ mass was obtained at LEP ($m_{\nu_\tau} < 18.2$ MeV [5]) from the decay $\tau \to 5\pi\nu_\tau$.

As we shall see, there is now strong evidence that neutrinos are indeed massive, although extremely light, below the stringent bound of Eq. (6).

Neutrino masses can be easily accommodated in the SM. A massive fermion necessarily has two states of helicity, since it is always possible to reverse the helicity of a state that moves at a slower speed than light by looking at it from a boosted reference frame. In fact a mass can be thought of as the strength of the coupling between the two helicity states:

$$m\ \overline{\psi_L}\psi_R + \text{h.c.} \tag{7}$$

In order to include such a coupling in the SM for the neutrinos we need to identify the neutrino right-handed state, which in the SM is absent. It turns out there are two ways to proceed:

Dirac massive neutrinos

We can enlarge the SM by adding a set of three right-handed neutrino states, which would be singlets under $SU(3)\times SU(2)\times U_Y(1)$, but coupled to matter just through the neutrino masses. This coupling has to be of the Yukawa type to preserve the gauge symmetry in such a way that the masses are proportional to the vacuum expectation value of the Higgs field, v, exactly like for the remaining fermions [1]:

$$\lambda_\nu\ \overline{L_L}\ \tilde{\Phi}\ \nu_R + \text{ h.c. } \to m_\nu = \lambda_\nu\ v, \tag{8}$$

where $L_L = (\nu_L\ l_L)$ is the lepton doublet and $\tilde{\Phi}$ is the scalar doublet that gets a vacuum expectation value $\langle\tilde{\Phi}\rangle = (v\ 0)$. There are two important consequences of proceeding in this way. Firstly there is a new hierarchy problem in the SM to be explained: why neutrinos are much lighter than the remaining leptons, even those in the same family (see Fig. 5). Secondly, lepton number, L, which counts the number of leptons minus that of antileptons, remains an exactly conserved global symmetry at the classical level [1], just as baryon number, B, is.

[1] As usual $B + L$ is broken by the anomaly and only $B - L$ remains exact at all orders.

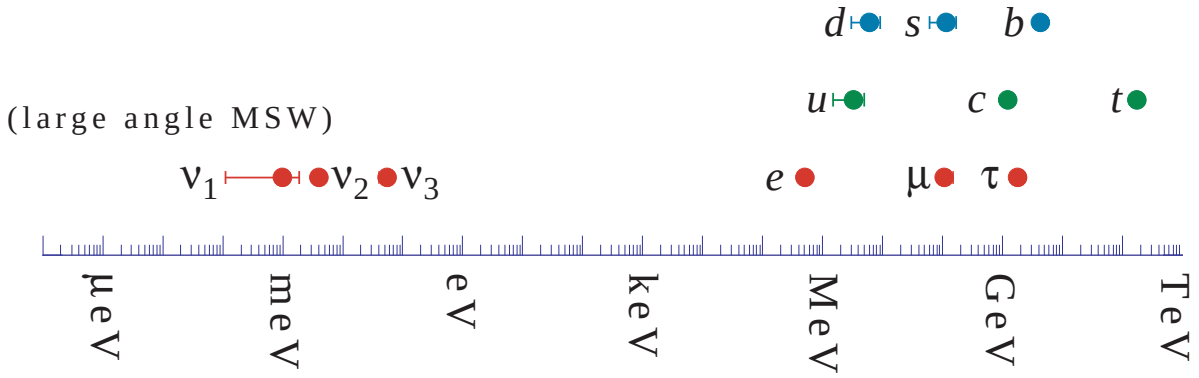


Fig. 5: Fermion spectrum in the Standard Model

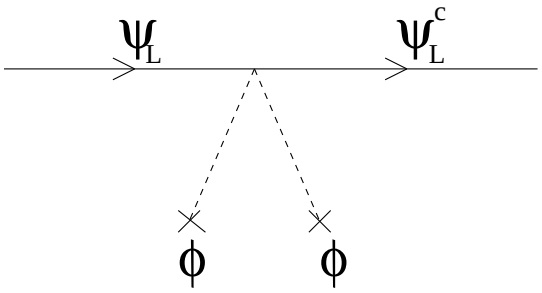


Fig. 6: Majorana coupling of the light neutrinos to the Higgs field

Majorana massive neutrinos

For neutral particles, Majorana realized that one can get rid of half of the degrees of freedom in a massive Dirac spinor in a Lorenzt-invariant way by identifying the right-handed state with the antiparticle of the left-handed state:

$$\nu_R \to (\nu_L)^c = C\bar{\nu}_L^T = C\gamma_0\nu_L^*, \tag{9}$$

where C is the operator of charge conjugation in spinor space.

Neutrinos are the only particles for which this possibility is compatible with charge conservation, because they are charged only under the spontaneously broken subgroup of the SM and thus a Majorana mass term can be written in a gauge invariant way by including two Higgs fields, as shown in Fig. 6:

$$\frac{1}{M} L_L^T C\ \alpha_\nu \tilde{\Phi}^T\ \tilde{\Phi}\ L_L + \text{h.c.}\ , \tag{10}$$

where an energy scale, M, has been introduced for dimensional reasons, so that the coupling α_ν is adimensional. Upon spontaneous symmetry breaking, these couplings become Majorana neutrino masses of the form

$$m_\nu = \alpha_\nu \frac{v^2}{M}. \tag{11}$$

If the scale M is much higher than the electroweak scale v, a strong hierarchy between the neutrino and the charged lepton masses arises naturally.

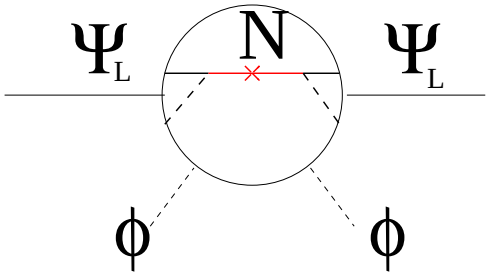


Fig. 7: Neutrino masses in the see-saw model

2.1 See-saw models

It is interesting to consider the simplest example to explain the origin of the scale M in the Majorana masses. This is the famous see-saw model of Gell-Mann, Ramond, Slansky, and Yanagida [6]. In this model, the Majorana effective interaction of Eq. (10) results from the interchange of very heavy right-handed Majorana neutrinos, as depicted in Fig. 7. The SM Lagrangian is enlarged with the terms

$$\delta\mathcal{L}_Y^\nu \quad = \quad \bar{L}_L\tilde{\lambda}_\nu\ \tilde{\Phi}\ N_R + \frac{1}{2}N_R^T C\ M_R\ N_R + \text{ h.c. }, \tag{12}$$

that is a Yukawa coupling of the lepton doublet and the heavy singlets plus a Majorana mass term for the singlets. Upon spontaneous symmetry breaking these couplings become mass terms:

$$\delta\mathcal{L}_Y^\nu \quad \to \quad \frac{1}{2}\left(\nu_L^T \quad {N_R}^T\right) C \begin{pmatrix} 0 & \tilde{\lambda}_\nu v \\ \tilde{\lambda}_\nu^T v & M_R \end{pmatrix} \begin{pmatrix} \nu_L \\ N_R \end{pmatrix} . \tag{13}$$

When $v \ll M_R$, the diagonalization of the mass matrix can be done in perturbation theory:

$$\mathcal{M} = \mathcal{M}^{(0)} + \mathcal{M}^{(1)} \equiv \begin{pmatrix} 0 & 0 \\ 0 & M_R \end{pmatrix} + \begin{pmatrix} 0 & \tilde{\lambda}_\nu v \\ \tilde{\lambda}_\nu^T v & 0 \end{pmatrix} . \tag{14}$$

To second order we find:

$$U^T\mathcal{M}U = \begin{pmatrix} -v^2\tilde{\lambda}_\nu\frac{1}{M_R}\tilde{\lambda}_\nu^T & 0 \\ 0 & M_R \end{pmatrix} \quad U = \begin{pmatrix} 1 & \tilde{\lambda}_\nu\frac{v}{M_R} \\ -\frac{v}{M_R}\tilde{\lambda}_\nu^T & 1 \end{pmatrix} . \tag{15}$$

There are three light Majorana neutrinos ($\nu_L' \simeq \nu_L + \tilde{\lambda}_\nu\frac{v}{M_R}N_R$) with a mass matrix:

$$v^2\tilde{\lambda}_\nu\frac{1}{M_R}\tilde{\lambda}_\nu^T, \tag{16}$$

and three heavy ones ($N_R' \simeq N_R - \frac{v}{M_R}\tilde{\lambda}_\nu^T\nu_L$) with the mass matrix M_R.

Equivalently we say that the heavy Majoranas can be integrated out leaving a trace of higher dimensional operators:

$$\mathcal{L}_{eff}^{d=5} = \frac{1}{2}L_L^T C\ \ \tilde{\Phi}^T\ \left(\tilde{\lambda}_\nu\frac{1}{M_R}\tilde{\lambda}_\nu^T\right)\ \tilde{\Phi}\ L_L \tag{17}$$

$$\mathcal{L}_{eff}^{d=6} = \mathcal{O}\left(\frac{1}{M_R^2}\right)... \tag{18}$$

The one with lowest dimension is the one we obtained from symmetry arguments in Eq. (10).

A few observations are in place:

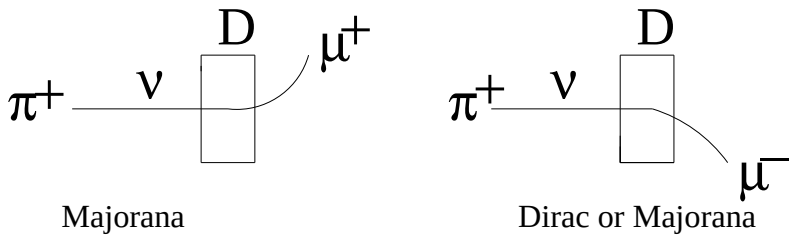


Fig. 8: A neutrino beam from π^+ decay (ν_μ) could interact in the magnetized detector producing a μ^+ only if neutrinos are Majorana.

– The new physics scale M in Eq. (10) is simply related to the masses of the heavy Majorana neutrinos and the Yukawa couplings:

$$\frac{\alpha_\nu}{M} \to \tilde{\lambda}_\nu \frac{1}{M_R} \tilde{\lambda}_\nu^T. \tag{19}$$

As we shall see, data imply there is at least one $m_\nu \geq 0.05$ eV. If $\tilde{\lambda}_\nu \sim O(1)$ then:

$$v \; < \; M_R \sim 10^{15} \text{ GeV} < M_{\text{Planck}}, \tag{20}$$

and the masses are close to the typical Grand Unification (GUT) scale.

– In order to give non-vanishing masses to all the three left-handed neutrinos, the number of Majorana singlets must satisfy $N_R \geq N_L = 3$. The reason is that the matrix $\underbrace{\tilde{\lambda}_\nu}_{N_L \times N_R} \underbrace{\frac{1}{M_R}}_{N_R \times N_R} \underbrace{\tilde{\lambda}_\nu^T}_{N_R \times N_L}$ has $N_L - N_R$ zero modes.

2.2 Majorana versus Dirac

The consequences of the SM neutrinos being massive Majorana particles are profound:

– A new physics scale M must exist and is accessible in an indirect way through neutrino masses.
– Lepton number is not conserved: a Majorana mass violates the conservation of all the charges carried by the fermion, including the global charges such as lepton number. As we shall see in Section 6, the dynamics associated to the scale M could be responsible for the generation of the baryon asymmetry in the Universe.
– The anomaly cancellation conditions fix all the hypercharges (i.e., there is only one possible choice for the hypercharges that satisfies Eq. (3)), which implies that electromagnetic charge quantization is the only possibility in a field theory with the same matter content as the SM.

It is clear that establishing the Majorana nature of neutrinos is of great importance. In principle there are very clear signatures, such as the one depicted in Fig. 8, where a ν_μ beam from π^+ decay is intercepted by a detector. In the Dirac case, the interaction of neutrinos on the detector via a charged current interaction will produce a μ^- in the final state. If neutrinos are Majorana, a wrong-sign muon in the final state is also possible. Unfortunately the rate for μ^+ production is suppressed by m_ν/E in amplitude with respect to the μ^-. For example, for $E_\nu = \mathcal{O}(1)$ GeV and $m_\nu \sim \mathcal{O}(1)$ eV the cross–section for this process will be roughly 10^{-18} times the usual CC neutrino cross-section, which means it is impossible to detect.

The best hope of observing a rare process of this type seems to be the search for neutrinoless double–beta decay ($2\beta0\nu$), the right diagram of Fig. 9. The background to this process is the standard

Experiment	Nucleus	$\lvert m_{ee}\rvert$
Heidelberg-Moscow I	^{76}Ge	< 0.34–1.1 eV$(90\%$ CL) [9]
Heidelberg-Moscow II	^{76}Ge	0.2–0.6 eV [10]
CUORICINO	^{120}Te	< 0.2–1.1 eV$(90\%$ CL) [11]
NEMO-3	^{100}Mo	< 0.6–2 eV$(90\%$ CL) [12]

Table 2: Present bounds from various neutrinoless double-beta-decay experiments

double–beta decay depicted on the left of Fig. 9, which has been observed to take place with a lifetime of $T_{2\beta 2\nu} > 10^{19}$–$10^{21}$ years.

If the source of L violation is just the Majorana ν mass, the inverse lifetime for this process is given by

$$T^{-1}_{2\beta 0\nu} \simeq \underbrace{G^{0\nu}}_{\text{Phase}} \underbrace{\left|M^{0\nu}\right|^2}_{\text{NuclearM.E.}} \underbrace{\left|\sum_i \left(V^{ei}_{\text{MNS}}\right)^2 m_i\right|^2}_{|m_{ee}|^2}, \tag{21}$$

where m_{ee} is the 11 entry in the neutrino mass matrix in the flavour basis. In spite of the suppression in the neutrino mass (over the energy of this proccess), the neutrinoless mode has a larger phase factor than the 2ν mode, and as a result the lifetime is expected to be of the order

$$T^{-1}_{2\beta 0\nu} \sim \left(\frac{m_\nu}{E}\right)^2 10^9 \; T^{-1}_{2\beta 2\nu}, \tag{22}$$

which could be observable for neutrino masses in the eV range. Several experiments have set stringent upper bounds on $|m_{ee}|$ and there is even a controversial positive signal, as shown in Table 2.

Fig. 9: 2β decay: normal (left) and neutrinoless (right)

2.3 Neutrino mixing

Generically, neutrino masses imply neutrino mixing [7, 8], because the Yukawa couplings need not be flavour diagonal:

$$\mathcal{L}^{\text{Dirac}}_m = \overline{\nu^i_L}\,(\lambda_\nu v)_{ij}\,\nu^j_R + \text{ h.c.} \tag{23}$$

$$\mathcal{L}^{\text{Majorana}}_m = \frac{1}{2}\frac{v^2}{M}\nu^{i\,T}_L C\,(\alpha_\nu)_{ij}\,\nu^j_L + \text{ h.c.} \tag{24}$$

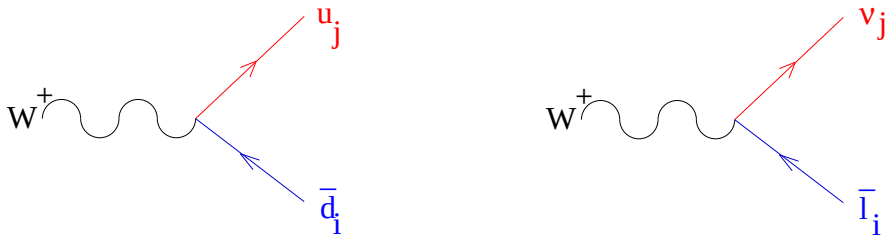


Fig. 10: Quark and lepton mixing

Instead, in the mass eigenbasis for all the leptons, the charged weak couplings are not diagonal, in complete analogy with the quark flavour sector (see Fig. 10):

$$\mathcal{L}^{\text{Dirac}} = \bar{l}^i_L \gamma_\mu W^+_\mu V^{ij}_{\text{MNS}} \nu^j_L + \frac{1}{2}\overline{\nu^i_L}\gamma_\mu Z_\mu \nu^i_L + \overline{\nu^i_L} m_i \nu^i_R + \text{ h.c.} \tag{25}$$

$$\mathcal{L}^{\text{Majorana}} = \bar{l}^i_L \gamma_\mu W^+_\mu \tilde{V}^{ij}_{\text{MNS}} \nu^j_L + \frac{1}{2}\overline{\nu^i_L}\gamma_\mu Z_\mu \nu^i_L + \frac{1}{2}{\nu^i_L}^T C m_i \nu^i_L + \text{ h.c.} \tag{26}$$

The number of parameters that are in principle observable in the lepton mixing matrix (V_{MNS} for Dirac and $\tilde{V}_{\text{MNS}}$ for Majorana) can easily be computed by counting the number of independent real and imaginary elements of the Yukawa matrices and eliminating those that can be absorbed in field redefinitions. The allowed field redefinitions are the unitary rotations of the fields that leave the Lagrangian invariant in the absence of lepton masses, but are not symmetries of the full Lagrangian when lepton masses are included.

In the Dirac case, it is possible to rotate independently the left-handed lepton doublet, together with the right-handed charged leptons and neutrinos, that is $U(n)^3$, for a generic number of families n. However, this includes total lepton number which remains a symmetry of the massive theory and thus cannot be used to reduce the number of physical parameters in the mass matrix. The parameters that can be absorbed in field redefinitions are thus the parameters of the group $U(n)^3/U(1)$ (that is $\frac{3(n^2-n)}{2}$ real, $\frac{3(n^2+n)-1}{2}$ imaginary).

In the case of Majorana neutrinos, there is no independent right-handed neutrino field, nor is lepton number a good symmetry. Therefore the number of field redefinitions is the number of parameters of the elements in $U(n)^2$ (that is n^2-n real and n^2+n imaginary).

The resulting real physical parameters are the mass eigenstates and the mixing angles, while the resulting imaginary parameters are CP-violating phases. All this is summarized in Table 3. Dirac and Majorana neutrinos differ only in the number of observables phases. For three families ($n=3$), there is just one Dirac phase and three in the Majorana case.

A standard parametrization of the mixing matrices is given by

$$V_{\text{MNS}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_{23} & s_{23} \\ 0 & -s_{23} & c_{23} \end{pmatrix} \begin{pmatrix} c_{13} & 0 & s_{13} \\ 0 & 1 & 0 \\ -s_{13} & 0 & c_{13} \end{pmatrix} \begin{pmatrix} c_{12} & s_{12}e^{i\delta} & 0 \\ -s_{12}e^{i\delta} & c_{12} & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{27}$$

$$\tilde{V}_{\text{MNS}} = V_{\text{MNS}}(\theta_{12}, \theta_{13}, \theta_{23}, \delta) \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{i\alpha_1} & 0 \\ 0 & 0 & e^{i\alpha_2} \end{pmatrix} . \tag{28}$$

3 Neutrino oscillations

The fact that neutrinos are such weakly interacting particles allows them to have coherence over very long distances. For example, a neutrino with an energy of $\mathcal{O}(1$ MeV) moving in lead, which has a

Table 3: Number of real and imaginary parameters in the Yukawa matrices, of those that can be absorbed in field redefinitions. The difference between the two is the number of observable parameters: the lepton masses (m), mixing angles (θ), and phases (ϕ).

	Yukawas	Field redefinitions	$No.\ m$	$No.\ \theta$	$No.\ \phi$
Dirac	$\lambda_l,\ \lambda_\nu$ $4n^2$	$U(n)^3/U(1)_L$ $\frac{3(n^2-n)}{2},\ \frac{3(n^2+n)-1}{2}$	$2n$	$\frac{n^2-n}{2}$	$\frac{(n-2)(n-1)}{2}$
Majorana	$\lambda_l,\ \alpha_\nu^T=\alpha_\nu$ $3n^2+n$	$U(n)^2$ $n^2-n,\ n^2+n$	$2n$	$\frac{n^2-n}{2}$	$\frac{n^2-n}{2}$

Fig. 11: Neutrino oscillations

density of $\rho = 7.9$ g/cm^3, has a mean free path $l \sim \frac{1}{\sigma\rho} \sim 4\times10^{16}$ metres ~ 4 light-years.

Neutrinos are necessarily produced in a flavour eigenstate, that is, in a precise combination of the mass eigenstates, which are the true eigenstates of the free Hamiltonian. After some distance L, where neutrinos have evolved freely, the mass eigenstate components in the original flavour state get different phases and, as a result, there is a non-zero probability that the flavour measured at L is a different one [7], as shown in Fig. 11.

There has been a lot of discussion about what is the rigorous way to define such a transition probability. This is not straightforward because, in quantum field theory (which is required since neutrinos are relativistic), we are used to considering processes in which there is no knowledge of the position in space or time where the interaction took place, and it is then a good approximation to consider asymptotic states that are simply plane waves, with well-defined energy–momentum. In this case this is not possible, because we must distinguish the macroscopic distance that separates the source of neutrinos and the detector. This implies that it cannot be a good approximation to consider asymptotic states of well-defined momentum at least in the direction between source and detector. This fact has often confused the derivation and even led to incorrect results.

Let us consider that neutrinos are produced as wave packets localized around the source position $x_0 = (t_0, \mathbf{x_0})$ in a flavour state α:

$$|\nu_\alpha(x)\rangle = \sum_j V_{\alpha j} \int \frac{d^3k}{(2\pi)^3} f_j(\mathbf{k}) e^{-ik_0^j(t-t_0)} e^{i\mathbf{k}(\mathbf{x}-\mathbf{x_0})} |\nu_j\rangle\ , \tag{29}$$

where ${k_0^j}^2 = \mathbf{k}^2 + m_j^2$, since the state being asymptotic must be on-shell and $V_{\alpha j}$ is the mixing matrix. The wave packets $f_j(\mathbf{k})$ depend on the production process (uncertainty in momentum of the initial states, kinematics), but we do not need to know the exact form. For example we can consider a Gaussian:

$$f_i(\mathbf{k}) \sim e^{-(\mathbf{k}-\bar{\mathbf{q}}^i)^2/(2\sigma_i^2)}\ . \tag{30}$$

We expect that, neglecting neutrino masses, the wave packets are the same for all the mass eigenstates:

$$f_i(\mathbf{k}) \sim f(\mathbf{k}) + O\left(m_i/|\mathbf{k}|\right) \sim e^{-(\mathbf{k}-\mathbf{q})^2/(2\sigma^2)} \,. \tag{31}$$

Let us forget about the proper normalization of the state for the time being. Let us consider that the neutrino produced is moving in the direction of a detector located at some distance down the beam line L in the $\hat{z}$ direction (therefore $\mathbf{q} = (0, 0, q_z)$), where we want to measure the flavour of the state in Eq. (29). The probability that we measure a state with flavour β at any point x is $\sim |\langle \nu_\beta | \nu_\alpha(x) \rangle|^2$, where

$$|\nu_\beta\rangle = \sum_j V_{\beta j} |\nu_j\rangle. \tag{32}$$

The amplitude is then

$$\langle \nu_\beta | \nu_\alpha(x) \rangle = \sum_i \, V^*_{\beta i} V_{\alpha i} \int d^3k \; f_i(\mathbf{k}) e^{-ik_0^i(t-t_0)} e^{i\mathbf{k}(\mathbf{x}-\mathbf{x_0})}. \tag{33}$$

Note that we measure neither the time of the measurement nor the spatial $\hat{x}$ and $\hat{y}$ components, so we can integrate over them:

$$P(\nu_\alpha \to \nu_\beta) \sim \int dt \; d\hat{x} \; d\hat{y} |\langle \nu_\beta | \nu_\alpha(x) \rangle|^2 = \sum_{i,j} V^*_{\beta i} V_{\alpha i} V_{\beta j} V^*_{\alpha j} \times$$
$$\int_{\mathbf{k}} \int dk'_z \; f_i(\mathbf{k}) f_j^*(\mathbf{k}') \delta \left(\sqrt{m_i^2 + k_z^2 + k_x^2 + k_y^2} - \sqrt{m_j^2 + {k'_z}^2 + k_x^2 + k_y^2} \right) \; e^{i(k_z - k'_z)L}. \tag{34}$$

Up to exponentially small terms and neglecting effects of $O(m_i/|\mathbf{k}|)$ everywhere else than in the phase factor (where they are enhanced by L), we obtain

$$P(\nu_\alpha \to \nu_\beta) \sim \sum_{i,j} V^*_{\beta i} V_{\alpha i} V_{\beta j} V^*_{\alpha j} \int_{\mathbf{k}} |f(\mathbf{k})|^2 \frac{|\mathbf{k}|}{|k_z|} e^{-i \frac{\Delta m^2_{ji} L}{2|k_z|}}, \tag{35}$$

where $\Delta m^2_{ji} = m_i^2 - m_j^2$.

Now, we have to care about the normalization. The simplest way to compute it is by requiring that the probability be one if $\alpha = \beta$ in the case of zero or equal neutrino masses (i.e., $\Delta m^2_{ji} = 0$). Doing this we finally obtain

$$P(\nu_\alpha \to \nu_\beta) = \sum_{i,j} V^*_{\beta j} V_{\alpha j} V_{\beta i} V^*_{\alpha i} \int_{\mathbf{k}} \; e^{-i \frac{\Delta m^2_{ij} L}{2|k_z|}} \frac{|\mathbf{k}|}{|k_z|} |f(\mathbf{k})|^2 / \int_{\mathbf{k}} \frac{|\mathbf{k}|}{|k_z|} |f(\mathbf{k})|^2$$
$$\simeq \sum_{i,j} V^*_{\beta j} V_{\alpha j} V_{\beta i} V^*_{\alpha i} e^{-i \frac{\Delta m^2_{ij} L}{2|q_z|}} \;, \tag{36}$$

where in the last equality we have assumed that the phase factor does not change very much in the range of momenta of the wave packet, so that it can be taken out of the integral. The probability for the flavour transition is thus a periodic function of the distance between source and detector, hence the name *neutrino oscillations* first described by Pontecorvo [7].

Defining $W^{jk}_{\alpha\beta} \equiv [V_{\alpha j} V^*_{\beta j} V^*_{\alpha k} V_{\beta k}]$ and using the unitarity of the mixing matrix, we can rewrite the probability in the more familiar way:

$$P(\nu_\alpha \;\; \to \;\; \nu_\beta) \; = \delta_{\alpha\beta} \; - 4 \sum_{k>j} \mathrm{Re}[\mathrm{W}^{\mathrm{jk}}_{\alpha\beta}] \; \sin^2 \left(\frac{\Delta \mathrm{m}^2_{\mathrm{jk}} \, \mathrm{L}}{4 \mathrm{E}_\nu} \right) \tag{37}$$

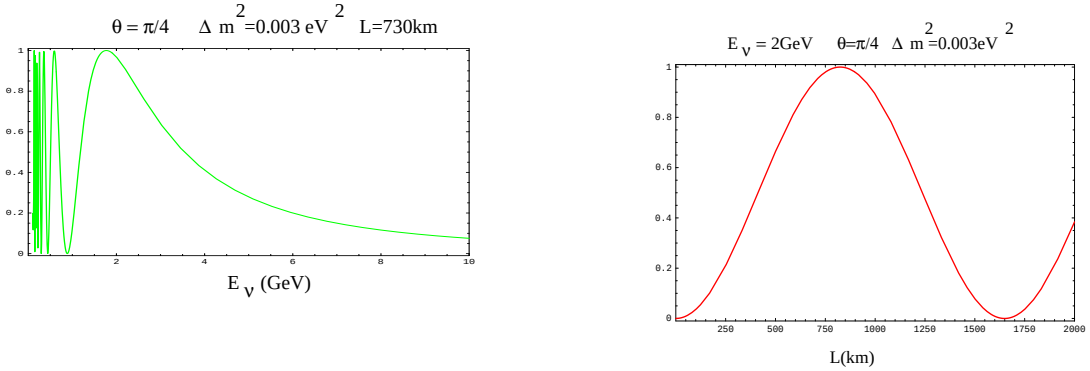


Fig. 12: Two-family oscillation probability as a function of the neutrino energy at fixed baseline of $L = 730$ km (left) and as a function of the baseline at fixed neutrino energy $E_\nu = 2$ GeV (right)

$$\pm \quad 2 \sum_{k>j} \mathrm{Im}[\mathrm{W}^{\mathrm{jk}}_{\alpha\beta}] \, \sin\left(\frac{\Delta \mathrm{m}^2_{\mathrm{jk}} \, \mathrm{L}}{2\mathrm{E}_\nu}\right), \tag{38}$$

where the $\pm$ refers to neutrinos/antineutrinos and $|\mathbf{q}| = |q_z| \simeq E_\nu$.

We refer to an *appearance* or *disappearance* oscillation probability when the initial and final flavours are different ($\alpha \neq \beta$) or the same ($\alpha = \beta$), respectively. Note that oscillation probabilities show the expected GIM suppression of any flavour changing process: they vanish if the neutrinos are degenerate.

In the simplest case of two-family mixing, the mixing matrix depends on just one mixing angle:

$$V_{\mathrm{MNS}} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} , \tag{39}$$

and there is only one mass square difference Δm^2. The oscillation probability of Eq. (38) simplifies to the well-known expression

$$P(\nu_\alpha \to \nu_\beta) = \sin^2 2\theta \, \sin^2\left(\frac{\Delta m^2 \, L}{4E_\nu}\right), \quad \alpha \neq \beta \, . \tag{40}$$

The probability is the same for neutrinos and antineutrinos because there are no imaginary entries in the mixing matrix. It is a sinusoidal function of the distance between source and detector, with a period determined by the oscillation length:

$$L_{\mathrm{osc}} \, (\mathrm{km}) = 2\pi \frac{E_\nu(\mathrm{GeV})}{1.27 \Delta m^2 (\mathrm{eV}^2)} \, , \tag{41}$$

which is proportional to the neutrino energy and inversely proportional to the neutrino mass square difference. The amplitude of the oscillation is determined by the mixing angle. It is maximal for $\sin^2 2\theta = 1$ or $\theta = \pi/4$. This oscillation probability as a function of the neutrino energy and the baseline is shown in Fig. 12

It is important to stress that there is an intrinsic limit to coherence, since the size of the wave packet is non-zero. Indeed the last equality of Eq. (36) requires that the phase factor varies slowly in the range of momenta of the wave packet. This condition is not satisfied when L becomes too large. The decoherence length, L_D, can be estimated as

$$\left| \frac{\Delta m^2_{ij} L_D}{2} \left(\frac{1}{|q_z|} - \frac{1}{|q_z| + \sigma} \right) \right| \sim 2\pi \Rightarrow L_D \sim L_{\mathrm{osc}} \frac{|q_z|}{\sigma} \, . \tag{42}$$

that is the phase factor changes by 2π when the momentum in the $\hat{z}$ direction varies within one σ from the central value, where σ is the width of the wave packet in momentum space [see Eq. (30)] When the baseline satisfies $L \gg L_D$, neutrinos do not oscillate because the phase factor averages to zero all the terms with $i \neq j$ in Eq. (36). The flavour transition probability then becomes independent of L:

$$P\left(\nu_\alpha \to \nu_\beta\right) = \sum_i |V_{\alpha i} V_{\beta i}|^2 = 2\cos^2\theta \sin^2\theta = \frac{1}{2}\sin^2 2\theta. \tag{43}$$

In practice, the smearing in L and E_ν produces the same effect. When $L \gg L_{\rm osc}$, the oscillations are so fast that any real experiment will measure the average:

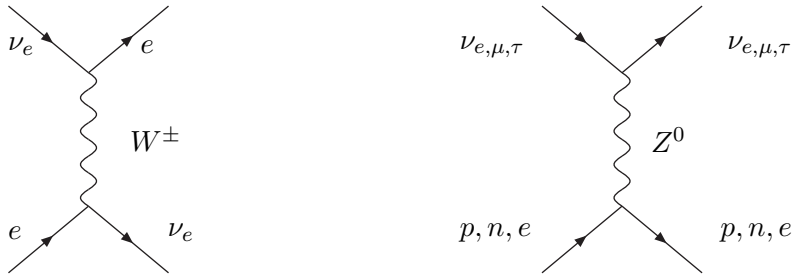
$$\langle P\left(\nu_\alpha \to \nu_\beta\right)\rangle = \frac{1}{2}\sin^2 2\theta, \tag{44}$$

which is exactly the same result as in the case of no coherence.

Note that the 'smoking gun' for neutrino oscillations is not the flavour transition, which can occur in the presence of neutrino mixing without oscillations, but the peculiar L/E_ν dependence. An idealized experiment looking for neutrino oscillations should then be able to tell flavour on one hand and should be performed at a baseline such that $L \sim L_{\rm osc}(E_\nu)$ in order to observe the oscillatory pattern, which measures the neutrino mass square difference. Note that neutrino oscillations are not sensitive to the absolute mass scale though.

3.1 Matter effects

When neutrinos propagate in matter (Earth, Sun, etc.), the amplitude for their propagation is modified owing to coherent forward scattering on electrons and nucleons [13]:



The effective Hamiltonian density resulting from the charged current interaction is

$$\mathcal{H}_{\rm CC} = \sqrt{2} G_F \left[\bar{e}\gamma_\mu P_L \nu_e\right]\left[\bar{\nu}_e \gamma^\mu P_L e\right] = \sqrt{2} G_F \left[\bar{e}\gamma_\mu P_L e\right]\left[\bar{\nu}_e \gamma^\mu P_L \nu_e\right]. \tag{45}$$

Since the medium is not polarized, the expectation value of the electron current is simply the number density of electrons:

$$\langle \bar{e}\gamma_\mu P_L e\rangle_{\rm unpol.medium} \quad = \quad \delta_{\mu 0} N_e. \tag{46}$$

Including also the neutral current interactions in the same way, the effective Hamiltonian for neutrinos in the presence of matter is

$$\mathcal{H}^{eff} = \mathcal{H}^{\rm vac} + \bar{\nu} V_m \gamma^0 (1-\gamma_5)\nu \tag{47}$$

$$V_m = \begin{pmatrix} (\frac{G_F}{\sqrt{2}}\left(N_e - \frac{N_n}{2}\right) & 0 & 0 \\ 0 & \frac{G_F}{\sqrt{2}}\left(-\frac{N_n}{2}\right) & 0 \\ 0 & 0 & \frac{G_F}{\sqrt{2}}\left(-\frac{N_n}{2}\right) \end{pmatrix}, \tag{48}$$

where N_n is the number density of neutrons. The matter potential in the center of the Sun is $V_e \sim 10^{-11}$ eV and in the Earth $V_e \sim 10^{-13}$ eV. In spite of these tiny values, these effects are non-negligible in neutrino oscillations.

The plane wave solutions to the modified Dirac equation satisfy a different dispersion relation and as a result, the phases of neutrino oscillation phenomena change. The new dispersion relation becomes

$$E - V_m - M_\nu = (\pm|\mathbf{p}| - V_m)\,\frac{1}{E + M_\nu - V_m}\,(\pm|\mathbf{p}| - V_m) \quad h = \pm, \tag{49}$$

where $h = \pm$ indicate the two helicity states and we have neglected effects of $\mathcal{O}(V M_\nu)$. This is a reasonable approximation since $m_\nu \gg V_m$. For the positive energy states we then have

$$E > 0 \quad E^2 = |\mathbf{p}|^2 + M_\nu^2 + 4EV_m \quad h = - \quad E^2 = |\mathbf{p}|^2 + M_\nu^2,\ h = +, \tag{50}$$

while for the negative energy ones $V_m \to -V_m$ and $h \to -h$.

The effect of matter can be simply accommodated in an effective mass matrix:

$$\tilde{M}_\nu^2 = M_\nu^2 \pm 4EV_m. \tag{51}$$

The effective mixing matrix $\tilde{V}_{\text{MNS}}$ is the one that takes us from the original flavour basis to that which diagonalizes this effective mass matrix:

$$\begin{pmatrix} \tilde{m}_1^2 & 0 & 0 \\ 0 & \tilde{m}_2^2 & 0 \\ 0 & 0 & \tilde{m}_3^2 \end{pmatrix} = \tilde{V}_{\text{MNS}}^\dagger \left(M_\nu^2 \pm 4E \begin{pmatrix} V_e & 0 & 0 \\ 0 & V_\mu & 0 \\ 0 & 0 & V_\tau \end{pmatrix} \right) \tilde{V}_{\text{MNS}}. \tag{52}$$

Note that the number of physical parameters is the same but the effective mixing angles and masses depend on the energy.

3.2 Neutrino oscillations in constant matter

In the case of two flavours, the effective mass and mixing angle have relatively simple expressions:

$$\sin^2 2\tilde{\theta} = \frac{\left(\Delta m^2 \sin 2\theta\right)^2}{\left(\Delta m^2 \cos 2\theta \mp 2\sqrt{2}\, G_F E\, N_e\right)^2 + \left(\Delta m^2 \sin 2\theta\right)^2} \tag{53}$$

$$\Delta\tilde{m}^2 = \sqrt{\left(\Delta m^2 \cos 2\theta \mp 2\sqrt{2} E\, G_F\, N_e\right)^2 + \left(\Delta m^2 \sin 2\theta\right)^2}, \tag{54}$$

where the sign $\mp$ corresponds to neutrinos/antineutrinos. The corresponding oscillation amplitude has a resonance [13, 14], when the neutrino energy satisfies

$$\sqrt{2}\, G_F\, N_e \mp \frac{\Delta m^2}{2E} \cos 2\theta = 0 \quad \Rightarrow \quad \sin^2 2\tilde{\theta} = 1 \quad \Delta\tilde{m}^2 = \Delta m^2 \sin 2\theta. \tag{55}$$

The oscillation amplitude is therefore maximal independently of the value of the vacuum mixing angle.

We also note that

- oscillations vanish at $\theta = 0$, because the oscillation length becomes infinite for $\theta = 0$;
- the resonance is only there for ν or $\bar{\nu}$ but not both;
- the resonance condition depends on the sign$(\Delta m^2 \cos 2\theta)$:
  resonance observed in $\nu \to$ sign$(\Delta m^2 \cos 2\theta) > 0$,
  resonance observed in $\bar{\nu} \to$ sign$(\Delta m^2 \cos 2\theta) < 0$.

3.3 Neutrino oscillations in variable matter

In the Sun the density of electrons is not constant. However, if the variation is sufficiently slow, the eigenstates of H_{eff} change slowly with the density and we can assume that the neutrino produced in a local eigenstate remains in the same eigenstate along the trajectory. This is the so-called *adiabatic approximation*.

Let us suppose that neutrinos are crossing the Sun. We consider here two-family mixing for simplicity. At any point in the trajectory, it is possible to diagonalize the Hamiltonian fixing the matter density to that at the given point. The resulting eigenstates can be written as

$$|\tilde{\nu}_1\rangle = |\nu_e\rangle \cos\tilde{\theta} - |\nu_\mu\rangle \sin\tilde{\theta}, \tag{56}$$

$$|\tilde{\nu}_2\rangle = |\nu_e\rangle \sin\tilde{\theta} + |\nu_\mu\rangle \cos\tilde{\theta}. \tag{57}$$

Neutrinos are produced close to the centre $x = 0$ where the electron density, $N_e(0)$, is very large. Let us suppose that it satisfies

$$2\sqrt{2}G_F N_e(0) \gg \Delta m^2 \cos 2\theta. \tag{58}$$

Then the diagonalization of the mass matrix at this point gives

$$\tilde{\theta} \simeq \frac{\pi}{2} \Rightarrow |\nu_e\rangle \simeq |\tilde{\nu}_2\rangle \tag{59}$$

in such a way that an electron neutrino is mostly the second mass eigenstate. When neutrinos exit the Sun, at $x = R_\odot$, the matter density falls to zero, $N_e(R_\odot) = 0$, and the local effective mixing angle is the one in vacuum, $\tilde{\theta} = \theta$. If θ is small, the eigenstate $\tilde{\nu}_2$ is mostly ν_μ according to Eq. (57).

Therefore an electron neutrino produced at $x = 0$ is mostly the eigenstate $\tilde{\nu}_2$, but this eigenstate outside the Sun is mostly ν_μ. There is maximum $\nu_e \to \nu_\mu$ conversion if the adiabatic approximation is a good one. This is the famous MSW effect [13, 14]. The evolution of the eigenstates is shown in Fig. 13: the MSW effect would occur when there is a level crossing in the absence of mixing. The conditions for this to happen are:

– *Resonant condition*: the density at the production is above the critical one

$$N_e(0) > \frac{\Delta m^2 \cos 2\theta}{2\sqrt{2}EG_F}. \tag{60}$$

– *Adiabaticity*: the splitting of the levels is large compared to energy injected in the system by the variation of $N_e(r)$. A measurement of this is given by γ which should be much larger than one:

$$\gamma = \frac{\sin^2 2\theta}{\cos 2\theta} \frac{\Delta m^2}{2E} \frac{1}{|\nabla \log N_e(r)|} > \gamma_{\min} > 1, \tag{61}$$

where $\nabla = \partial/\partial r$.

At fixed energy both conditions give the famous MSW triangles, if plotted on the plane $(\log(\sin^2 2\theta), \log(\Delta n$

$$\log\left(\Delta m^2\right) < \log\left(\frac{2\sqrt{2}G_F N_e(0)E}{\cos 2\theta}\right) \tag{62}$$

$$\log\left(\Delta m^2\right) > \log\left(\gamma_{\min} 2E\nabla \log N_e \frac{\cos 2\theta}{\sin^2 2\theta}\right). \tag{63}$$

For example, taking $N_e(r) = N_c \exp(-r/R_0), R_0 = R_\odot/10.54, N_c = 1.6 \times 10^{26}$ cm$^{-3}, E = 1$ MeV, these curves are shown in Fig. 14.

As we shall see, the deficit of electron neutrinos coming from the Sun has been interpreted in terms of an MSW effect in neutrino propagation in the Sun. Before the recent experiments SNO and KamLAND that we shall discuss in Section 4.1, there were several solutions possible inside the expected MSW triangle: SMA, LMA and LOW as shown in Fig. 15. The famous SMA and LOW solutions are now history.

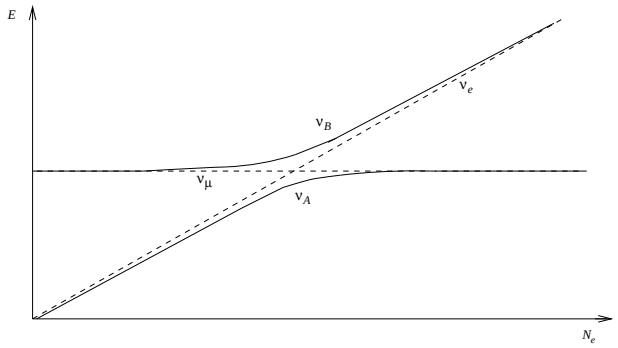


Fig. 13: Evolution of the eigenstates as a function of the distance to the centre of the Sun

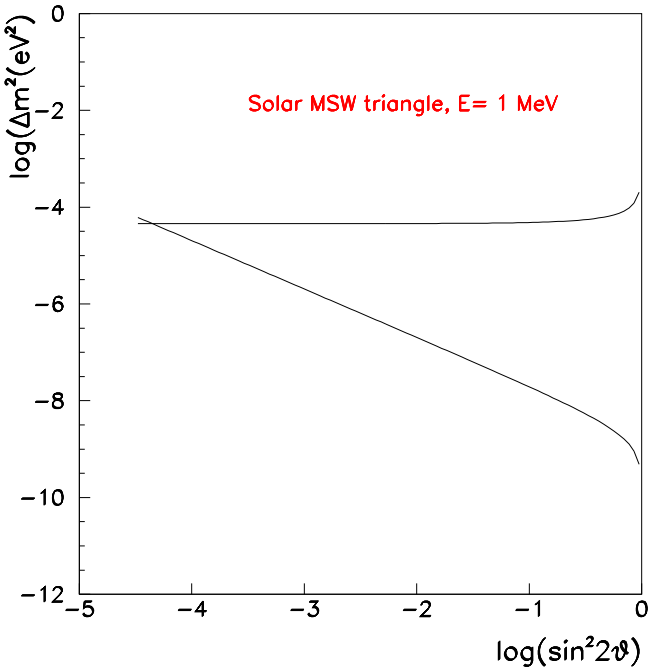


Fig. 14: MSW triangle: in the region between the two lines the resonance and adiabaticity conditions are both satisfied for neutrinos of energy 1 MeV

4 Evidence for neutrino oscillations

Nature has been kind enough to provide us with two natural sources of neutrinos (the Sun and the atmosphere) where neutrino flavour transitions have been observed in a series of ingenious experiments, that started back in the 1960s with the pioneering experiment of R. Davies. This effort has already been rewarded once with the Nobel prize of 2002.

4.1 The solar puzzle

The Sun is an intense source of neutrinos produced in the chain of nuclear reactions that burn hydrogen into helium:

$$4p \longrightarrow {}^4\mathrm{He} + 2e^+ + 2\nu_e. \tag{64}$$

The expected spectral flux of ν_e in the absence of oscillations is shown in Fig. 16. The prediction of this flux obtained by J. Bahcall and collaborators [16] is the result of a detailed simulation of the solar

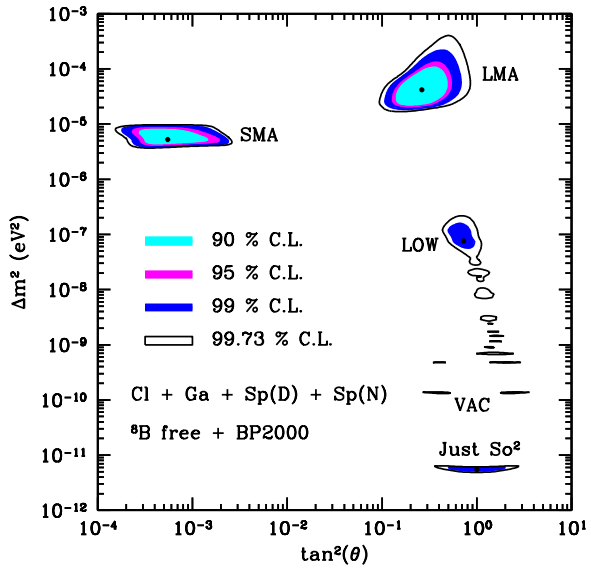


Fig. 15: Neutrino oscillation solutions to the solar neutrino deficit in year 2000 (taken from Ref. [15])

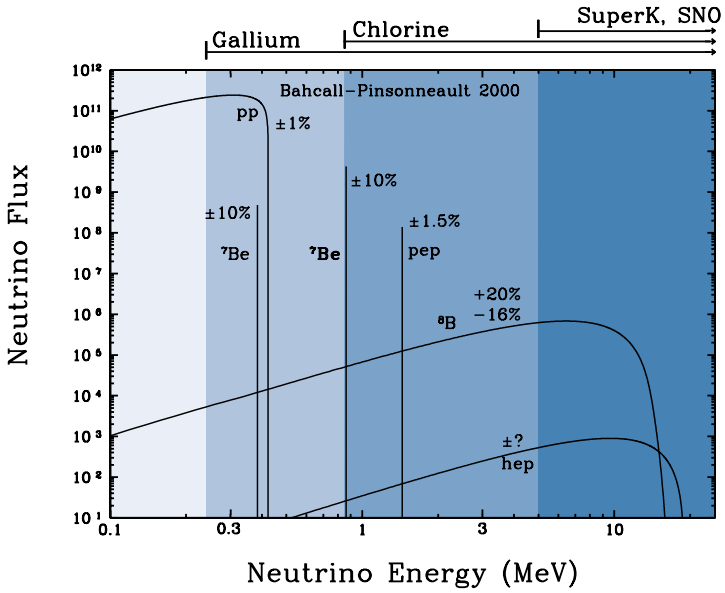


Fig. 16: Spectrum of solar neutrinos. The different bands indicate the threshold of the different detection techniques.

interior and has been improved over many years. It is the so-called standard solar model (SSM).

Neutrinos coming from the Sun have been detected with several experimental techniques that have a different neutrino energy threshold as indicated in Fig. 16. On the one hand, the radiochemical techniques, used in the experiments Homestake (chlorine, ^{37}Cl)[17], Gallex/GNO [18] and Sage [19] (using gallium, ^{71}Ga, and germanium, ^{71}Ge, respectively), can count the total number of neutrinos with a rather low threshold ($E_\nu > 0.81$ MeV in Homestake and $E_\nu > 0.23$ MeV in Gallex and Sage), they cannot get any information on the directionality, the energy of the neutrinos, nor the time of the event. On the other hand, Kamiokande [20] pioneered a new technique to observe solar neutrinos using water Cherenkov detectors. The signal comes from elastic neutrino scattering on electrons (ES), $\nu_e + e^- \rightarrow$

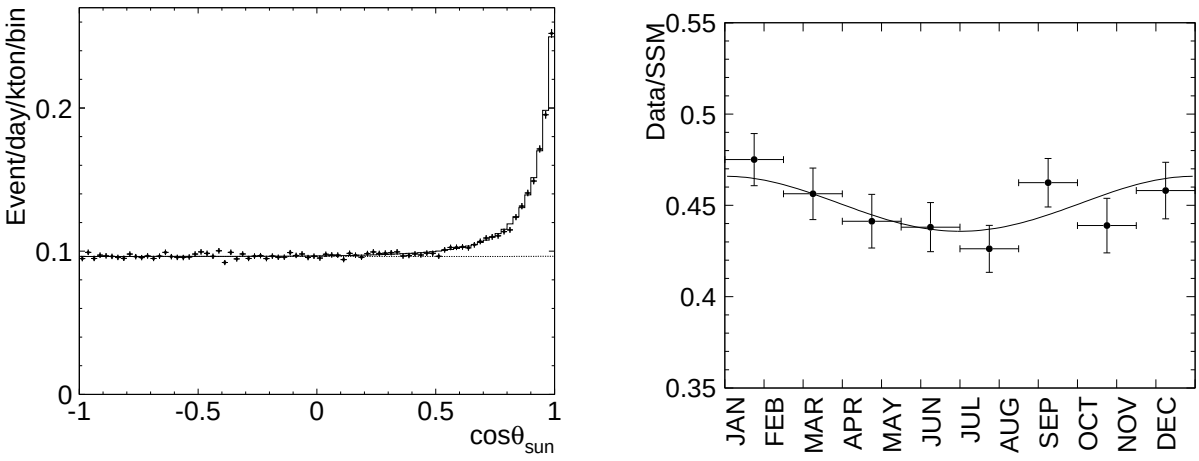


Fig. 17: Left: distribution of solar neutrino events as a function of the zenith angle of the Sun. Right: seasonal variation of the solar neutrino flux in SuperKamiokande.

$\nu_e + e^-$, that can be observed from the Cherenkov radiation emitted by the recoiling electrons. These are real-time experiments that provide information on the directionality and the energy of the neutrinos by measuring the recoiling electron. Unfortunately, the threshold for these types of experiments is much higher, ≥ 5 MeV. All these experiments have consistently observed a number of solar neutrinos between 1/3 and 1/2 of the number expected in the SSM and for a long time this was referred to as the *solar neutrino problem or deficit*.

The progress in this field over the past ten years has been enormous culminating in a solution to this puzzle that no longer relies on the predictions of the standard solar model.

There have been three milestones.

1998: SuperKamiokande [21] measured the solar neutrino deficit with unprecedented precision. Furthermore the measurement of the direction of the events demonstrated that the neutrinos measured definitely come from the Sun: the left plot of Fig. 17 shows the distribution of the events as a function of the zenith angle of the Sun. A seasonal variation of the flux is expected since the distance between the Earth and the Sun varies seasonally. The right plot of Fig. 17 shows that the measured variation is in perfect agreement with that expectation. If the deficit of ν_e in the Sun is interpreted in terms of neutrino oscillations, two very important observables to discriminate between different solutions are the spectral distribution of the events shown in the left plot of Fig. 18, which shows a rather flat spectrum, and the day/night asymmetry. The latter is important because neutrinos arriving from the Sun at night have to cross the Earth and some of the possible solutions are such that matter effects in neutrino propagation in the Earth are relevant. The analysis of solar data in year 2000 in terms of neutrino oscillations of the ν_e into some other type indicated a number of possible solutions as shown in Fig. 15.

2001: The SNO experiment [22] measured the flux of solar neutrinos using the three reactions:

$$(\text{CC}) \qquad \nu_e + d \to p + p + e^- \qquad E_{\text{thres}} > 5 \text{ MeV} \tag{65}$$

$$(\text{NC}) \qquad \nu_x + d \to p + n + \nu_x \quad x = e, \mu, \tau \qquad E_{\text{thres}} > 2.2 \text{ MeV} \tag{66}$$

$$(\text{ES}) \qquad \nu_e + e^- \to \nu_e + e^- \qquad E_{\text{thres}} > 5 \text{ MeV} \tag{67}$$

Since the CC reaction is only sensitive to electron neutrinos, while the NC one is sensitive to all the types that couple to the Z^0 boson, the comparison of the fluxes measured with both reactions can establish if there are ν_μ and ν_τ in the solar flux independently of the normalization given by the SSM. The neutrino

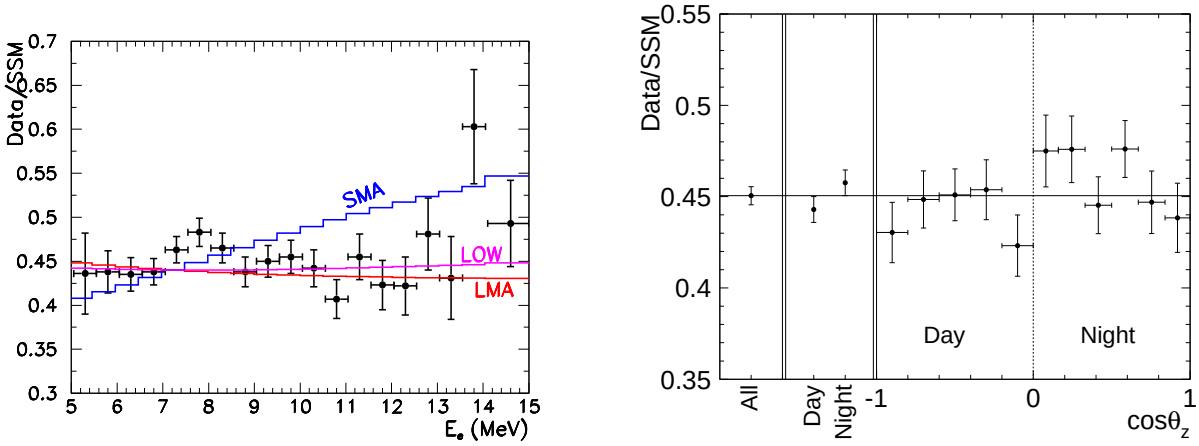


Fig. 18: Left: Distribution of the solar neutrino events as a function of the electron energy. Right: Day–night distribution of the solar neutrino events in SuperKamiokande.

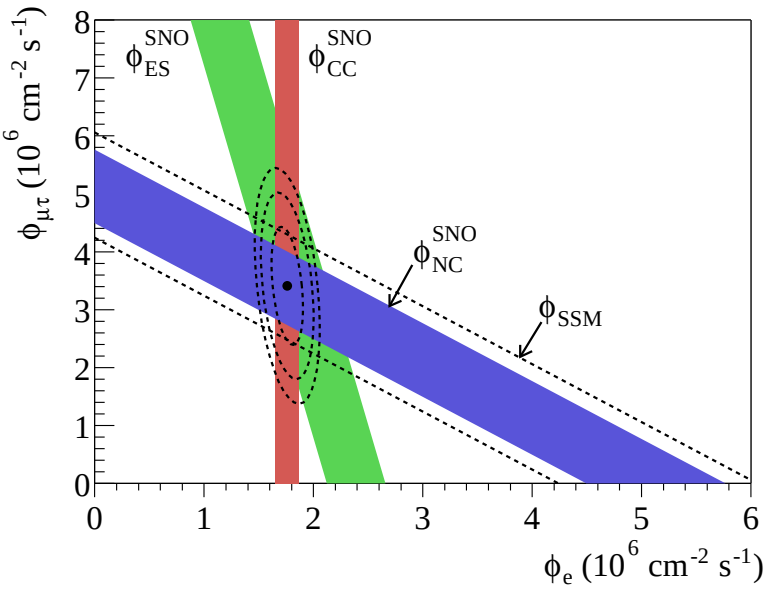


Fig. 19: Flux of ν_μ and ν_τ versus the flux of ν_e in the solar neutrino flux as measured from the three reactions observable in the SNO experiment. The dashed band shows the prediction of the SSM, which agrees perfectly with the flux measured with the NC reaction (from Ref.[23]).

fluxes measured by the three reactions by SNO are:

$$\phi^{\rm CC} = 1.67(9) \times 10^6 \text{ cm}^{-2}\text{s}^{-1}, \quad \phi^{\rm NC} = 5.54(48) \times 10^6 \text{ cm}^{-2}\text{s}^{-1}, \quad \phi^{\rm ES} = 1.77(26) \times 10^6 \text{ cm}^{-2}\text{s}^{-1}. \tag{68}$$

These measurements demonstrate that the Sun shines (ν_μ, ν_τ) about two times more than it shines ν_e, which constitutes the first direct demonstration of flavour transitions in the solar flux! Furthermore the NC flux that measures all active species in the solar flux, is compatible with the total ν_e flux expected according to the SSM as shown in Fig. 19.

The post-SNO global fits of all solar data shown in Fig. 20 (left) in terms of neutrino oscillations are quite different from those in Fig. 15. Of all the possible solutions, only the one at the largest mixing angle and mass square difference survives, the famous LMA solution.

2002: The solar oscillation is confirmed with reactor neutrinos in the KamLAND experiment [24].

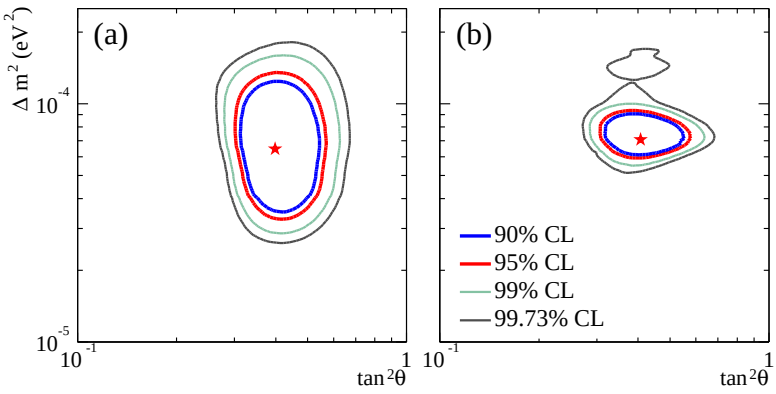


Fig. 20: Left: Analysis of all solar data in terms of neutrino oscillations. Right: Analysis including also KamLAND data (from Ref. [22]).

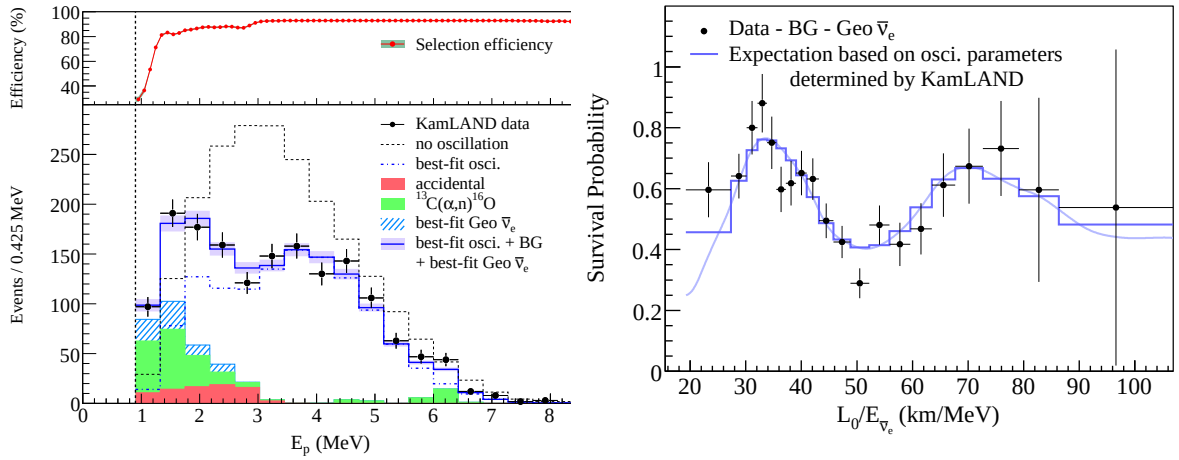


Fig. 21: Spectral distribution of the $\bar{\nu}_e$ events in KamLAND (left) and E_ν/L dependence (right). The data are compared to the expectation in the absence of oscillations and to the best fit oscillation hypothesis (from Ref. [25]).

This is 1kton of liquid scintillator which measures the flux of reactor neutrinos produced in a cluster of nuclear plants around Kamioka. The average distance is $\langle L \rangle = 175$ km. Neutrinos are detected via inverse β-decay which has a threshold energy of about 2.6 MeV:

$$\bar{\nu}_e + p \to e^+ + n \qquad E_{\rm th} > 2.6 \text{ MeV} \,. \tag{69}$$

The fortunate circumstance that

$$\langle E_\nu (1 \text{ MeV}) \rangle / L(100 \text{ km}) \sim 10^{-5} \text{ eV}^2 \tag{70}$$

is in the range indicated by solar data, and that the expected mixing angle is large, implies that a large depletion of the expected antineutrino flux (which is known to a few per cent accuracy) should be observed together with a significant energy dependence.

Figure 21 shows the latest KamLAND results [25] for the spectral distribution of events as well as as a function of the ratio E_ν/L. They have recently lowered the energy threshold and have sensitivity to geoneutrinos. The measurements of geoneutrinos could have important implications in geophysics. Concerning the sensitivity to the oscillation parameters, Fig. 22 shows the present determination of the solar oscillation parameters from KamLAND and other solar experiments. The precision in the determination of $\Delta m^2_{\rm solar}$ is spectacular and shows that neutrino experiments are entering the era of precision physics.

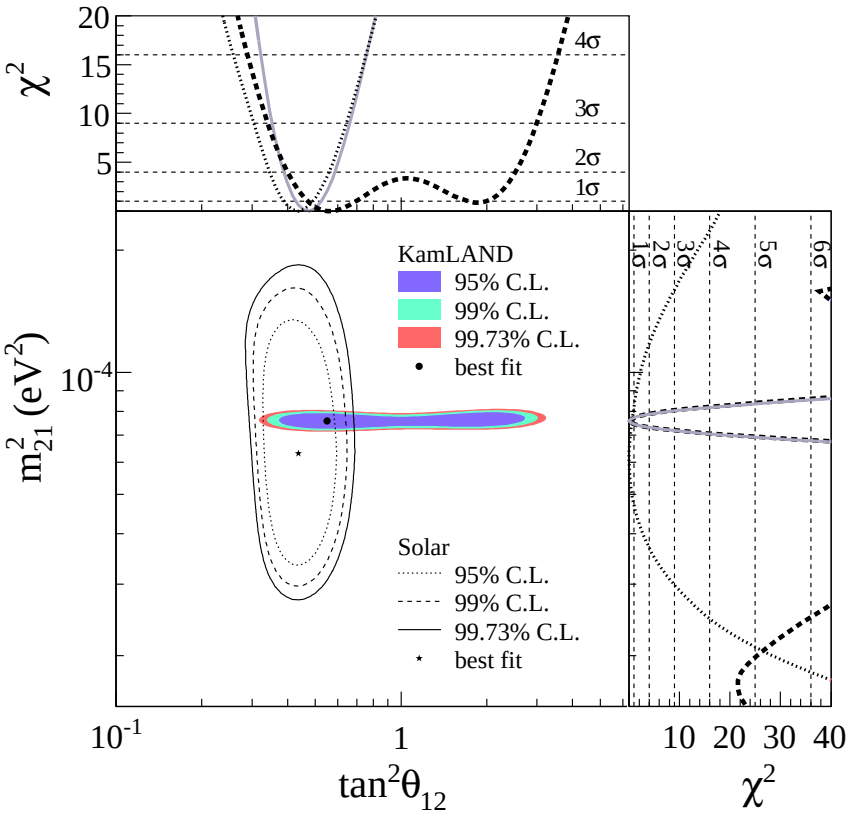


Fig. 22: Analysis of all solar and KamLAND data in terms of oscillations (from Ref. [25])

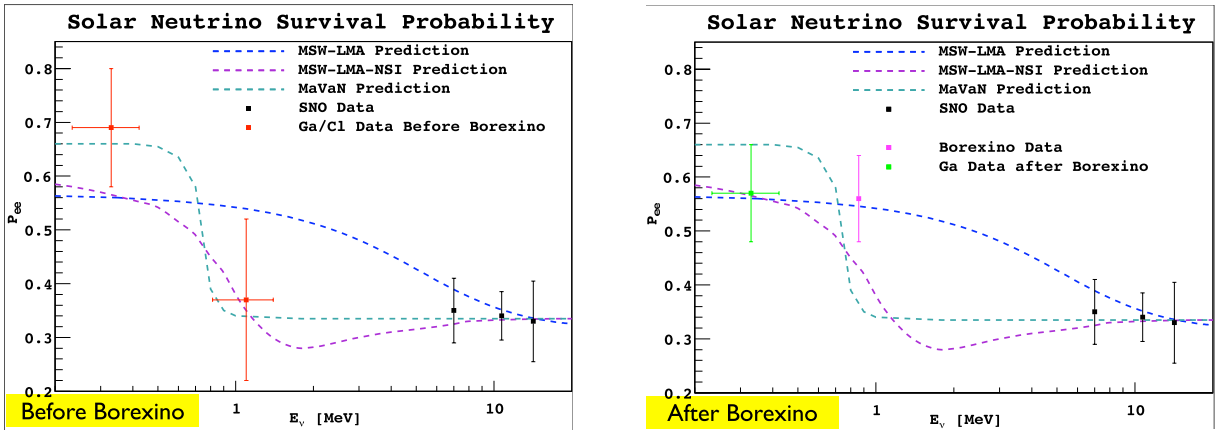


Fig. 23: Comparison of solar neutrino fluxes measured by the different experiments before Borexino (left) and after (right). Presented by the Borexino Collaboration at Neutrino 2008.

Last year new data was presented by a new solar neutrino experiment Borexino [26]. It is the lowest-threshold real-time solar neutrino experiment and the only one that could measure the flux of the monocromatic ^{7}Be neutrinos:

$$\Phi(^7\mathrm{Be}) = 5.08(25) \times 10^9 \ \mathrm{cm}^{-2}\mathrm{s}^{-1} \ .$$

The relevance of Borexino is illustrated in Fig. 23. The result is in agreement with the oscillation interpretation of other solar and reactor experiments and it adds further information to disfavour alternative exotic interpretations of the data.

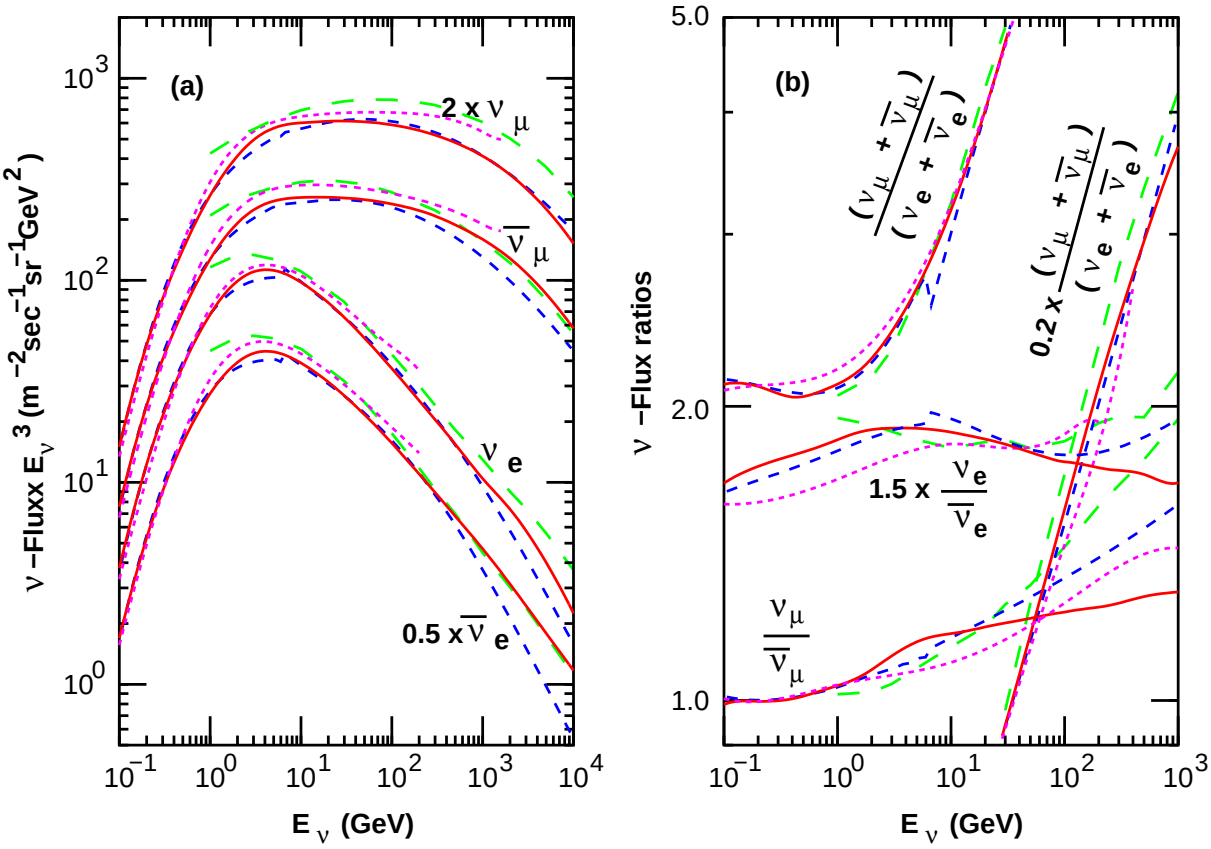


Fig. 24: Comparison of the predictions of different Monte Carlo simulations of the atmospheric neutrino fluxes averaged over all directions (left) and of the flux ratios $(\nu_\mu + \bar{\nu}_\mu)/(\nu_e + \bar{\nu}_e)$, $\nu_\mu/\bar{\nu}_\mu$, and $\nu_e/\bar{\nu}_e$ (right). The solid line corresponds to a recent full 3D simulation. Taken from the last reference in Ref. [27].

In summary, solar neutrinos experiments have made fundamental discoveries in particle physics and are now becoming useful for other applications, such as a precise understanding of the Sun and the Earth.

4.2 Atmospheric neutrino anomaly

Neutrinos are also produced in the atmosphere when primary cosmic rays impinge on it producing K, π that subsequently decay. The fluxes of such neutrinos can be predicted within a 10–20% accuracy to be those in the left plot of Fig. 24.

Clearly, atmospheric neutrinos are an ideal place to look for neutrino oscillation since the E_ν/L span several orders of magnitude, with neutrino energies varying from a few hundred MeV to 10^3 GeV and distances between production and detection varying from 10–10^4 km, as shown in Fig. 25 (right).

Many of the uncertainties in the predicted fluxes cancel when the ratio of muon to electron events is considered. The first indication of a problem was found when a deficit was observed precisely in this ratio by several experiments: Kamiokande [28], IMB [29], Soudan2 [30], Macro [31].

In 1998, SuperKamiokande clarified to a large extent the origin of this anomaly [32]. This experiment can distinguish muon and electron events, measure the direction of the outgoing lepton (the zenith angle with respect to the Earth's axis) which is correlated to that of the neutrino (the higher the energy the higher the correlation), in such a way that they could measure the variation of the flux as a function of the distance travelled by the neutrinos. Furthermore, they considered different samples of events: sub-GeV (lepton with energy below 1 GeV)), multi-GeV (lepton with energy above 1 GeV), together

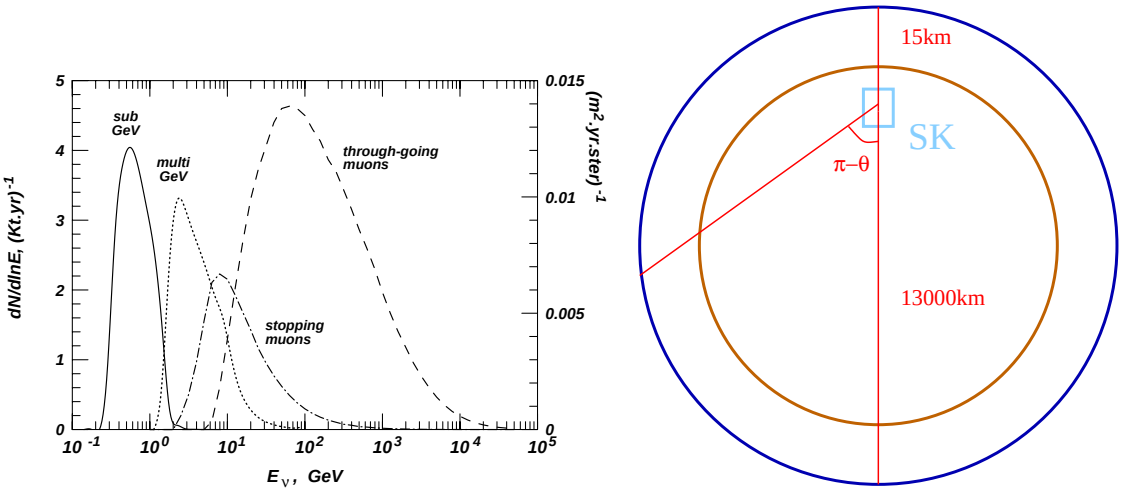


Fig. 25: Left: Parent neutrino energies of the different samples considered in Superkamiokande: sub-GeV, multi-GeV, stopping and through-going muons. Right: Distances travelled by atmospheric neutrinos as a function of the zenith angle.

with stopping and through-going muons that are produced on the rock surrounding Superkamiokande. The different samples correspond to different parent neutrino energies as can be seen in Fig. 25 (left). The number of events for the different samples as a function of the zenith angle of the lepton are shown in Fig. 26.

While the electron events observed are in agreement with predictions, a large deficit of muon events was found with a strong dependence on the zenith angle: the deficit was almost 50% for those events corresponding to neutrinos coming from below $\cos\theta = -1$, while there is no deficit for those coming from above. The quality of the fit to the neutrino oscillation hypothesis $\nu_\mu \to \nu_\tau$ is shown in the plot. The perfect fit to the oscillation hypothesis is rather non-trivial given the sensitivity of this measurement to the E_ν (different samples) and L (zenith angle) dependence. The significance of the E_ν/L dependence has been presented recently by the SuperKamiokande Collaboration [34], as shown in Fig. 27.

Appropriate neutrino beams to search for the atmospheric oscillation can easily be produced at accelerators if the detector is located at a long baseline of a few hundred kilometres, since

$$|\Delta m^2_{\text{atmos}}| \sim \frac{E_\nu(1-10\text{ GeV})}{L(10^2-10^3\text{ km})}. \tag{71}$$

A *conventional* neutrino beam is produced from protons hitting a target and producing π and K:

$$p \;\to\; \text{Target} \to \pi^+, K^+ \to \quad \nu_\mu(\%\nu_e, \bar\nu_\mu, \bar\nu_e) \tag{72}$$

$$\nu_\mu \to \nu_x. \tag{73}$$

Those of a selected charge are focused and are left to decay in a long decay tunnel producing a neutrino beam of mostly muon neutrinos (or antineutrinos) with a contamination of electron neutrinos of a few per cent. The atmospheric oscillation can be established by studying, as a function of the energy, either the disappearance of muon neutrinos or, if the energy of the beam is large enough, the appearance of τ neutrinos.

There are three such conventional beams: KEK–Kamioka (L = 235 km), Fermilab–Soudan (L = 730 km), CERN-Gran Sasso (L = 730 km). The latter being the only one sensitive to ν_τ appearance. The K2K experiment at Kamioka has already presented a positive signal for ν_μ disappearance [35], confirming the atmospheric oscillation. Their result is shown in Fig. 28. More recently also the MINOS experiment has presented a positive result as shown in Fig. 29.

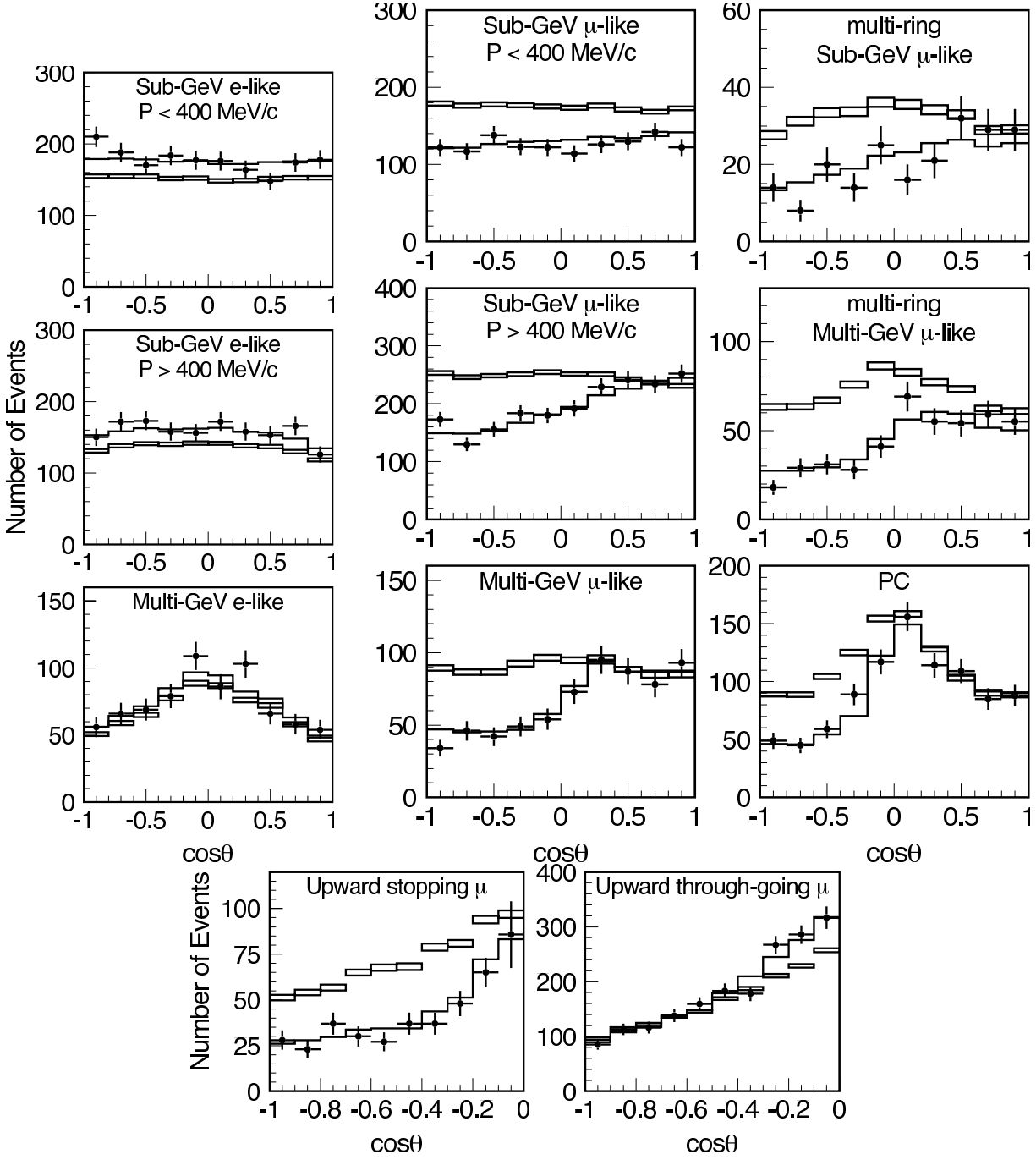


Fig. 26: Zenith angle distribution for fully-contained single-ring e-like and μ-like events, multi-ring μ-like events, partially contained events, and upward-going muons. The points show the data and the solid lines show the Monte Carlo events without neutrino oscillation. The dashed lines show the best-fit expectations for $\nu_\mu \leftrightarrow \nu_\tau$ oscillations (from Ref. [33]).

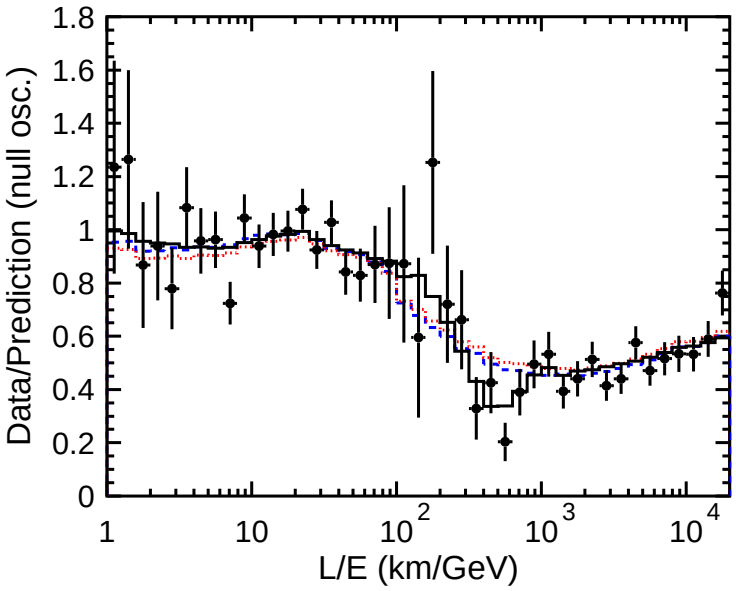


Fig. 27: Ratio of the data to the non-oscillated Monte Carlo events (points) with the best-fit expectation for 2-flavour $\nu_\mu \leftrightarrow \nu_\tau$ oscillations (solid line) as a function of E_ν/L (from Ref. [34]).

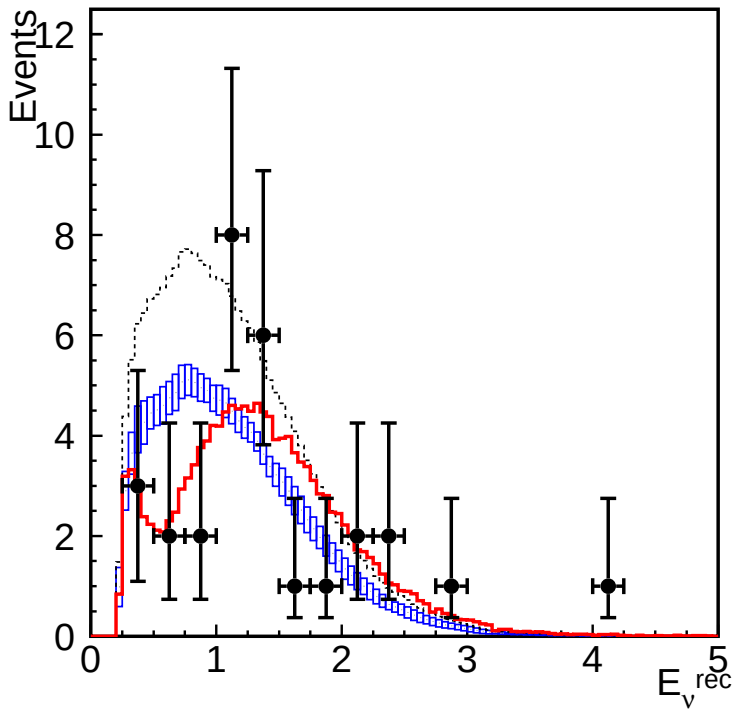


Fig. 28: Distribution of ν_μ events in K2K as a function of the *reconstructed* neutrino energy (from Ref. [35])

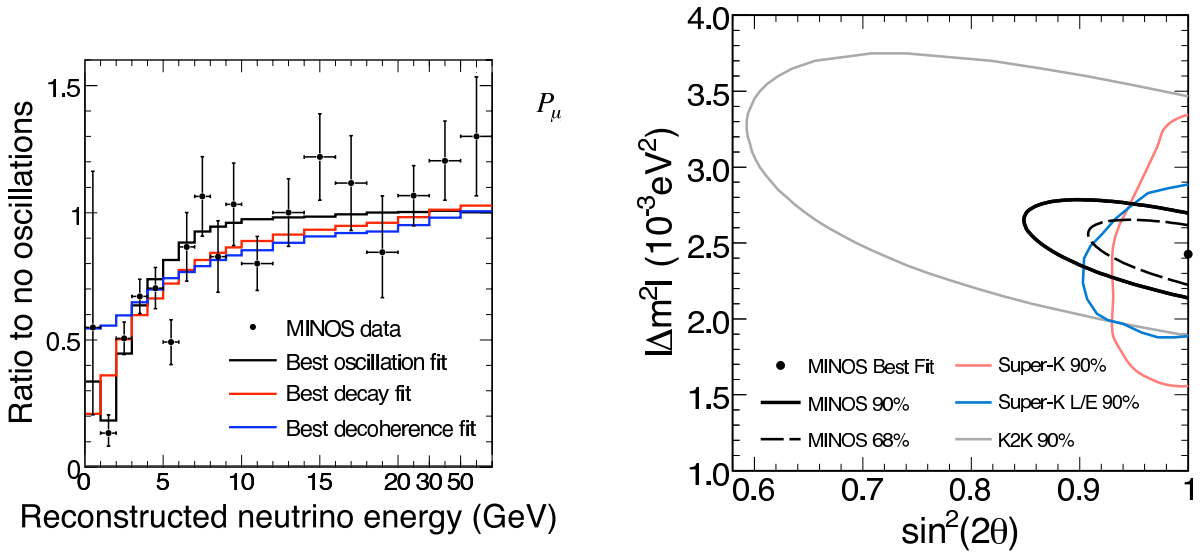


Fig. 29: Left: Ratio of measured to expected (in absence of oscillations) neutrino events in MINOS as a functions of neutrino energy. Right: Determination of oscillation parameters from MINOS data compared to K2K and Super-K.

4.3 Reactor experiments in the atmospheric range

Experiments that look for the disappearance of reactor $\bar{\nu}_e$ at an $E_\nu/L \sim \Delta m^2_{\text{atmos}}$ have also been performed [36, 37, 38]. The most sensitive of these has been Chooz [38]. No disappearance of $\bar{\nu}_e$ was observed, which excludes the parameter range shown in Fig. 30. Although SuperKamiokande had already established that atmospheric $\nu_e/\bar{\nu}_e$ do not seem to oscillate in the atmospheric range, the sensitivity of SuperKamiokande to this oscillation turns out to be much worse than that of Chooz because of the presence of electron and muon neutrinos in the atmospheric flux. It is in the context of three-neutrino mixing that the negative signal of Chooz has been most relevant, as we shall see.

4.4 LSND

Finally, an accelerator experiment, LSND, has found an appearance signal that could be interpreted in terms of neutrino flavour transitions [39]. They observed a surplus of electron events in a muon neutrino beam from π^+ decaying in flight (DIF) and a surplus of positron events in a neutrino beam from μ^+ decaying at rest (DAR). The interpretation of this data in terms of neutrino oscillations gives the range shown by a coloured band in Fig. 31:

$$\begin{aligned} \pi^+ \to \quad & \mu^+ \quad \nu_\mu \\ & \qquad \nu_\mu \to \nu_e \quad \text{DIF } (28 \pm 6/10 \pm 2) \\ & \mu^+ \quad \to e^+ \nu_e \bar{\nu}_\mu \\ & \qquad\qquad \bar{\nu}_\mu \to \bar{\nu}_e \quad \text{DAR } (64 \pm 18/12 \pm 3) \end{aligned}$$

Part of this region was already excluded by the experiment KARMEN [40] that has unsuccessfully searched for $\bar{\nu}_\mu \to \bar{\nu}_e$ in a similar range.

In 2006 the first results from MiniBOONE were presented. This experiment was designed to search for $\nu_\mu \to \nu_e$ transitions in the region of the LSND signal. They did not find confirmation of LSND as shown in Fig. 31

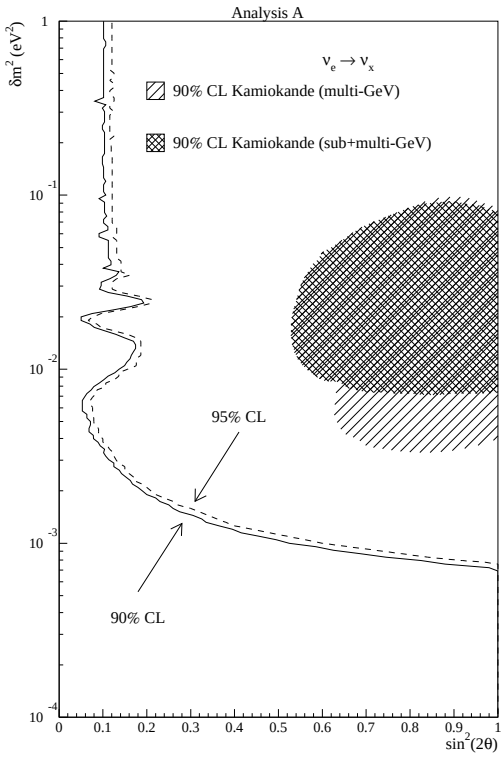


Fig. 30: Range of oscillation parameters for the oscillation $\bar{\nu}_e \rightarrow \bar{\nu}_x$ excluded by the Chooz data (from Ref. [38])

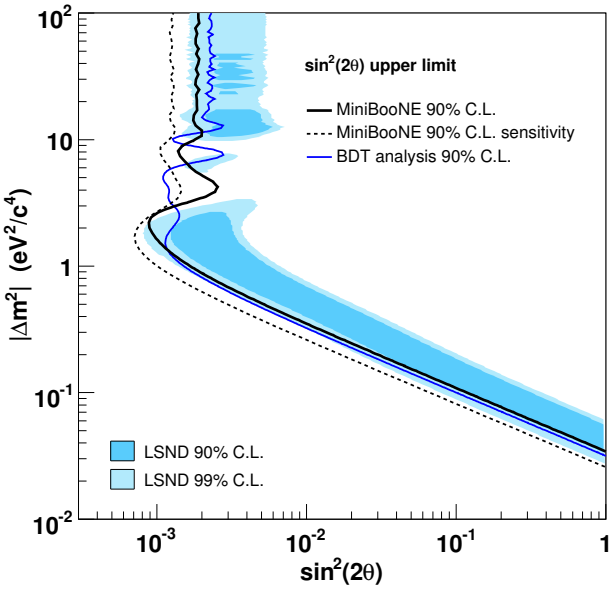


Fig. 31: Range of parameters for the oscillation $\nu_\mu \rightarrow \nu_e$ that could explain LSND data and those excluded by MiniBOONE (from Ref. [41])

4.5 Three-neutrino mixing

As we have seen there is experimental evidence for neutrino oscillation pointing to three distinct neutrino mass square differences:

$$\underbrace{|\Delta m^2_{\text{Sun}}|}_{\sim 8\cdot 10^{-5}\ \text{eV}^2} \ll \underbrace{|\Delta m^2_{\text{atmos}}|}_{\sim 2.5\cdot 10^{-3}\ \text{eV}^2} \ll \underbrace{|\Delta m^2_{\text{LSND}}|}_{>0.1\ \text{eV}^2} \tag{74}$$

Clearly the mixing of the three standard neutrinos ν_e, ν_μ, ν_τ can only explain two of the anomalies, so the explanation of the three sets of data would require the existence of a sterile ν species, since only three light neutrinos can couple to the Z^0 boson.

The existence of extra light sterile neutrinos could accomodate a third splitting, but all such scenarios give a very poor fit to all data.

It is now the standard scenario to consider three-neutrino mixing dropping the LSND result. The two independent neutrino mass square differences are assigned to the solar and atmospheric ones:

$$\Delta m^2_{13} = m^2_3 - m^2_1 = \Delta m^2_{\text{atmos}}, \qquad \Delta m^2_{12} = m^2_2 - m^2_1 = \Delta m^2_{\text{Sun}}\ . \tag{75}$$

With this convention, the mixing angles θ_{23} and θ_{12} in the parametrization of Eq. (28) correspond approximately to the ones measured in atmospheric and solar oscillations, respectively. This is because solar and atmospheric anomalies approximately decouple as independent 2-by-2 mixing phenomena thanks to the hierarchy between the two mass splittings, $|\Delta m^2_{\text{atmos}}| \gg |\Delta m^2_{\text{Sun}}|$, on the one hand and the fact that the angle θ_{13}, which measures the electron component of the third mass eigenstate element $\sin\theta_{13} = (V_{\text{MNS}})_{e3}$, is small.

To see this, let us first consider the situation in which $E_\nu/L \sim \Delta m^2_{13}$. We can thus neglect the solar mass square difference in front of the atmospheric one and E_ν/L. The oscillation probabilities obtained in this limit are given by

$$P(\nu_e \to \nu_\mu) \simeq s^2_{23}\ \sin^2 2\theta_{13}\ \sin^2\left(\frac{\Delta m^2_{13} L}{4E_\nu}\right), \tag{76}$$

$$P(\nu_e \to \nu_\tau) \simeq c^2_{23}\ \sin^2 2\theta_{13}\ \sin^2\left(\frac{\Delta m^2_{13} L}{4E_\nu}\right), \tag{77}$$

$$P(\nu_\mu \to \nu_\tau) \simeq c^4_{13}\ \sin^2 2\theta_{23}\ \sin^2\left(\frac{\Delta m^2_{13} L}{4E_\nu}\right). \tag{78}$$

Only two angles enter these formulae: θ_{23} and θ_{13}. The latter is the only one that enters the disappearance probability for ν_e in this regime:

$$P(\nu_e \to \nu_e) = 1 - P(\nu_e \to \nu_\mu) - P(\nu_e \to \nu_\tau) \simeq \sin^2 2\theta_{13}\ \sin^2\left(\frac{\Delta m^2_{13} L}{4E_\nu}\right)\ . \tag{79}$$

This is precisely the measurement of the Chooz experiment. Therefore the result of Chooz constrains the angle θ_{13} to be unobservably small.

If θ_{13} is set to zero in Eq. (78), the only probability that survives is the $\nu_\mu \to \nu_\tau$ one, which has the same form as a 2-family mixing formula Eq. (40) if we identify

$$(\Delta m^2_{\text{atmos}}, \theta_{\text{atmos}}) \to (\Delta m^2_{13}, \theta_{23})\ . \tag{80}$$

Instead if $E_\nu/L \sim \Delta m^2_{12}$, the atmospheric oscillation its too rapid and gets averaged out. The survival probability for electrons in this limit is given by:

$$P(\nu_e \to \nu_e) \simeq c^4_{13}\left(1 - \sin^2 2\theta_{12}\ \sin^2\left(\frac{\Delta m^2_{12} L}{4E_\nu}\right)\right) + s^4_{13}. \tag{81}$$

Again it depends only on two angles, θ_{12} and θ_{13}, and in the limit in which the latter is zero, the survival probability measured in solar experiments has the form of two-family mixing if we identify

$$(\Delta m^2_{\text{Sun}}, \theta_{\text{Sun}}) \to (\Delta m^2_{12}, \theta_{12}) \; . \tag{82}$$

The results that we have shown of solar and atmospheric experiments have been analysed in terms of 2-family mixing. The previous argument indicates that when fits are done in the context of 3-family mixing nothing changes very much, thanks to the strong constrain set by Chooz on θ_{13}.

Figure 32 shows the result of a recent global analysis of all data for the different parameters. The 2σ limits are

$$\begin{aligned} \theta_{23} = 36.9^\circ - 51.3^\circ \qquad \theta_{12} = 32.3^\circ - 37.8^\circ \qquad \theta_{13} < 10.3^\circ \\ \Delta m^2_{12} = 7.66(35) \times 10^{-5}\ \text{eV}^2 \qquad \Delta m^2_{23} = 2.38(27) \times 10^{-3}\ \text{eV}^2 \; . \end{aligned} \tag{83}$$

In summary, all the data, except LSND, can be explained if the neutrino spectrum has a structure as shown in Fig. 33. The neutrino mixing matrix is approximately given by

$$|V_{\text{MNS}}| \simeq \begin{pmatrix} 0.77 - -0.86 & 0.5 - -0.63 & 0 - -0.22 \\ 0.22 - -0.56 & 0.44 - -0.73 & 0.57 - -0.80 \\ 0.21 - -0.55 & 0.40 - -0.71 & 0.59 - -0.82 \end{pmatrix} \; , \tag{84}$$

and we do not know anything about the phases $(\delta, \alpha_1, \alpha_2)$. Note the striking difference between this mixing matrix and the CKM matrix which is approximately diagonal:

$$V_{\text{CKM}} \simeq \begin{pmatrix} 1 & O(\lambda) & O(\lambda^3) \\ O(\lambda) & 1 & O(\lambda^2) \\ O(\lambda^3) & O(\lambda^2) & 1 \end{pmatrix} \quad \lambda \sim 0.2. \tag{85}$$

The main features are

- Large mixing angles, in particular one is close to maximal.
- There is an intriguing near tri-bimaximal mixing pattern

$$V_{\text{tri-bi}} \simeq \begin{pmatrix} \sqrt{\frac{2}{3}} & \sqrt{\frac{1}{3}} & 0 \\ -\sqrt{\frac{1}{6}} & \sqrt{\frac{1}{3}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{3}} & \sqrt{\frac{1}{2}} \end{pmatrix} .$$

5 Prospects in neutrino physics

After the next generation of neutrino experiments that are under construction, we shall probably still be far from having complete knowledge of the neutrino mass matrix. There remain several fundamental questions to be answered:

1. Are neutrinos Dirac or Majorana particles?
2. Is total lepton number conserved or violated?
3. What is the absolute neutrino mass scale? Is it a new physics scale?
4. What is the neutrino mass spectrum: i.e., $\Delta m^2_{\text{atmos}} > \text{or} < 0$?
5. Is there CP violation in the lepton sector?
6. What is the value of θ_{13}?

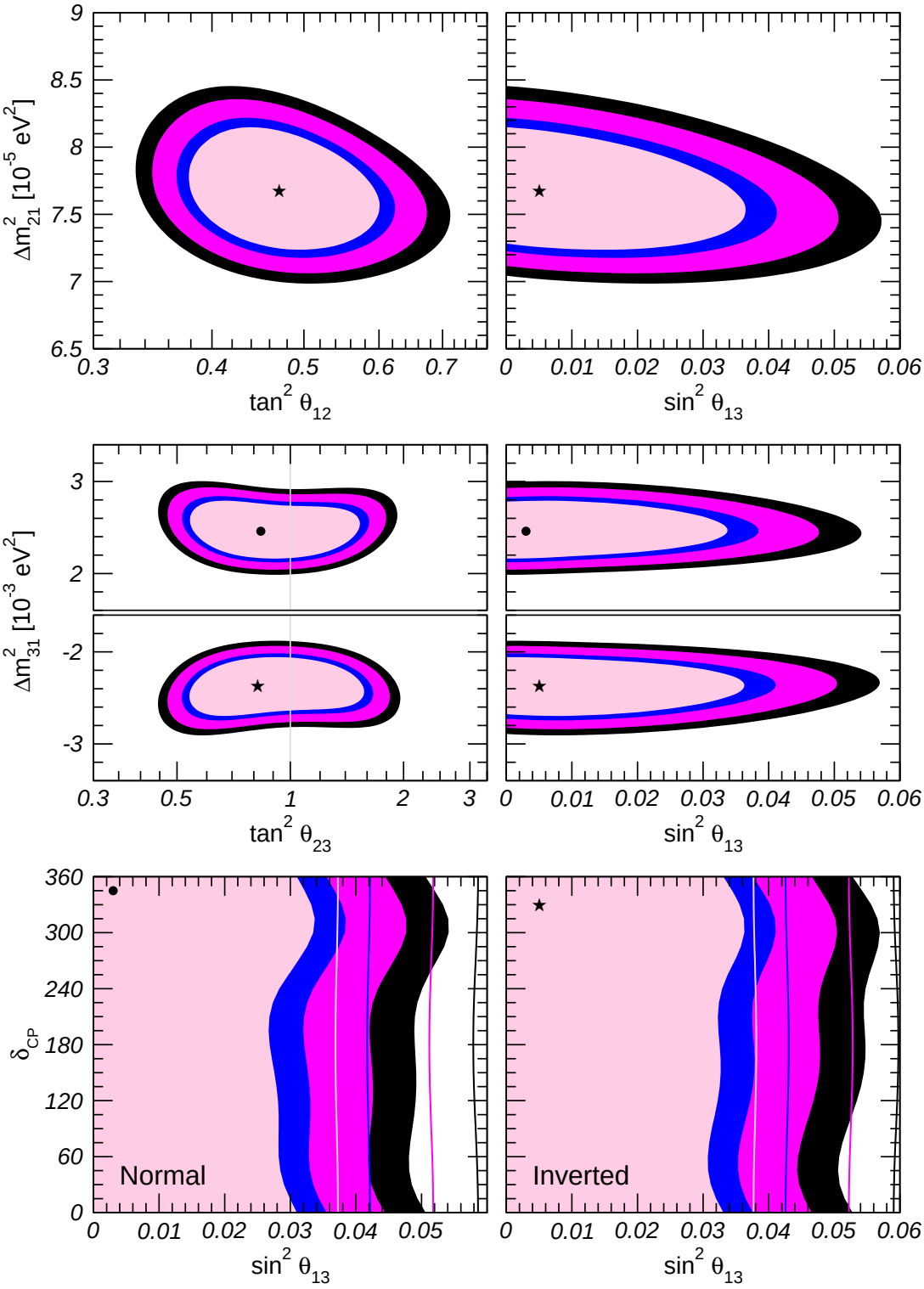


Fig. 32: Fits to the standard 3ν-mixing scenario including all available neutrino oscillation data (from Ref. [42])

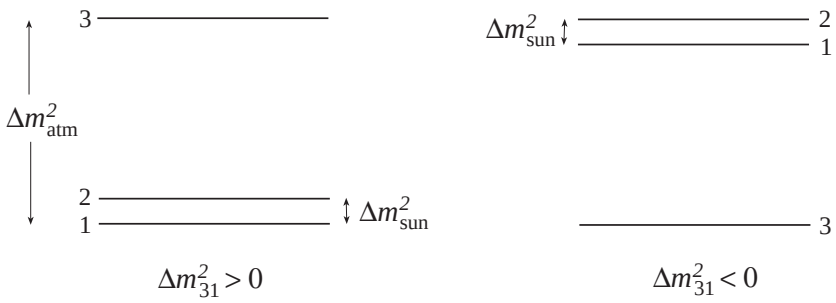


Fig. 33: Possible neutrino spectra consistent with solar and atmospheric data

The best hope addressing the first three questions lies in more precise experiments searching for neutrinoless double-β decay, measuring the end-point of β decay as well as cosmological measurements. Figure 34 shows the present constraints on the combination of parameters that is directly measured in $2\beta0\nu$ experiments:

$$m_{\beta\beta} \equiv |m_{ee}| = |c^2_{13}(m_1 c^2_{12} + m2e^{i\alpha_1} s^2_{12}) + m_3 e^{i\alpha_2} s^2_{13}| \,, \tag{86}$$

and in cosmology:

$$\Sigma \equiv m_1 + m_2 + m_3 \,. \tag{87}$$

The cosmological data included in this fit is only that from the cosmic microwave background (CMB).

Note that a lot of information on $m_{\beta\beta}$ is already provided by neutrino oscillation experiments. If the hierarchy is inverse ($m_3 \ll m_1, m_2 \sim \sqrt{|\Delta m^2_{\rm atmos}|}$), there is a lower bound on $m_{\beta\beta} \geq 10^{-2}$ eV, as shown by the red (I.H.) band. Instead, if the hierarchy is normal $m_3 \sim \sqrt{|\Delta m^2_{\rm atmos}|} \gg m_1, m_2$, there is no lower bound because neither θ_{13} nor m_1 is bounded from below, as shown by the blue (N.H.) band. The horizontal band shows the controversial claim of a positive signal [10].

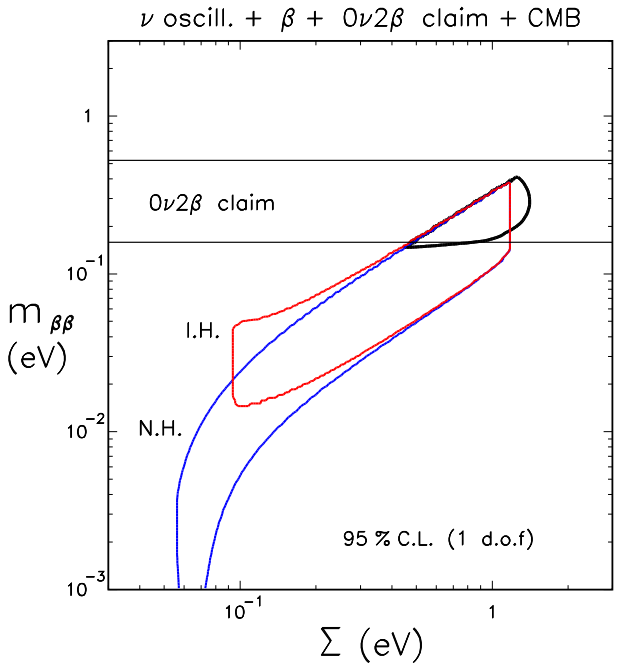


Fig. 34: Present constraints on $m_{\beta\beta}$ and Σ from neutrino experiments and CMB data (from Ref. [43])

A plethora of forthcoming experiments that will improve these constraints are under construction.

KATRIN [44] is an experiment to measure the spectrum of tritium β decay that is expected to improve the sensitivity to the element:

$$m_e \equiv \sqrt{m_1^2 c_{12}^2 c_{13}^2 + m_2^2 s_{12}^2 c_{13}^2 + m_3^2 s_{13}^2} \tag{88}$$

to about 0.2 eV, which is an improvement of one order of magnitude with respect to the present limit in Eq. (6). Concerning $0\nu\beta\beta$ [45] the next step of several experiments using different detector techniques (CUORE, EXO, GENIUS, Majorana, etc.) is to reach the level of precision of $m_{\beta\beta} \sim 0.1$ eV, which would allow testing the positive claim in a definite way. Further in the future there are also proposals to improve this precision by another order of magnitude reaching the 10^{-2} eV level, which could be sufficient to explore the full parameter space in the case of the inverse hierarchy. The measurement of a non-zero $m_{\beta\beta}$ would not only prove that neutrinos are Majorana and that lepton number is violated, but might give the best determination of the lightest neutrino mass, and even help in establishing the neutrino mass hierarchy.

Concerning cosmology, it is quite impressive that the sensitivity to the neutrino matter component of the Universe has already reached the eV range. Further significant improvements are expected in the near future (e.g., by PLANCK) that can push present limits by at least one order of magnitude.

Concerning the last three fundamental questions above, they can be studied in more precise neutrino oscillation experiments in the atmospheric range (i.e., $\langle E_\nu \rangle / L \sim \Delta m^2_{\text{atmos}}$) optimized to measure the subleading transitions involving ν_e. In particular, $\nu_e \leftrightarrow \nu_\mu$ and $\bar\nu_e \leftrightarrow \bar\nu_\mu$ are the so-called *golden* measurements [46], while the $\nu_e \leftrightarrow \nu_\tau$ and $\bar\nu_e \leftrightarrow \bar\nu_\tau$, being experimentally more challenging, are the *silver* ones [47].

5.1 CP violation in neutrino oscillations

As in the quark sector, the mixing matrix of three neutrinos has CP violating phases. The so-called Dirac phase, δ, induces CP violation in neutrino oscillations, that is a difference between $P(\nu_\alpha \to \nu_\beta)$ and $P(\bar\nu_\alpha \to \bar\nu_\beta)$, for $\alpha \neq \beta$. As we saw in the general expression of Eq. (38), CP violation is possible if there are imaginary entries in the mixing matrix that make $\text{Im}[W^{jk}_{\alpha\beta} \equiv [U_{\alpha j} U^*_{\beta j} U^*_{\alpha k} U_{\beta k}] \neq 0$. By CPT, disappearance probabilities cannot violate CP however, because under CPT

$$P(\nu_\alpha \to \nu_\beta) = P(\bar\nu_\beta \to \bar\nu_\alpha)\ , \tag{89}$$

so in order to observe a CP or T-odd asymmetry the initial and final flavour must be different, $\alpha \neq \beta$:

$$A^{CP}_{\alpha\beta} \equiv \frac{P(\nu_\alpha \to \nu_\beta) - P(\bar\nu_\alpha \to \bar\nu_\beta)}{P(\nu_\alpha \to \nu_\beta) + P(\bar\nu_\alpha \to \bar\nu_\beta)}, \quad A^{T}_{\alpha\beta} \equiv \frac{P(\nu_\alpha \to \nu_\beta) - P(\nu_\beta \to \nu_\alpha)}{P(\nu_\alpha \to \nu_\beta) + P(\nu_\beta \to \nu_\alpha)}\,. \tag{90}$$

In the case of 3-family mixing it is easy to see that the CP(T)-odd terms in the numerator are the same for all transitions $\alpha \neq \beta$:

$$A^{\text{CP(T)-odd}}_{\nu_\alpha \nu_\beta} = \frac{\sin\delta c_{13} \sin 2\theta_{13} \overbrace{\sin 2\theta_{12} \frac{\Delta m^2_{12} L}{4E_\nu}}^{\text{solar}} \overbrace{\sin 2\theta_{23} \sin^2 \frac{\Delta m^2_{13} L}{4E_\nu}}^{\text{atmos}}}{P^{\text{CP-even}}_{\nu_\alpha \nu_\beta}}. \tag{91}$$

As expected, the numerator is GIM suppressed in all the Δm^2_{ij} and all the angles, because if any of them is zero, the CP-odd phase becomes unphysical.

In order to maximize this asymmetry, it is necessary to perform experiments in the atmospheric range $\langle E_\nu \rangle / L \sim \Delta m^2_{\text{atmos}}$, so that the GIM suppression is minimized. In this case, only two small parameters remain in the CP-odd terms: the solar splitting, Δm^2_{Sun} (i.e., small compared to the other scales,

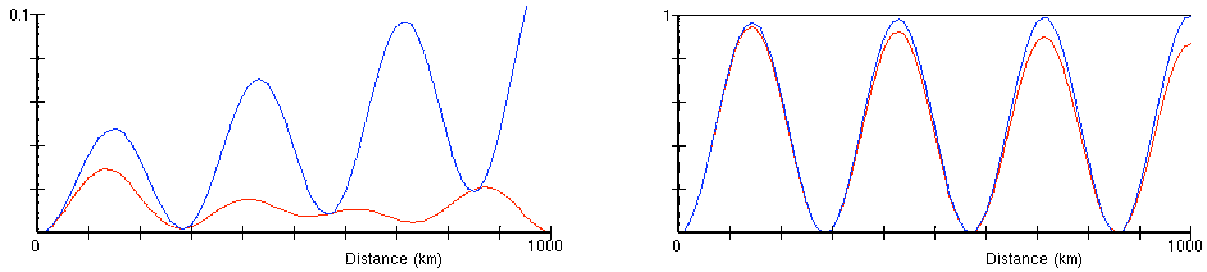


Fig. 35: Comparison of the $\nu_e \leftrightarrow \nu_\mu/\bar{\nu}_e \leftrightarrow \bar{\nu}_\mu$ (left) and $\nu_\mu \leftrightarrow \nu_\tau/\bar{\nu}_\mu \leftrightarrow \bar{\nu}_\tau$ (right) oscillation probabilities for $E_\nu = 500$ MeV, $\theta_{13} = 8^\circ$ and $\delta = 90^\circ$ as a function of the distance

$\Delta m^2_{\text{atmos}}$ and $\langle E_\nu \rangle / L$), and the angle θ_{13}. The asymmetry is then larger in the subleading transitions: $\nu_e \to \nu_\mu(\nu_\tau)$, because the CP-even terms in the denominator are also suppressed by the same small parameters. Indeed a convenient approximation for the $\nu_e \leftrightarrow \nu_\mu$ transitions is obtained expanding to second order in both small parameters [46]:

$$\begin{aligned} P_{\nu_e\nu_\mu(\bar{\nu}_e\bar{\nu}_\mu)} &= s^2_{23}\,\sin^2 2\theta_{13}\,\sin^2\left(\frac{\Delta m^2_{13}\,L}{4E_\nu}\right) \quad \equiv P^{\text{atmos}} \\ &+ c^2_{23}\,\sin^2 2\theta_{12}\,\sin^2\left(\frac{\Delta m^2_{12}\,L}{4E_\nu}\right) \quad \equiv P^{\text{solar}} \\ &+ \tilde{J}\cos\left(\pm\delta - \frac{\Delta m^2_{13}\,L}{4E_\nu}\right)\frac{\Delta m^2_{12}\,L}{4E_\nu}\sin\left(\frac{\Delta m^2_{13}\,L}{4E_\nu}\right) \quad \equiv P^{\text{inter}}, \end{aligned} \tag{92}$$

where $\tilde{J} \equiv c_{13}\ \sin 2\theta_{13}\ \sin 2\theta_{12}\ \sin 2\theta_{23}$. This approximate formula is obtained as an expansion to second order in the parameters θ_{13} and Δm^2_{Sun}. The first term corresponds to the atmospheric oscillation, the second one is the solar one and there is an interference term which has the information on the phase δ. Depending on the value of θ_{13}, it is possible that the atmospheric term dominates over the other two, in such a way that the CP-even terms are suppressed in θ^2_{13}, or if it is the solar term that dominates, the suppression is in $(\Delta m^2_{\text{Sun}})^2$. The asymmetries in these two regimes show therefore the following dependence on the small parameters:

$$\begin{aligned} P^{\text{atmos}} \gg P^{\text{solar}} \quad &\to \quad A^{CP,T}_{\nu_e\nu_\mu(\nu_\tau)} \sim \frac{\Delta m^2_{12} L/E_\nu}{\sin 2\theta_{13}}, \\ P^{\text{solar}} \gg P^{\text{atmos}} \quad &\to \quad A^{CP,T}_{\nu_e\nu_\mu(\nu_\tau)} \sim \frac{\sin 2\theta_{13}}{\Delta m^2_{12} L/E_\nu}, \\ P^{\text{solar}} \simeq P^{\text{atmos}} \quad &\to \quad A^{CP,T}_{\nu_e\nu_\mu(\nu_\tau)} = O(1)\ . \end{aligned} \tag{93}$$

Therefore asymmetries in the subleading transitions are expected to be rather large, specially when the solar and atmospheric terms are comparable.

In contrast, the asymmetries in the leading $\nu_\mu \to \nu_\tau$ transition in the atmospheric range are much smaller, because the CP-even terms are unsuppressed in each of the two small parameters. The difference between the neutrino and antineutrino oscillation probabilities for the leading and subleading channels are shown in Fig. 35.

5.2 The neutrino spectrum

The oscillation probabilities in matter can also be approximated by an expansion to second order in the two small parameters: θ_{13} and Δm^2_{12} [46]. The result has the same structure as in vacuum:

$$P_{\nu_e\nu_\mu(\bar{\nu}_e\bar{\nu}_\mu)} = s^2_{23}\,\sin^2 2\theta_{13}\left(\frac{\Delta_{13}}{B_\pm}\right)^2\ \sin^2\left(\frac{B_\pm L}{2}\right)$$

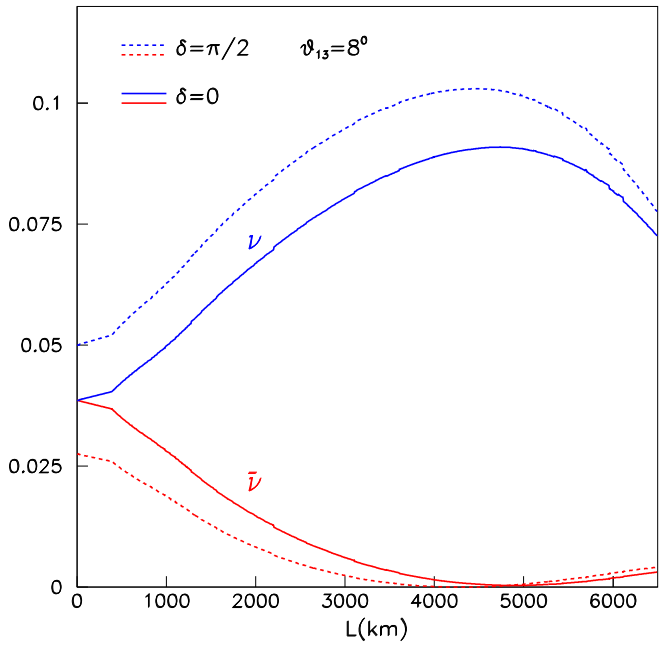


Fig. 36: $P(\nu_e \to \nu_\mu)$ and $P(\bar{\nu}_e \to \bar{\nu}_\mu)$ as a function of the baseline L in kilometres, at a neutrino energy $E_\nu/L = |\Delta m^2_{13}|/2\pi$ and for $\theta_{13} = 8°$ and $\delta = 0$ (solid) and $90°$ (dashed)

$$+c^2_{23}\,\sin^2 2\theta_{12}\left(\frac{\Delta_{12}}{A}\right)^2 \sin^2\left(\frac{A\,L}{2}\right)$$
$$+\tilde{J}\,\frac{\Delta_{12}}{A}\,\sin(\frac{AL}{2})\,\frac{\Delta_{13}}{B_\pm}\,\sin\left(\frac{B_\pm L}{2}\right)\,\cos\left(\pm\delta - \frac{\Delta_{13}\,L}{2}\right)\,, \tag{94}$$

where

$$B_\pm = |A \pm \Delta_{13}| \qquad \Delta_{ij} = \frac{\Delta m^2_{ij}}{2E_\nu} \qquad A = \sqrt{2} G_F N_e\,. \tag{95}$$

This formula shows a resonant enhancement of the atmospheric term in the the neutrino or antineutrino oscillation probability (depending on the sign of Δm^2_{13}) channel when

$$2E_\nu A \sim |\Delta m^2_{13}|\,. \tag{96}$$

Considering the electron number density in the Earth, the resonant energy is $E_\nu \sim 10 - -20$ GeV. This resonance is illustrated in Fig. 36, which shows the $\nu_e \to \nu_\mu$ oscillation probability for neutrinos and antineutrinos, as a function of the baseline, for neutrino energy constrained to the first atmospheric peak, i.e., $E_\nu/L = |\Delta m^2_{13}|/2\pi$. The difference between the neutrino and anti-neutrino oscillation probabilities induced by matter effects becomes comparable to that due to maximal CP-violation for $L = \mathcal{O}(1000)$ km. This is approximately the baseline where matter effects and CP violation can both be measured simultaneously. At much longer distances, matter effects completely hide CP-violation effects and vice versa.

5.3 The measurement of θ_{13} and δ

5.3.1 Theoretical challenge

In the future, we shall face the challenge of extracting simultaneously θ_{13}, δ and also the hierarchy from the measurement of the oscillation probabilities $\nu_\mu \leftrightarrow \nu_e$ and $\bar{\nu}_\mu \leftrightarrow \bar{\nu}_e$. This turns out to be non-

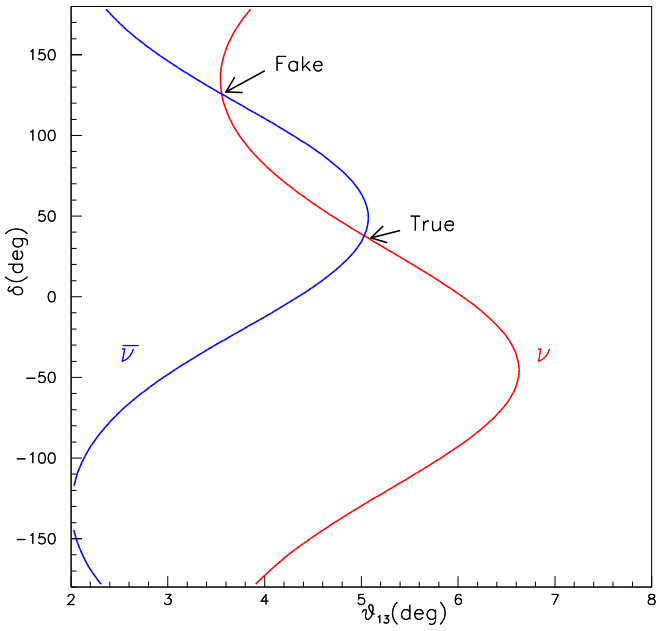


Fig. 37: Equiprobability curves $P_{\nu_e\nu_\mu}(E_\nu/L, \theta_{13}, \delta)$ = Meas$_1$ and $P_{\bar{\nu}_e\bar{\nu}_\mu}(E_\nu/L, \theta_{13}, \delta)$ = Meas$_2$ on the plane (θ_{13}, δ). They generically cross at two points: the true solution (θ_{13}, δ) and a fake one.

trivial even in principle, because of the existence of degeneracies [48]. In fact, at fixed E_ν, L there are generically two solutions for (θ_{13}, δ) that give the same probabilities for neutrinos and antineutrinos.

This is due to the periodicity in δ: if the equiprobability curves for neutrinos and antineutrinos on the plane (θ_{13}, δ) cross at one point (at the true solution), they must cross at least once more as shown in Fig. 37.

The fake solution has a strong dependence on the ratio E_ν/L in vacuum.

Normally neutrino beams are not monochromatic, so E_ν/L is not fixed. If we consider as the measurement the integrated signals (after integrating in energy the probability $\times$ flux $\times$ cross section), the same argument holds and a fake solution appears generically although it has a more complicated dependence on $\langle E_\nu \rangle$ and L.

Besides, the fact that other oscillation parameters will also not be known at the time of this measurement, such as the sign(Δm^2_{13}) or sign($\cos\theta_{23}$), increases the difficulty further: these unknowns will also bias the extraction of θ_{13} and δ leading to additional fake solutions, the so-called eight-fold degeneracy [49].

Several strategies for resolving these degeneracies have been proposed. Given the energy dependence of the fake solutions, it is very useful to have a detector with good neutrino energy resolution. Figure 38 shows the oscillation probability as a function of the neutrino energy for some values of (θ_{13}, δ) with that corresponding to the fake solution $(\theta_{13}^{\text{fake}}(\langle E_\nu \rangle/L), \delta^{\text{fake}}(\langle E_\nu \rangle/L))$. The curves cross at $\langle E_\nu \rangle$ but differ quite significantly at other energies.

Another possibility is to consider performing several experiments with differing $\langle E_\nu \rangle/L$ or with different matter effects.

Finally, the measurement of other oscillation probabilities beside the golden one can help. For example, if a precise measurement of the disappearance probability for ν_e is done in the atmospheric range, with an improved Chooz-type experiment, this could provide a measurement of θ_{13} that does not

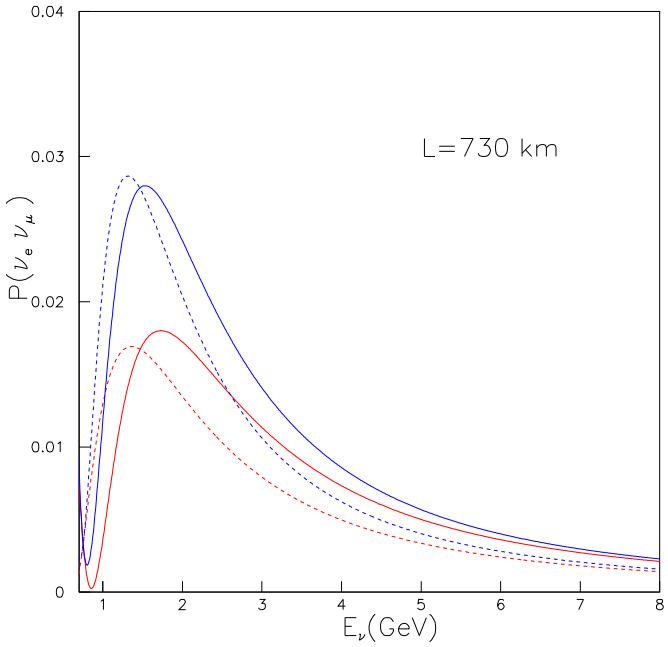


Fig. 38: Oscillation probability for neutrinos and antineutrinos as a function of the energy, for some true values of θ_{13} and δ, and for the fake solutions (dashed curves)

depend on δ at all [50].

Similarly, if we combine the golden measurement with the silver one: $\nu_e \to \nu_\tau$ and $\bar{\nu}_e \to \bar{\nu}_\tau$, the fake solutions can be excluded [47].

5.3.2 Experimental challenge

The challenge is to measure for the first time the *small* subleading transitions $\nu_e \leftrightarrow \nu_\mu$ and $\bar{\nu}_e \leftrightarrow \bar{\nu}_\mu$ with $\langle E_\nu \rangle / L \sim |\Delta m^2_{\text{atmos}}|$. The need to be above the muon threshold implies that rather long baselines are required as shown in Fig. 39. There are many ideas being pursued. Let us briefly describe the different proposals.

5.3.3 Future reactor experiments

Reactor neutrinos have an energy in the range of MeV and therefore can only look at the disappearance channel $\bar{\nu}_e \to \bar{\nu}_e$. It has been pointed out before that reactor neutrinos have provided the most stringent limit on the angle θ_{13}. A future upgrade of this type of experiments is possible, by increasing the detector size and reducing the systematics by intercepting the beam with both a near and a far detector. The experiment Double-Chooz is under construction and expects to reach a sensitivity limit of $\sin^2 2\theta_{13} \geq 0.03$, with the advantage that being a disappearance measurement, there is no ambiguity due to the CP phase δ or any other parameter.

5.3.4 Future superbeam experiments

Neutrino beams produced at accelerators have already been constructed to measure the disappearance of ν_μ in the atmospheric range (K2K and MINOS), as well as the apperance channel $\nu_\mu \to \nu_\tau$ (OPERA). As we have seen, these experiments have confirmed the leading atmospheric oscillation, but they will improve the sensitivity to the unknowns very little.

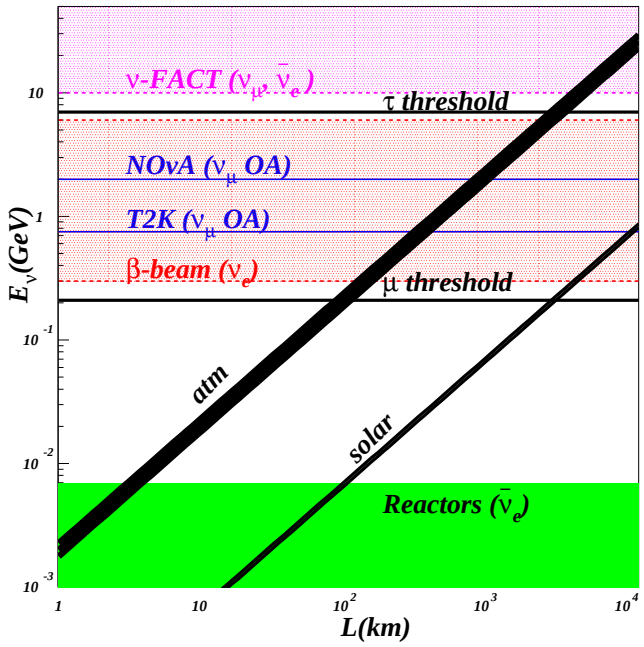


Fig. 39: Energy of the proposed future neutrino oscillation experiments: Nufact, β-beam, superbeams (T2K and NOvA) and reactors. The *atm* and *solar* black bands correspond to the first atmospheric and solar oscillation peaks, respectively.

These *conventional* beams result from the decay of pions and kaons produced from an intense proton beam that hits a target. They are thus mostly ν_μ (or $\bar{\nu}_\mu$ depending on the polarity) with a per cent contamination of ν_e. Neutrino beams of this type but with much higher intensity, the so-called *super-beams*, could be obtained with new megawatt proton sources, however, the sensitivity to the subleading transition $\nu_\mu \to \nu_e$ is limited by systematics. Not only can the flavour and spectral composition of these beams not be determined with good accuracy, but the irreducible background of ν_e is the limiting factor. One way to reduce this background is to use an off-axis configuration. Pion decay kinematics implies that a detector located off-axis intercepts a beam with a much better defined energy, and this allows the beam background to be reduced below the 1% level.

Two projects using off-axis superbeams are being pursued. The first one is T2K in Japan [51], that is expected to start taking data in 2009. It will use the SuperKamiokande detector to intercept a beam produced in J-PARC, which corresponds to a baseline of 295 km. If $\sin^2 2\theta_{13} \geq 0.01 - -0.02$, an appearance of ν_e will be observed, although the experiment will have no sensitivity to CP violation nor to the mass hierarchy. The second project is NOvA in the USA [52]. The NUMI beam at Fermilab will be intercepted off-axis by a new detector located 810 km away. It is expected to reach a similar sensitivity to θ_{13} as T2K, but if $\sin^2 2\theta_{13} \geq 0.05$, the comparison of the ν and $\bar{\nu}$ appearance signals could provide the first determination of the neutrino hierarchy.

5.3.5 Neutrino factory and β beams

The measurement of leptonic CP violation will probably require a further step. New ideas to obtain neutrino beams with reduced systematics have been actively discussed in recent years. At the *Neutrino Factory* (NF) [53] neutrinos are produced from μ^+ or μ^- which are accelerated to some reference energy and are allowed to decay in a storage ring with long straight sections (see Fig. 40). Subleading transitions

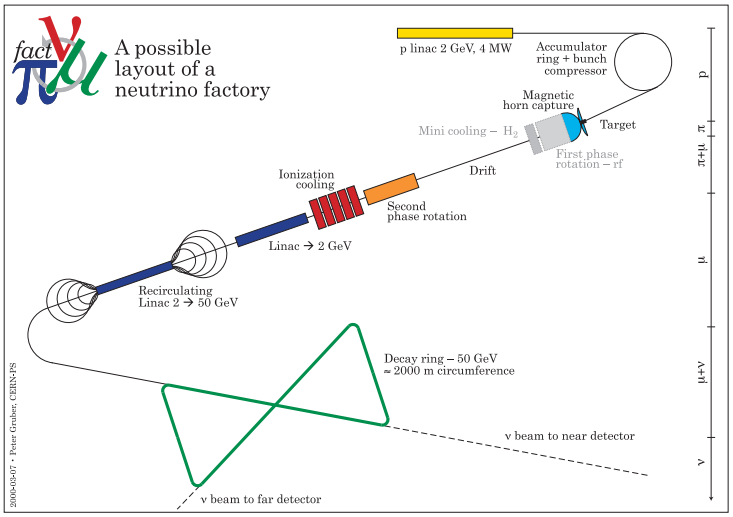


Fig. 40: Possible layout of a CERN-based Neutrino Factory complex

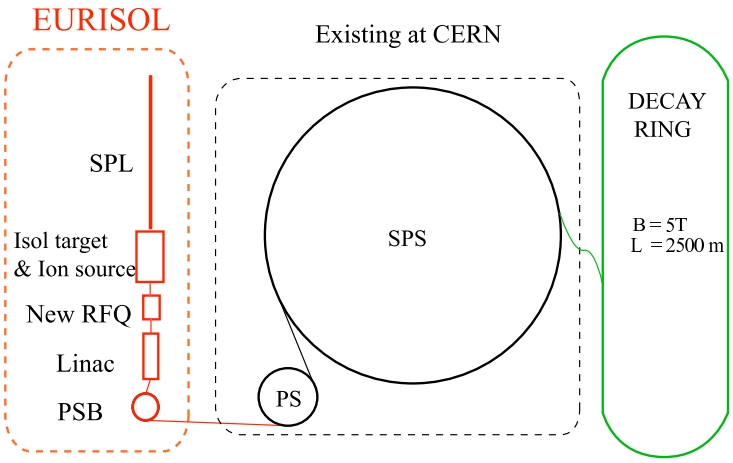


Fig. 41: Possible layout of a CERN-based β beam

can be searched for by looking for wrong-sign muons in a massive magnetized detector:

$$\begin{aligned} \mu^- \to e^- \quad \nu_\mu \quad \bar{\nu}_e\,; & \\ \bar{\nu}_e \to \bar{\nu}_\mu \to \mu^+ & \\ \nu_\mu \quad \to \nu_\mu \to \mu^-. & \end{aligned} \tag{97}$$

A similar situation is found in the case of the β *beam* (BB)[54]. This is a neutrino beam obtained from boosted radioactive ions, such as ${}^{18}_{10}\mathrm{Ne}$ or ${}^{6}\mathrm{He}^{++}$, which are accelerated and circulated in a storage ring where they decay, producing a pure ν_e or $\bar{\nu}_e$ beam, respectively (see Fig. 41):

$$\begin{aligned} {}^{6}\mathrm{He}^{++} \to {}^{6}_{3}\mathrm{Li}^{+++}\, e^- \qquad & \bar{\nu}_e \\ & \bar{\nu}_e \to \bar{\nu}_\mu \to \mu^+ \\ {}^{18}_{10}\mathrm{Ne} \to {}^{18}_{9}\mathrm{F}^-\, e^+ \qquad & \nu_e \\ & \nu_e \to \nu_\mu \to \mu^-. \end{aligned} \tag{98}$$

The golden transition can be searched for in this case by counting muons. It is not necessary to measure their charge, so the detector does not need to be magnetized.

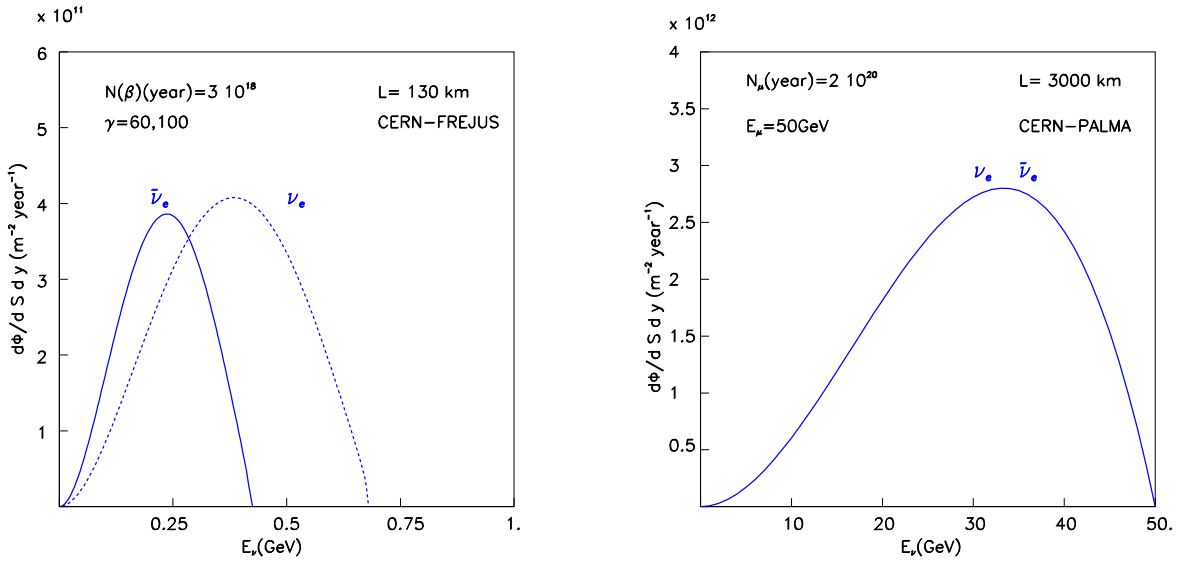


Fig. 42: Left: ν_e and $\bar{\nu}_e$ fluxes in the BB from 10^{18} ^{18}Ne/3×10^{18} ^{6}He ion decays per year at $\gamma = 100/60$ and $L = 130$ km. Right: ν_e and $\bar{\nu}_e$ fluxes at the NF from 2×10^{20} 50 GeVμ^-/μ^+ decays and $L = 3000$ km.

The neutrino fluxes ν_e and $\bar{\nu}_e$ at the NF or BB can be known with a very good accuracy, since they are easily obtained from the number of muons or ions decaying in the storage ring and the well-known muon or ion decay kinematics:

$$\left.\frac{d\Phi^{\rm NF}}{dSdy}\right|_{\theta\simeq 0} \simeq \frac{N_\mu}{\pi L^2} 12\gamma^2 y^2(1-y), \tag{99}$$

with $y = \frac{E_\nu}{E_\mu}$ and

$$\left.\frac{d\Phi^{\rm BB}}{dSdy}\right|_{\theta\simeq 0} \simeq \frac{N_\beta}{\pi L^2}\frac{\gamma^2}{g(y_e)} y^2(1-y)\sqrt{(1-y)^2-y_e^2}, \tag{100}$$

and $y = \frac{E_\nu}{2\gamma E_0}, y_e = m_e/E_0, g(y_e) \equiv \frac{1}{60}\left\{\sqrt{1-y_e^2}(2-9y_e^2-8y_e^4) + 15y_e^4 \log\left[\frac{y_e}{1-\sqrt{1-y_e^2}}\right]\right\}$. N_μ and N_β are the muons or ions decaying per year. Note that both fluxes increase with the γ factor of the parent particle as γ^2.

These fluxes are shown in Fig. 42 for two standard setups for the NF and the BB. Although the fluxes at the neutrino factory are larger by at least one order of magnitude, the need to magnetize the detector in the NF is a big limitation to how massive it can be in practice. In the case of the β beam no magnetization is needed, which opens the possibility to use very massive water Cherenkov detectors, like those that have been proposed to improve the limits on proton decay and to study supernova neutrinos [55].

In both the Neutrino Factory and the β-beam designs, the energy of the parent muon or ion (which is proportional to the average neutrino energy) can be optimized within a rather large range, since this is fixed by the acceleration scheme that is part of the machine design. Once the energy is fixed, the baseline is also fixed by the atmospheric oscillation length. This optimization is, however, a complex problem because there are often contradicting requirements in the maximization of the intensity, the minimization of backgrounds, having useful spectral information, measuring the *silver* channel in addition to the *golden* one, having sizeable matter effects, etc. This optimization was done for the NF some years ago and a muon energy of a few tens of GeV and a baseline of a few thousand kilometres is considered a

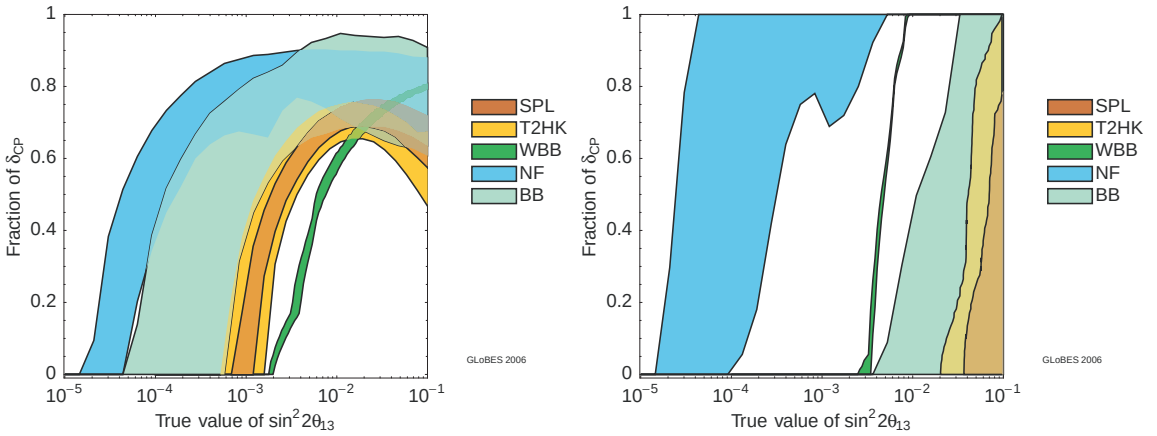


Fig. 43: Left: Sensitivity limit to leptonic CP violation in the plane $(\sin^2 2\theta_{13}, \delta)$ of superbeams (SPL, T2KHK), the wide band beam (WBB), Neutrino Factory (NF) and β beams (BB). The bands correspond to most/least conservative assumptions concerning the facility/detectors. Right: Sensitivity limits to the neutrino mass hierarchy in the same facilities. Taken from Ref. [56].

reference setup [46]. For the BB, a scenario with a neutrino beam of a few GeV and distances of a few hundred kilometers is close to optimal.

Figure 43 shows a comparison of the physics reach for CP violation and the neutrino hierarchy of the NF and BB complexes with other second-generation superbeams that have also been proposed as alternatives (SPL, T2HK, WBB). Even though this is probably not yet the end of the story as regards optimization/comparison, these plots show that reaching the realm of $\sin^2 2\theta_{13} \sim 10^{-4}$ will be possible in the future, both for leptonic CP violation and the neutrino hierarchy.

6 Leptogenesis

The Universe is made of matter. The matter–antimatter asymmetry is measured to be

$$\eta_B \equiv \frac{N_b - N_{\bar{b}}}{N_\gamma} \sim 6.15(25) \times 10^{-10} \, . \tag{101}$$

It has been known for a long time that all the ingredients to generate dynamically such an asymmetry from a symmetric initial state are present in the laws of particle physics. These ingredients were first put forward by Sakharov:

Baryon number violation

$B + L$ is anomalous in the SM [57] both with and without massive neutrinos, while $B - L$ is preserved if the light neutrinos are Dirac particles. At high T in the early Universe, $B + L$ violating transitions could be in thermal equilibrium [58] due to the thermal excitation of configurations with topological charge called sphalerons, see Fig. 44.

These processes violate baryon and lepton numbers by the same amount:

$$\Delta B = \Delta L. \tag{102}$$

If there are heavy Majorana singlets, as in the see-saw models, there is an additional source of L violation (and $B - L$). If a lepton charge is generated at temperatures where the sphalerons are still in thermal equilibrium, a baryon charge can be generated.

Deviation from thermal equilibrium

Sphalerons are in equilibrium for $T \geq 100$ GeV [59], which means that in order to get these processes out of equilibrium it is necessary to go to the electroweak phase transition.

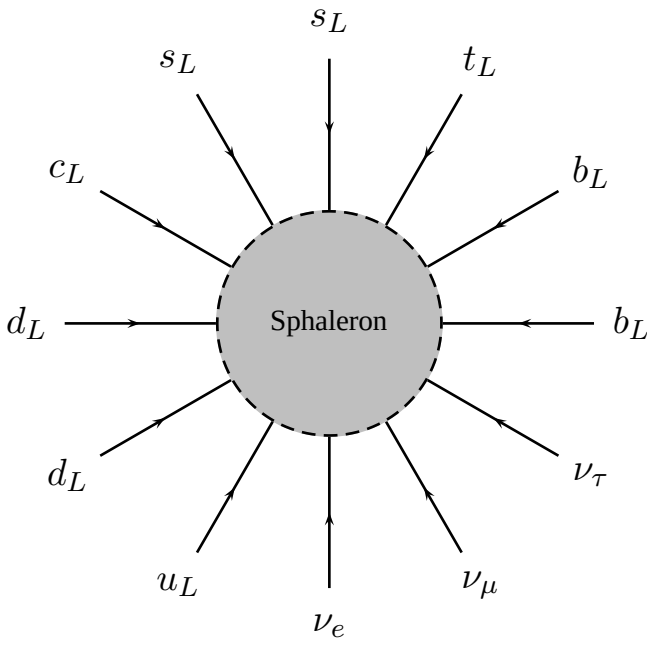


Fig. 44: Artistic view of a sphaleron

Electroweak baryogenesis which has been extensively studied both in the SM and in the most popular extensions like the MSSM, is currently disfavoured in the SM because the out-of-equilibrium condition is not well met: the electroweak phase transition is not strongly first order.

A different out-of-equilibrium condition is met in the L violation processes associated to the heavy Majorana singlets [60]. These singlets are in equilibrium until they decouple at a temperature similar to their masses. Since their masses must be significantly larger than the electroweak scale if we are to explain the smallness of neutrino masses, sphalerons are still in equilibrium when the heavy Majorana singlets decouple. Therefore if a lepton number is generated in their decay, inducing a lepton number abundance Y_L, the equilibrium of sphaleron processes implies that a baryon abundance will also be present [61]:

$$Y_B = aY_{B-L} = \frac{a}{a-1}Y_L \quad a = \frac{28}{79} \quad \text{in SM} . \tag{103}$$

C and CP violation

In order for lepton number to be generated in the decay of these Majorana singlets, it is necessary that CP and C be violated in the decays:

$$\epsilon_1 = \frac{\Gamma(N \to \Phi l) - \Gamma(N \to \Phi \bar{l})}{\Gamma(N \to \Phi l) + \Gamma(N \to \Phi \bar{l})} \neq 0 . \tag{104}$$

In fact this is generically the case since, as we have seen, there are new CP-violating phases in the neutrino mixing matrices which induce an asymmetry at the one-loop level (see Fig. 45).

These processes can then produce a net lepton asymmetry if the number distributions of the Majorana singlets, N_N, differ from the thermal ones. This can occur close to the decoupling temperature, when the density of the heavy neutrinos gets exponentially suppressed, but they are so weakly interacting that they cannot follow the fast depletion (in other words if the decay rate is slower than the expansion of the Universe close to the decoupling temperature) and

$$N_N > N_N^{\text{thermal}}. \tag{105}$$

This is shown in Fig. 46. The final asymmetry is given by

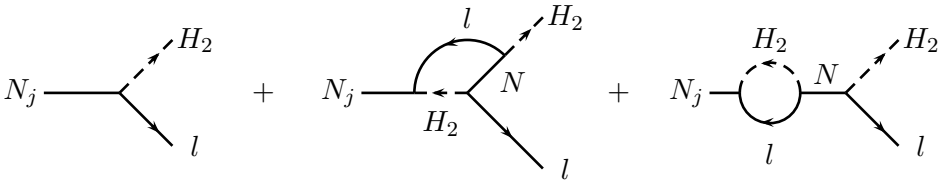


Fig. 45: Tree-level and one-loop diagrams contributing to heavy neutrino decays

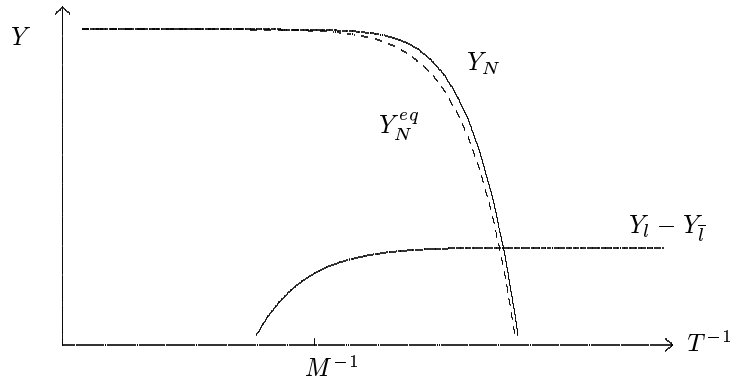


Fig. 46: Abundance of the heavy Majorana singlets at the decoupling temperature and the lepton number generated in the decay

$$Y_B = 10^{-2} \overbrace{\epsilon_1}^{\text{CP-asym}} \overbrace{\kappa}^{\text{eff. factor}} , \tag{106}$$

where κ is an efficiency factor which depends on the non-equilibrium dynamics. Therefore a relation between the baryon number of the Universe and the neutrino flavour parameters in ϵ_1 exists.

An interesting question is whether the baryon asymmetry can be predicted quantitatively from the measurements at low energies of the neutrino mass matrix. Unfortunately this is not the case generically because the asymmetry ϵ_1 depends on more parameters than those that are observable at low energies.

As we saw in Section 2.1, at least three heavy Majorana neutrinos of masses M_i are needed to give masses to the three light neutrinos. The asymmetry in the decay of the lighest of them in the minimal model with $M_{2,3} \gg M_1$ is [62]

$$\epsilon_1 = -\frac{3}{16\pi} \sum_i \frac{\mathrm{Im}[(\tilde{\lambda}_\nu^\dagger \tilde{\lambda}_\nu)^2_{i1}]}{(\tilde{\lambda}^\dagger \tilde{\lambda})_{11}} \frac{M_1}{M_i} . \tag{107}$$

Instead, at low energies, there is sensitivity only to the neutrino mass matrix:

$$\tilde{\lambda}_\nu \frac{1}{M_R} \tilde{\lambda}_\nu^T , \tag{108}$$

where M_R is the heavy Majorana mass matrix. The two combinations are different and the measurement of the matrix in Eq. (108) does not allow one to compute ϵ_1. This is because in general the number of parameters measurable at high energies in the see-saw model is larger than at low energies. The counting of parameters for n generations before and after integrating out the heavy fields is shown Table 4 (see Section 2.3 for explanations).

If the prediction of the lepton asymmetry is not possible, it should at least be possible to constrain the neutrino mass matrix, assuming that the lepton asymmetry explains the measured baryon asymmetry.

Table 4: Number of physical parameters in the see-saw model with n families and the same number of right-handed Majorana neutrinos at high and low energies

	Yukawas	Field redefinitons	$No.\ m$	$No.\ \theta$	$No.\ \phi$
see-saw $E \geq M_i$	$Y_l, Y_\nu, M_R = M_R^T$ $5n^2 + n$	$U(n)^3$ $\frac{3(n^2-n)}{2}, \frac{3(n^2+n)}{2}$	$3n$	$n^2 - n$	$n^2 - n$
see-saw $E \ll M_i$	$Y_l, \alpha_\nu^T = \alpha_\nu$ $3n^2 + n$	$U(n)^2$ $n^2 - n, n^2 + n$	$2n$	$\frac{n^2-n}{2}$	$\frac{n^2-n}{2}$

Indeed, various upper bounds can be derived on the generated asymmetry, through a bound on ϵ_1 or on κ. In particular ϵ_1 has been shown to satisfy

$$|\epsilon_1| \leq \frac{8}{16\pi} \frac{M_1}{v^2} |\Delta m^2_{\text{atm}}|^{1/2} \ , \tag{109}$$

and therefore leptogenesis in this model requires that the lightest heavy neutrino is rather heavy:

$$M_1 \geq \mathcal{O}(10^9 \text{ GeV}) \ . \tag{110}$$

A sufficiently large κ implies an upper bound on the lightest neutrino mass:

$$m_i \leq \mathcal{O}(\text{eV}). \tag{111}$$

For further details and references see Ref. [62].

7 Outlook for theory

One of the most important questions to resolve in neutrino physics is whether the origin of neutrino masses is a new physics scale and if so what this scale is. One can envisage various possibilities for such new physics, and the simplest is to assume that its associated energy scale is above the electroweak scale. It is well known, since the pioneering work of Weinberg [63], that the appropriate language to describe the low-energy effects of such new physics, no matter what it is, is that of *effective field theory*. The effects of *any* beyond-the-standard-model dynamics with a characteristic energy scale, $\Lambda \gg v$, can be described at low-energies, i.e., $E < \Lambda$, by the SM Lagrangian plus a tower of operators with mass dimension, $d > 4$, constructed out of the SM fields and satisfying all the gauge symmetries. Even though the number of such operators is infinite, they can be classified according to their dimension, d, since an operator of dimension d must be suppressed by the scale Λ^{d-4}, and therefore higher dimensionality means stronger suppression in the high-energy scale:

$$\mathcal{L} = \mathcal{L}_{\text{SM}} + \sum_i \frac{\alpha_i}{\Lambda} \mathcal{O}_i^{d=5} + \sum_i \frac{\beta_i}{\Lambda^2} \mathcal{O}_i^{d=6} + ... \tag{112}$$

Different fundamental theories correspond to different values for the *low-energy couplings* $\alpha_i, \beta_i, ...$, but the structure of the effective interactions is the same.

It turns out that the first operator in the list is the famous Weinberg operator of Eq. (10):

$$\mathcal{O}^{d=5} = \bar{L}_L^c \tilde{\Phi}^T \tilde{\Phi} L_L \ , \tag{113}$$

where $\tilde{\Phi}, L$ are the SM Higgs and lepton doublets, respectively. This operator is the only one with $d = 5$ in the SM, and, as we have seen, brings in three essential new features to the minimal SM:

– neutrino masses,
– lepton mixing,
– lepton number violation.

Upon spontaneous symmetry breaking, such an operator induces a neutrino mass matrix of the form

$$m_\nu = \alpha \frac{v^2}{\Lambda} \,, \tag{114}$$

where α is generically a matrix in flavour space. Neutrino masses are therefore expected to be naturally small if $\Lambda \gg v$.

If we assume that the neutrino masses we have measured are the result of this leading operator, one could ask the question: What type of new physics would induce such an interaction ? In the same way that one can conjecture the presence of a massive gauge boson from the Fermi four-fermion interaction, one can classify the extra degrees of freedom that can induce at tree-level Weinberg's interaction. It turns out that there are the three well-known possibilities as depicted in Fig. 47:

– type I see-saw: SM+ heavy singlet fermions [6],
– type II see-saw: SM + heavy triplet scalar [64],
– type III see-saw: SM + heavy triple fermions [65],

or combinations. The masses of the extra states define the scale Λ.

It is also possible that Weinberg's interaction is generated by new physics at higher orders, such as in the famous Zee model [66] and related ones [67]. In this case, the coupling α in Eq. (112) will be suppressed by loop factors $1/(16\pi^2)$.

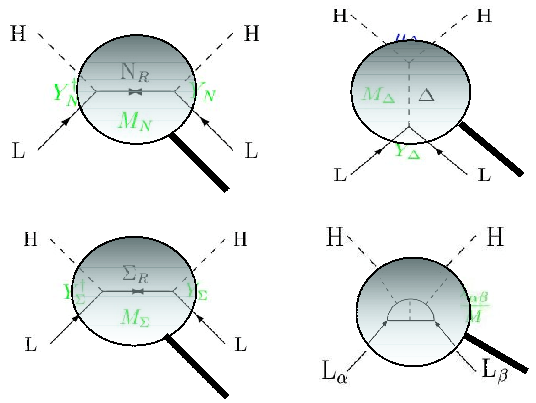


Fig. 47: Magnifying-glass view of Weinberg operator in see-saws Type I (top left), Type II (top right), Type III (bottom left) and Zee–Babu model (bottom right)

Unfortunately the measurement of neutrino masses alone will not tell us which of these possibilities is the one chosen by Nature. In particular, the measurement of Weinberg's interaction leaves behind an unresolved $\alpha \leftrightarrow \Lambda$ *degeneracy* that makes it impossible to know what the scale of the new physics is, even if we were to know the absolute value of neutrino masses.

Generically, however, the new physics will give other signals beyond Weinberg's operator. The next in importance are the $d = 6$ operators of Eq. (112) [68]. Recently the $d = 6$ operators induced at tree level in see-saw models of Types I to III have been worked out [69]. They give rise to a rich phenomenology that could discriminate between the models. In particular, they could induce beyond-the-standard-model signals in Z and W decays, deviations in the ρ parameter or the W mass, and mediate

rare lepton decays, as well as violations of universality and unitarity of the neutrino mass matrix. It would therefore be extremely important to search for these effects. Whether they are large enough to be observed or not depends strongly on how high the scale Λ is, since all these effects are suppressed by two powers of Λ.

As mentioned before, neutrino masses alone do not tell us what Λ is, but there are several theoretical prejudices of what this scale should be. The most popular one is to relate Λ to a grand-unification scale, given the intriguing fact that the seesaw-type ratio $\frac{v^2}{M_{\rm GUT}} \sim 0.01 - -0.1$ eV, in the right ballpark of a neutrino mass scale. Recently, however, it has been pointed out [70] that within see-saw models, and without supersymmetry, this choice would destabilize the electroweak scale, since the Higgs mass would receive quadratic loop corrections in Λ. A naturalness argument would then imply that $\Lambda < 10^7$ GeV, at least if there is no supersymmetry.

Another possibility is to consider Λ to be related to the electroweak scale, i.e., not far from it. After all, the electroweak scale is the only scale we are sure exits. The question is then if such a choice would be testable via the measurement of the $d = 6$ operators. The answer to this question is no in the simplest type I see-saw model, because in order to get neutrino masses in the right ballpark when $\Lambda \sim$ TeV, it is necessary to have extremely small Yukawa couplings, which suppress also the $d = 6$ operators to an unobservable level. Several recent works have discussed the possibility to have larger effects of the $d = 6$ operators [72, 71, 69]. One possibility is that realized in Zee-type models where $d = 5$ operators are forbidden at tree level and are therefore suppressed by loop factors, while $d = 6$ operators are allowed at tree level and therefore unsuppressed. A more radical possibility is the existence of two independent scales in Eq. (112), one that suppresses $d = 6$ operators, Λ_6, and another one, $\Lambda_5 \gg \Lambda_6$, that suppresses the $d = 5$ one. This possibility is not unnatural, because the $d = 5$ and $d = 6$ operators can be classified according to a a global symmetry: total lepton number. If we therefore assume that the scale at which lepton number is broken, $\Lambda_{\rm LN}$, is much higher than the scale at which lepton flavour violation, $\Lambda_{\rm LFV}$, is relevant, we can ensure that the $d = 5$ operator, that breaks lepton number, is suppressed by the former scale, $\Lambda_5 \sim \Lambda_{\rm LN}$, while the lepton-flavour effects induced by operators of $d = 6$ would be suppressed only by a lower scale $\Lambda_6 \sim \Lambda_{\rm LFV} << \Lambda_{\rm LN}$. The effective field theory describing such a possibility would look therefore like

$$\mathcal{L} = \mathcal{L}_{\rm SM} + \sum_i \frac{\alpha_i}{\Lambda_{\rm LN}} \mathcal{O}_i^{d=5} + \sum_i \frac{\beta_i}{\Lambda^2_{\rm LFV}} \mathcal{O}_i^{d=6} + ..., \tag{115}$$

where the operators that break lepton number and those that preserve this symmetry are generically suppressed by different scales. Such a possibility has recently been considered in the context of the popular Minimal Flavour Violation hypothesis [72]. The underlying rationale for such an assumption is not completely ad hoc, since in this context one could hope to explain two apparently contradictory facts

- common origin of lepton and quark family mixing at a scale $\Lambda_{\rm LFV}$,
- large gap between neutrino masses and remaining fermions since neutrino masses would be suppressed by $\Lambda_{\rm LN}$.

In fact this separation of scales is built-in in several of the models mentioned before. The simplest example being the type II see-saw model, where the scalar-triplet mass, M_Δ, is directly connected with the $\Lambda_{\rm LFV}$, while the scale of lepton number violation is M^2_Δ/μ, where μ is a dimensionful coupling in the scalar potential of the triplet. In fact, it is the separation of scales that makes the phenomenology of this model much richer at low energies than that of type I see-saw models in their simplest version.

If this possibility is realized, there would be many interesting consequences:

- lepton flavour violation could be measurable beyond neutrino oscillations,
- the scale of lepton flavour violation, $\Lambda_{\rm LFV}$, could be reached at the LHC.

In recent years a lot of activity has been devoted to studying possible signals of neutrino masses at the the LHC. Lepton number violation could give rise to spectacular signals at LHC, like same-charge lepton pairs [73]. This signal has been studied in detail recently in various see-saw models. In one-scale models of type I, neutrino masses restrict these processes to being highly suppressed beyond detectable levels [74]. However, the separation of scales mentioned before, allows light enough triplets in the type II see-saw to be pair-produced at LHC:

$$pp \to H^{++}H^{--} \to l^+l^+l^-l^-, \tag{116}$$

leading to the powerful signal of same-charge lepton pairs. Not only can the invariant mass be reconstructed from the two leptons pairs, but the flavour structure of the branching ratios to different leptons is in one-to-one correspondence with the flavour structure of the neutrino mass matrix. Therefore the putative measurement of these processes would provide direct information on the neutrino mass matrix [75].

Solving the flavour problem of the Standard Model is surely a quixotic enterprise and we shall need to explore as many avenues as we can. In recent years it has become increasingly clear that in addition to quark flavour factories, we can obtain very valuable information on different aspects of this puzzle also from LHC and lepton flavour factories.

8 Conclusions

The results of many beautiful experiments in the last decade have demonstrated beyond doubt that neutrinos are massive and mix. The standard 3ν scenario can explain in terms of four fundamental parameters all available data, except that of the unconfirmed signal of LSND. The lepton flavour sector of the Standard Model is expected to be at least as complex as the quark one, even though we know it only partially.

The structure of the neutrino spectrum and mixing is quite different from the one that has been observed for the quarks: there are large leptonic mixing angles and the neutrino masses are much smaller than those of the remaining leptons. These peculiar features of the lepton sector strongly suggest that leptons and quarks constitute two complementary approaches to understanding the origin of flavour in the Standard Model. In fact, the smallness of neutrino masses can be naturally understood if there is new physics beyond the electroweak scale.

Many fundamental questions remain to be answered in future neutrino experiments, and these can have very important implications for our understanding of the Standard Model and of what lies beyond: Are neutrinos Majorana particles? Are neutrino masses the result of a new physics scale? Is CP violated in the lepton sector? Could neutrinos be the seed of the matter–antimatter asymmetry in the Universe?

A rich experimental programme lies ahead where fundamental physics discoveries are very likely (almost warrantied). We can only hope that neutrinos will keep up with their old tradition and provide a window to what lies beyond the Standard Model.

References

[1] Lectures, *Field Theory and the Standard Model*, V. Rubakov at this School.

[2] C. Amsler *et al.*, Phys. Lett. B667 (2008) 1. http://pdg.lbl.gov.

[3] Ch. Weinheimer *et al.*, Phys. Lett. B460 (1999) 219; Erratum ibid. 464 (1999) 332; M. Lobashev *et al.*, Phys. Lett. B464 (1999) 227.

[4] K.A. Assamagan *et al.*, Phys. Rev. D53 (1996) 6065.

[5] R. Barate *et al.*, Eur. Phys. J. C2 (1998) 395.

[6] P. Minkowski, Phys. Lett. B 67 (1977) 421; M. Gell-Mann, P. Ramond and R. Slansky, in *Supergravity*, edited by P. van Nieuwenhuizen and D. Freedman, (North-Holland, Amsterdam, 1979),

p. 315; T. Yanagida, in *Proceedings of the Workshop on Unified Theories and Baryon Number in the Universe*, edited by O. Sawada and A. Sugamoto (KEK Report No. 79-18, Tsukuba, 1979), p. 95; R.N. Mohapatra and G. Senjanović, Phys. Rev. Lett. 44 (1980) 912.

[7] B. Pontecorvo, Zh. Eksp. Teor. Fiz. 33 (1957) 549; JETP, 6 (1958) 429; Zh. Eksp. Teor. Fiz. 33 (1958) 247.

[8] Z. Maki, M. Nakagawa and S. Sakata, Prog. Theor. Phys. 28 (1962) 870.

[9] L. Baudis *et al.*, Phys. Rev. Lett. 83(1999) 41.

[10] H.V. Klapdor-Kleingrothaus and I.V.Krivosheina, Mod. Phys. Lett. A21 (2006) 1547.

[11] C. Arnaboldi *et al.*, Phys. Rev. Lett. 95 (2005) 142501.

[12] R. Arnold *et al.*, Nucl. Phys. A781 (2007) 209.

[13] L. Wolfenstein, Phys. Rev. D17 (1978) 2369.

[14] S.P. Mikheev and A.Yu. Smirnov, Yad. Fiz. 42 (1985) 1441; Nuov. Cim. 9C (1986) 17.

[15] J. Bahcall, P.I. Krastev and A. Yu. Smirnov, JHEP 0105 (2001) 015. J.N. Bahcall, M.C. González-García and C. Peña-Garay, JHEP 0108 (2001) 014.

[16] J.N. Bahcall, M.H. Pinsonneault and S. Basu, Astrophys. J. 555 (2001) 990.

[17] B.T. Cleveland *et al.*, Astrophys. J. 496 (1998) 505.

[18] W. Hampel *et al.*, Phys. Lett. B447 (1999) 127.

[19] J.N. Abdurashitov *et al.*, J. Exp. Theor. Phys. 95 (2002) 181.

[20] Y. Fukuda *et al.*, Phys. Rev. Lett. 77 (1996) 1683.

[21] S. Fukuda *et al.*, Phys. Rev. Lett. 86 (2001) 5651 and 5656; Phys. Lett. B539 (2002) 179; M.B. Smy *et al.*, Phys. Rev. D69 (2004) 011104.

[22] Q.R. Ahmad *et al.*, Phys. Rev. Lett. 87 (2001) 071301; 89 (2002) 011301 and 011302; 92 (2004) 181301.

[23] B. Aharmim *et al.*, Phys. Rev. C72 (2005) 055502. B. Aharmim *et al*, Phys. Rev. Lett. 101 (2008) 111301.

[24] K. Eguchi *et al.*, Phys. Rev. Lett. 90 (2003) 021802.

[25] S. Abe *et al.*, Phys. Rev. Lett. 100 (2008) 221803.

[26] C. Arpesella *et al.*, Phys. Rev. Lett. 101 (2008) 091302.

[27] M. Honda, *et al.*, Phys. Rev. D 54 (1995) 4985. V. Agrawal, T.K. Gaisser, P. Lipari, and T. Stanev, Phys. Rev. D 53 (1996) 1314. M. Honda *et al.*, Phys. Rev. D 70 (2004) 043008.

[28] Y. Fukuda *et al.*, Phys. Lett. B335 (1994) 237.

[29] R. Becker-Szendy *et al.*, Nucl. Phys. (Proc. Suppl.) B38 (1995) 331.

[30] M. Sánchez *et al.*, Phys. Rev. D 68 (2003) 113004.

[31] M. Ambrosio *et al.*, Phys. Lett. B566 (2003) 35.

[32] Y. Fukuda *et al.*, Phys. Rev. Lett 81 (1998) 1562.

[33] M. Ishitsuka for the Super-Kamiokande Collaboration, Proceedings of the XXXIXth Rencontres de Moriond on Electroweak Interactions and Unified Theories (2004), hep-ex/0406076.

[34] Y. Ashie *et al.*, Phys. Rev. Lett. 93 (2004) 101801.

[35] M.H. Ahn, Phys. Rev. Lett 90 (2003) 041801.

[36] B. Achkar *et al.*, Nucl. Phys. B434 (1995) 503.

[37] F. Boehm *et al.*, Phys. Rev. D64 (2001) 112001.

[38] M. Apollonio *et al.*, Phys. Lett. B 466 (1999) 415; Eur. Phys. J. C27 (2003) 331.

[39] A. Aguilar *et al.*, Phys. Rev. D 64 (2001) 112007.

[40] B. Armbruster *et al.*, Phys. Rev. D65 (2002) 112001.

[41] A. A. Aguilar-Arevalo *et al.* [MiniBooNE Collaboration], Phys. Rev. Lett. 100 (2008) 032301.

[42] M.C. González-García and M. Maltoni, Phys. Rep. 460 (2008) 1.

[43] G. L. Fogli *et al.*, Phys. Rev. D 78 (2008) 033010 .

[44] A. Osipowicz *et al.*, hep-ex/0109033.

[45] For a recent review see P. Vogel, arXiv:0807.2457[hep-ph].

[46] A. Cervera *et al.*, Nucl Phys. B579 (2000) 17.

[47] A. Donini, D. Meloni and P. Migliozzi, Nucl. Phys. B646 (2002) 321.

[48] J. Burguet-Castell *et al.*, Nucl Phys. B608 (2001) 301.

[49] H. Minakata and H. Nunokawa, JHEP 0110 (2001) 1. G.L. Fogli and E. Lisi, Phys. Rev. D54 (1996) 3667. V. Barger, D. Marfatia and K. Whisnant, Phys. Rev. D65 (2002) 073023.

[50] H. Minakata *et al.*, Phys. Rev. D68 (2003) 033017.

[51] Y. Itow *et al.*, Nucl. Phys. Proc. Suppl. 111 (2002) 146.

[52] A. Para and M. Szleper, hep-ex/0110032; D. Ayres *et al.*, hep-ex/0210005.

[53] S. Geer, Phys. Rev. D57 (1998) 6989. A. De Rújula, M.B. Gavela, and P. Hernández, Nucl. Phys. B547 (1999) 21.

[54] P. Zucchelli, Phys. Lett. B532 (2002) 166.

[55] M. Goodman, *et al.*, Physics Potential and feasability of UNO, ed. D. Casper, C.K. Jung, C. McGrew and C. Yanagisawa, SBHEP01-3 (July 2001).

[56] The ISS Physics Working Group, "Physics at a future Neutrino Factory and super-beam facility," Editors S.F. King, K. Long, Y. Nagashima, B.L. Roberts, and O. Yasuda, arXiv:0710.4947 [hep-ph].

[57] G. 't Hooft, Phys. Rev. Lett. 37 (1976) 8.

[58] V.A. Kumin, V.A. Rubakov, and M.E. Shaposhnikov, Phys. Lett. B155 (1985) 36.

[59] D. Bödecker, M. Laine, and K. Rummukainen, Phys. Rev. D 61 (2000) 056003.

[60] M. Fukugita, and T. Yanagida, Phys. Lett. B 174 (1986) 45.

[61] J.A. Harvey and M.S. Turner, Phys. Rev. D 42 (1990) 3344.

[62] For a recent review and further references see S. Davidson, E. Nardi, and Y. Nir, Phys. Rep. 466 (2008) 105 [arXiv:0802.2962 [hep-ph]].

[63] S. Weinberg, Phys. Rev. Lett. 43 (1979) 1566.

[64] M. Magg and C. Wetterich, Phys. Lett. B94 (1980) 61; J. Schechter and J. W. F. Valle, Phys. Rev. D 22 (1980) 2227; C. Wetterich, Nucl. Phys. B187 (1981) 343; G. Lazarides, Q. Shafi, and C. Wetterich, Nucl Phys. B181 (1981) 287; R.N. Mohapatra and G. Senjanović, Phys. Rev. D23 (1981) 165.

[65] R. Foot, H. Lew, X.-G. He, and G.C. Joshi, Z. Phys. C44 (1989) 441; E. Ma, Phys. Rev. Lett. 81 (1998) 1171.

[66] A. Zee, Phys. Lett. B93 (1980) 389.

[67] A. Zee, Phys. Lett. B161 (1985) 141 and Nucl.Phys. B264 (1986) 99. K.S. Babu, Phys. Lett. B203 (1988) 132.

[68] W. Buchmuller and D. Wyler, Nucl. Phys. B 268 (1986) 621.

[69] A. Abada, C. Biggio, F. Bonnet, M. B. Gavela, and T. Hambye, JHEP 0712 (2007) 061.

[70] F. Vissani, Phys. Rev. D57 (1998) 7027; J. A. Casas, J. R. Espinosa, and I. Hi dalgo, JHEP 0411 (2004) 057.

[71] J. Kersten and A. Y. Smirnov, Phys. Rev. D 76 (2007) 073005.

[72] V. Cirigliano, B. Grinstein, G. Isidori, and M.B. Wise, Nucl. Phys. B 728 (2005) 121.

[73] W. Y. Keung and G. Senjanovic, Phys. Rev. Lett. 50 (1983) 1427.

[74] F. del Aguila, J. A. Aguilar-Saavedra, and R. Pittau, JHEP 0710 (2007) 047.

[75] T. Han and B. Zhang, Phys. Rev. Lett. 97 (2006) 171804. T. Han, B. Mukhopadhyaya, Z. Si, and

K. Wang, Phys. Rev. D 76 (2007) 075013. J. Garayoa and T. Schwetz, JHEP 0803 (2008) 009. M. Kadastik, M. Raidal, and L. Rebane, Phys. Rev. D 77 (2008) 115023. A. G. Akeroyd, M. Aoki, and H. Sugiyama, Phys. Rev. D 77 (2008) 075010. P. Fileviez Perez, T. Han, G. y. Huang, T. Li, and K. Wang, Phys. Rev. D 78 (2008) 015018.

Flavour physics and CP violation

Y. Nir
Weizmann Institute of Science, Rehovot, Israel

Abstract
This is a written version of a series of lectures aimed at graduate students in particle theory/string theory/particle experiment familiar with the basics of the Standard Model. We explain the many reasons for the interest in flavour physics. We describe flavour physics and the related CP violation within the Standard Model, and explain how the B-factories proved that the Kobayashi-Maskawa mechanism dominates the CP violation that is observed in meson decays. We explain the implications of flavour physics for new physics. We emphasize the "new physics flavour puzzle". As an explicit example, we explain how the recent measurements of $D^0-\overline{D}^0$ mixing constrain the supersymmetric flavour structure. We explain how the ATLAS and CMS experiments can solve the new physics flavour puzzle and perhaps shed light on the standard model flavour puzzle. Finally, we describe various interpretations of the neutrino flavour data and their impact on flavour models.

1 What is flavour?

The term '**flavours**' is used, in the jargon of particle physics, to describe several copies of the same gauge representation, namely several fields that are assigned the same quantum charges. Within the Standard Model, when thinking of its unbroken $SU(3)_{\rm C}\times U(1)_{\rm EM}$ gauge group, there are four different types of particles, each coming in three flavours:

- Up-type quarks in the $(3)_{+2/3}$ representation: u, c, t.
- Down-type quarks in the $(3)_{-1/3}$ representation: d, s, b.
- Charged leptons in the $(1)_{-1}$ representation: e, μ, τ.
- Neutrinos in the $(1)_0$ representation: ν_1, ν_2, ν_3.

The term '**flavour physics**' refers to interactions that distinguish between flavours. By definition, gauge interactions, namely interactions that are related to unbroken symmetries and mediated therefore by massless gauge bosons, do not distinguish among the flavours and do not constitute part of flavour physics. Within the Standard Model, flavour physics refers to the weak and Yukawa interactions.

The term '**flavour parameters**' refers to parameters that carry flavour indices. Within the Standard Model, these are the nine masses of the charged fermions and the four 'mixing parameters' (three angles and one phase) that describe the interactions of the charged weak-force carriers ($W^\pm$) with quark–antiquark pairs. If one augments the Standard Model with Majorana mass terms for the neutrinos, one should add to the list three neutrino masses and six mixing parameters (three angles and three phases) for the $W^\pm$ interactions for lepton–antilepton pairs.

The term '**flavour universal**' refers to interactions with couplings (or to flavour parameters) that are proportional to the unit matrix in flavour space. Thus, the strong and electromagnetic interactions are flavour universal[1]. An alternative term for 'flavour universal' is '**flavour blind**'.

The term '**flavour diagonal**' refers to interactions with couplings (or to flavour parameters) that are diagonal, but not necessarily universal, in the flavour space. Within the Standard Model, the Yukawa interactions of the Higgs particle are flavour diagonal in the mass basis.

[1] In the interaction basis, the weak interactions are also flavour universal, and one can identify the source of all flavour physics in the Yukawa interactions among the gauge-interaction eigenstates.

The term '**flavour changing**' refers to processes where the initial and final flavour-numbers (that is, the number of particles of a certain flavour minus the number of antiparticles of the same flavour) are different. In 'flavour-changing charged current' processes, both up-type and down-type flavours, and/or both charged lepton and neutrino flavours are involved. Examples are (i) muon decay via $\mu \to e\bar{\nu}_i\nu_j$, and (ii) $K^- \to \mu^-\bar{\nu}_j$ (which corresponds, at the quark level, to $s\bar{u} \to \mu^-\bar{\nu}_j$). Within the Standard Model, these processes are mediated by the W bosons and occur at tree level. In '**flavour-changing neutral current**' (FCNC) processes, either up-type or down-type flavours but not both, and/or either charged lepton or neutrino flavours but not both, are involved. Examples are (i) muon decay via $\mu \to e\gamma$ and (ii) $K_L \to \mu^+\mu^-$ (which corresponds, at the quark level, to $s\bar{d} \to \mu^+\mu^-$). Within the Standard Model, these processes do not occur at tree level, and are often highly suppressed.

Another useful term is '**flavour violation**'. We shall explain it later in these lectures.

2 Why is flavour physics interesting?

- Flavour physics can discover new physics or probe it before it is directly observed in experiments. Here are some examples from the past:
  - The smallness of $\frac{\Gamma(K_L\to\mu^+\mu^-)}{\Gamma(K^+\to\mu^+\nu)}$ led to the prediction of a fourth (the charm) quark.
  - The size of Δm_K led to a successful prediction of the charm mass.
  - The size of Δm_B led to a successful prediction of the top mass.
  - The measurement of ε_K led to the prediction of the third generation.
- CP violation is closely related to flavour physics. Within the Standard Model, there is a single CP-violating parameter, the Kobayashi–Maskawa phase $\delta_{\rm KM}$ [1]. Baryogenesis tells us, however, that there must exist new sources of CP violation. Measurements of CP violation in flavour-changing processes might provide evidence for such sources.
- The fine-tuning problem of the Higgs mass, and the puzzle of dark matter imply that there exists new physics at, or below, the TeV scale. If such new physics had a generic flavour structure, it would contribute to flavour-changing neutral current (FCNC) processes orders of magnitude above the observed rates. The question of why this does not happen constitutes the *new physics flavour puzzle*.
- Most of the charged fermion flavour parameters are small and hierarchical. The Standard Model does not provide any explanation of these features. This is the *Standard Model flavour puzzle*. The puzzle became even deeper after neutrino masses and mixings were measured because, so far, neither smallness nor hierarchy in these parameters have been established.

3 Flavour in the Standard Model

A model of elementary particles and their interactions is defined by the following ingredients: (i) The symmetries of the Lagrangian and the pattern of spontaneous symmetry breaking; (ii) The representations of fermions and scalars. The Standard Model (SM) is defined as follows:
(i) The gauge symmetry is

$$G_{\rm SM} = SU(3)_{\rm C} \times SU(2)_{\rm L} \times U(1)_{\rm Y}. \tag{1}$$

It is spontaneously broken by the VEV of a single Higgs scalar, $\phi(1,2)_{1/2}$ ($\langle\phi^0\rangle = v/\sqrt{2}$):

$$G_{\rm SM} \to SU(3)_{\rm C} \times U(1)_{\rm EM}. \tag{2}$$

(ii) There are three fermion generations, each consisting of five representations of $G_{\rm SM}$:

$$Q_{Li}(3,2)_{+1/6},\ \ U_{Ri}(3,1)_{+2/3},\ \ D_{Ri}(3,1)_{-1/3},\ \ L_{Li}(1,2)_{-1/2},\ \ E_{Ri}(1,1)_{-1}. \tag{3}$$

3.1 The interactions basis

The Standard Model Lagrangian, $\mathcal{L}_{\rm SM}$, is the most general renormalizable Lagrangian that is consistent with the gauge symmetry (1), the particle content (3) and the pattern of spontaneous symmetry breaking (2). It can be divided into three parts:

$$\mathcal{L}_{\rm SM} = \mathcal{L}_{\rm kinetic} + \mathcal{L}_{\rm Higgs} + \mathcal{L}_{\rm Yukawa}. \tag{4}$$

For the kinetic terms, to maintain gauge invariance, one has to replace the derivative with a covariant derivative:

$$D^\mu = \partial^\mu + ig_s G_a^\mu L_a + igW_b^\mu T_b + ig' B^\mu Y. \tag{5}$$

Here G_a^μ are the eight gluon fields, W_b^μ the three weak interaction bosons, and B^μ the single hypercharge boson. The L_a's are $SU(3)_{\rm C}$ generators (the 3×3 Gell-Mann matrices $\frac{1}{2}\lambda_a$ for triplets, 0 for singlets), the T_b's are $SU(2)_{\rm L}$ generators (the 2×2 Pauli matrices $\frac{1}{2}\tau_b$ for doublets, 0 for singlets), and the Y's are the $U(1)_{\rm Y}$ charges. For example, for the quark doublets Q_L, we have

$$\mathcal{L}_{\rm kinetic}(Q_L) = i\overline{Q_{Li}}\gamma_\mu \left(\partial^\mu + \frac{i}{2}g_s G_a^\mu \lambda_a + \frac{i}{2}gW_b^\mu \tau_b + \frac{i}{6}g' B^\mu\right)\delta_{ij} Q_{Lj}, \tag{6}$$

while for the lepton doublets L_L^I, we have

$$\mathcal{L}_{\rm kinetic}(L_L) = i\overline{L_{Li}}\gamma_\mu \left(\partial^\mu + \frac{i}{2}gW_b^\mu \tau_b - \frac{i}{2}g' B^\mu\right)\delta_{ij} L_{Lj}. \tag{7}$$

The unit matrix in flavour space, δ_{ij}, signifies that these parts of the interaction Lagrangian are flavour universal. In addition, they conserve CP.

The Higgs potential, which describes the scalar self-interactions, is given by

$$\mathcal{L}_{\rm Higgs} = \mu^2 \phi^\dagger \phi - \lambda(\phi^\dagger \phi)^2. \tag{8}$$

For the Standard Model scalar sector, where there is a single doublet, this part of the Lagrangian is also CP conserving.

The quark Yukawa interactions are given by

$$-\mathcal{L}_{\rm Y}^q = Y_{ij}^d \overline{Q_{Li}}\phi D_{Rj} + Y_{ij}^u \overline{Q_{Li}}\tilde{\phi} U_{Rj} + \text{h.c.}, \tag{9}$$

(where $\tilde{\phi} = i\tau_2 \phi^\dagger$) while the lepton Yukawa interactions are given by

$$-\mathcal{L}_{\rm Y}^\ell = Y_{ij}^e \overline{L_{Li}}\phi E_{Rj} + \text{h.c.} \tag{10}$$

This part of the Lagrangian is, in general, flavour dependent (that is, $Y^f \not\propto \mathbf{1}$) and CP violating.

3.2 Global symmetries

In the absence of the Yukawa matrices Y^d, Y^u and Y^e, the SM has a large $U(3)^5$ global symmetry:

$$G_{\rm global}(Y^{u,d,e} = 0) = SU(3)_q^3 \times SU(3)_\ell^2 \times U(1)^5, \tag{11}$$

where

$$\begin{aligned} SU(3)_q^3 &= SU(3)_Q \times SU(3)_U \times SU(3)_D, \\ SU(3)_\ell^2 &= SU(3)_L \times SU(3)_E, \\ U(1)^5 &= U(1)_B \times U(1)_L \times U(1)_Y \times U(1)_{\rm PQ} \times U(1)_E. \end{aligned} \tag{12}$$

Out of the five $U(1)$ charges, three can be identified with baryon number (B), lepton number (L), and hypercharge (Y), which are respected by the Yukawa interactions. The two remaining $U(1)$ groups can be identified with the PQ symmetry whereby the Higgs and D_R, E_R fields have opposite charges, and with a global rotation of E_R only.

The point that is important for our purposes is that $\mathcal{L}_{\text{kinetic}} + \mathcal{L}_{\text{Higgs}}$ respect the non-Abelian flavour symmetry $S(3)_q^3 \times SU(3)_\ell^2$, under which

$$Q_L \to V_Q Q_L, \quad U_R \to V_U U_R, \quad D_R \to V_D D_R, \quad L_L \to V_L L_L, \quad E_R \to V_E E_R, \tag{13}$$

where the V_i are unitary matrices. The Yukawa interactions (9) and (10) break the global symmetry,

$$G_{\text{global}}(Y^{u,d,e} \neq 0) = U(1)_B \times U(1)_e \times U(1)_\mu \times U(1)_\tau. \tag{14}$$

(Of course, the gauged $U(1)_Y$ also remains a good symmetry.) Thus, the transformations of Eq. (13) are not a symmetry of $\mathcal{L}_{\text{SM}}$. Instead, they correspond to a change of the interaction basis. These observations also offer an alternative way of defining flavour physics: it refers to interactions that break the $SU(3)^5$ symmetry (13). Thus, the term '**flavour violation**' is often used to describe processes or parameters that break the symmetry.

One can think of the quark Yukawa couplings as spurions that break the global $SU(3)_q^3$ symmetry (but are neutral under $U(1)_B$),

$$Y^u \sim (3, \bar{3}, 1)_{SU(3)_q^3}, \quad Y^d \sim (3, 1, \bar{3})_{SU(3)_q^3}, \tag{15}$$

and of the lepton Yukawa couplings as spurions that break the global $SU(3)_\ell^2$ symmetry (but are neutral under $U(1)_e \times U(1)_\mu \times U(1)_\tau$),

$$Y^e \sim (3, \bar{3})_{SU(3)_\ell^2}. \tag{16}$$

The spurion formalism is convenient for several purposes: parameter counting (see below), identification of flavour suppression factors (see Section 5), and the idea of minimal flavour violation (see Section 7).

3.3 Counting parameters

How many independent parameters are there in $\mathcal{L}_{\text{Y}}^q$? The two Yukawa matrices, Y^u and Y^d, are 3×3 and complex. Consequently, there are 18 real and 18 imaginary parameters in these matrices. Not all of them are, however, physical. The pattern of G_{global} breaking means that there is freedom to remove 9 real and 17 imaginary parameters (the number of parameters in three 3×3 unitary matrices minus the phase related to $U(1)_B$). For example, we can use the unitary transformations $Q_L \to V_Q Q_L$, $U_R \to V_U U_R$, and $D_R \to V_D D_R$ to lead to the following interaction basis:

$$Y^d = \lambda_d, \quad Y^u = V^\dagger \lambda_u, \tag{17}$$

where $\lambda_{d,u}$ are diagonal,

$$\lambda_d = \text{diag}(y_d, y_s, y_b), \quad \lambda_u = \text{diag}(y_u, y_c, y_t), \tag{18}$$

while V is a unitary matrix that depends on three real angles and one complex phase. We conclude that there are 10 quark flavour parameters: 9 real ones and a single phase. In the mass basis, we shall identify the nine real parameters as six quark masses and three mixing angles, while the single phase is δ_{KM}.

How many independent parameters are there in $\mathcal{L}_{\text{Y}}^\ell$? The Yukawa matrix Y^e is 3×3 and complex. Consequently, there are 9 real and 9 imaginary parameters in this matrix. There is, however, freedom to remove 6 real and 9 imaginary parameters (the number of parameters in two 3×3 unitary matrices minus the phases related to $U(1)^3$). For example, we can use the unitary transformations $L_L \to V_L L_L$ and $E_R \to V_E E_R$ to lead to the following interaction basis:

$$Y^e = \lambda_e = \text{diag}(y_e, y_\mu, y_\tau). \tag{19}$$

We conclude that there are three real lepton flavour parameters. In the mass basis, we shall identify these parameters as the three charged lepton masses. We must, however, modify the model when we take into account the evidence for neutrino masses.

3.4 The mass basis

Upon the replacement $\mathcal{R}e(\phi^0) \to \frac{v+H^0}{\sqrt{2}}$, the Yukawa interactions (9) give rise to the mass matrices

$$M_q = \frac{v}{\sqrt{2}} Y^q. \tag{20}$$

The mass basis corresponds, by definition, to diagonal mass matrices. We can always find unitary matrices V_{qL} and V_{qR} such that

$$V_{qL} M_q V_{qR}^\dagger = M_q^{\rm diag} \equiv \frac{v}{\sqrt{2}} \lambda_q. \tag{21}$$

The four matrices V_{dL}, V_{dR}, V_{uL}, and V_{uR} are then the ones required to transform to the mass basis. For example, if we start from the special basis (17), we have $V_{dL} = V_{dR} = V_{uR} = \mathbf{1}$ and $V_{uL} = V$. The combination $V_{uL} V_{dL}^\dagger$ is independent of the interaction basis from which we start this procedure.

We denote the left-handed quark mass eigenstates as U_L and D_L. The charged-current interactions for quarks [that is the interactions of the charged $SU(2)_{\rm L}$ gauge bosons $W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp i W_\mu^2)$], which in the interaction basis are described by (6), have a complicated form in the mass basis:

$$-\mathcal{L}_{W^\pm}^q = \frac{g}{\sqrt{2}} \overline{U_{Li}} \gamma^\mu V_{ij} D_{Lj} W_\mu^+ + \text{h.c.}\,, \tag{22}$$

where V is the 3×3 unitary matrix ($VV^\dagger = V^\dagger V = \mathbf{1}$) that appeared in Eq. (17). For a general interaction basis,

$$V = V_{uL} V_{dL}^\dagger. \tag{23}$$

V is the Cabibbo–Kobayashi–Maskawa (CKM) *mixing matrix* for quarks [1, 2]. As a result of the fact that V is not diagonal, the $W^\pm$ gauge bosons couple to quark mass eigenstates of different generations. Within the Standard Model, this is the only source of *flavour-changing* quark interactions.

Exercise 1: *Prove that, in the absence of neutrino masses, there is no mixing in the lepton sector.*

Exercise 2: *Prove that there is no mixing in the Z couplings. (In the jargon of physics, there are no flavour-changing neutral currents at tree level.)*

The detailed structure of the CKM matrix, its parametrization, and the constraints on its elements are described in Appendix A.

4 Testing CKM

Measurements of rates, mixing, and CP asymmetries in B decays in the two B factories, BaBar and Belle, and in the two Tevatron detectors, CDF and D0, signified a new era in our understanding of CP violation. The progress is both qualitative and quantitative. Various basic questions concerning CP and flavour violation have, for the first time, received answers based on experimental information. These questions include, for example,

- Is the Kobayashi–Maskawa mechanism at work (namely, is $\delta_{\rm KM} \neq 0$)?
- Does the KM phase dominate the observed CP violation?

As a first step, one may assume the SM and test the overall consistency of the various measurements. However, the richness of data from the B factories allows us to go a step further and answer these questions model independently, namely allowing new physics to contribute to the relevant processes. We here explain the way in which this analysis proceeds.

4.1 $S_{\psi K_S}$

The CP asymmetry in $B \to \psi K_S$ decays plays a major role in testing the KM mechanism. Before we explain the test itself, we should understand why the theoretical interpretation of the asymmetry is exceptionally clean, and what are the theoretical parameters on which it depends, within and beyond the Standard Model.

The CP asymmetry in neutral meson decays into final CP eigenstates f_{CP} is defined as follows:

$$\mathcal{A}_{f_{CP}}(t) \equiv \frac{d\Gamma/dt[\overline{B}^0_{\rm phys}(t) \to f_{CP}\,] - d\Gamma/dt[B^0{}_{\rm phys}(t) \to f_{CP}\,]}{d\Gamma/dt[\overline{B}^0_{\rm phys}(t) \to f_{CP}\,] + d\Gamma/dt[B^0{}_{\rm phys}(t) \to f_{CP}\,]}\,. \tag{24}$$

A detailed evaluation of this asymmetry is given in Appendix B. It leads to the following form:

$$\begin{aligned} \mathcal{A}_{f_{CP}}(t) &= S_{f_{CP}}\,\sin(\Delta m t) - C_{f_{CP}}\,\cos(\Delta m t), \\ S_{f_{CP}} &\equiv \frac{2\,\mathcal{I}m(\lambda_{f_{CP}})}{1+|\lambda_{f_{CP}}|^2}, \quad C_{f_{CP}} \equiv \frac{1-|\lambda_{f_{CP}}|^2}{1+|\lambda_{f_{CP}}|^2}\,, \end{aligned} \tag{25}$$

where

$$\lambda_{f_{CP}} = e^{-i\phi_B}(\overline{A}_{f_{CP}}/A_{f_{CP}})\,. \tag{26}$$

Here ϕ_B refers to the phase of M_{12} [see Eq. (B.23)]. Within the Standard Model, the corresponding phase factor is given by

$$e^{-i\phi_B} = (V^*_{tb}V_{td})/(V_{tb}V^*_{td})\,. \tag{27}$$

The decay amplitudes A_f and $\overline{A}_f$ are defined in Eq. (B.1).

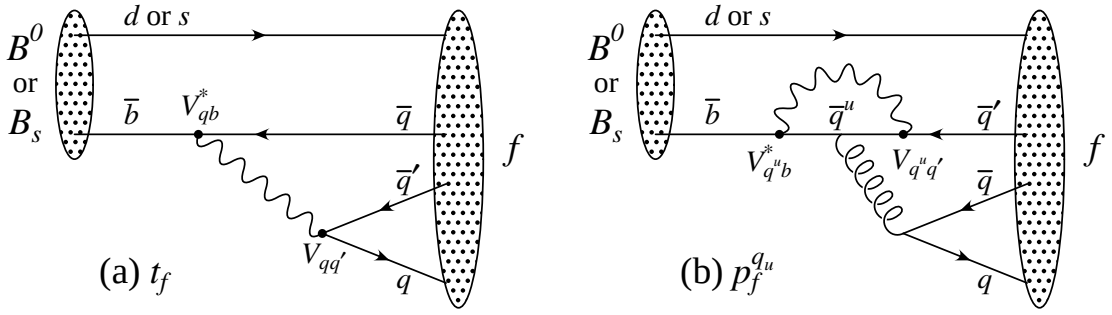


Fig. 1: Feynman diagrams for (a) tree and (b) penguin amplitudes contributing to $B^0 \to f$ or $B_s \to f$ via a $\bar{b} \to \bar{q}q\bar{q}'$ quark-level process

The $B^0 \to J/\psi K^0$ decay [3,4] proceeds via the quark transition $\bar{b} \to \bar{c}c\bar{s}$. There are contributions from both tree (t) and penguin (p^{q_u}, where $q_u = u, c, t$ is the quark in the loop) diagrams (see Fig. 1) which carry different weak phases:

$$A_f = (V^*_{cb}V_{cs})\,t_f + \sum_{q_u=u,c,t} \left(V^*_{q_u b}V_{q_u s}\right) p^{q_u}_f\,. \tag{28}$$

(The distinction between tree and penguin contributions is a heuristic one, the separation by the operator that enters is more precise. For a detailed discussion of the more complete operator product approach, which also includes higher order QCD corrections, see, for example, Ref. [5].) Using CKM unitarity, these decay amplitudes can always be written in terms of just two CKM combinations:

$$A_{\psi K} = (V^*_{cb}V_{cs})\,T_{\psi K} + (V^*_{ub}V_{us})\,P^u_{\psi K}, \tag{29}$$

where $T_{\psi K} = t_{\psi K} + p^c_{\psi K} - p^t_{\psi K}$ and $P^u_{\psi K} = p^u_{\psi K} - p^t_{\psi K}$. A subtlety arises in this decay that is related to the fact that $B^0 \to J/\psi K^0$ and $\overline{B}^0 \to J/\psi \overline{K}^0$. A common final state, e.g., $J/\psi K_S$, can

be reached via K^0–$\overline{K}^0$ mixing. Consequently, the phase factor corresponding to neutral K mixing, $e^{-i\phi_K} = (V_{cd}^* V_{cs})/(V_{cd} V_{cs}^*)$, plays a role:

$$\frac{\overline{A}_{\psi K_S}}{A_{\psi K_S}} = -\frac{(V_{cb}V_{cs}^*)\,T_{\psi K} + (V_{ub}V_{us}^*)\,P^u_{\psi K}}{\left(V_{cb}^*V_{cs}\right)\,T_{\psi K} + \left(V_{ub}^*V_{us}\right)\,P^u_{\psi K}} \times \frac{V_{cd}^*V_{cs}}{V_{cd}V_{cs}^*}. \tag{30}$$

The crucial point is that, for $B \to J/\psi K_S$ and other $\bar{b} \to \bar{c}c\bar{s}$ processes, we can neglect the P^u contribution to $A_{\psi K}$, in the SM, to an approximation that is better than one per cent:

$$|P^u_{\psi K}/T_{\psi K}| \times |V_{ub}/V_{cb}| \times |V_{us}/V_{cs}| \sim (\text{loop factor}) \times 0.1 \times 0.23 \lesssim 0.005. \tag{31}$$

Thus, to an accuracy of better than one per cent,

$$\lambda_{\psi K_S} = \left(\frac{V_{tb}^*V_{td}}{V_{tb}V_{td}^*}\right)\left(\frac{V_{cb}V_{cd}^*}{V_{cb}^*V_{cd}}\right) = -e^{-2i\beta}, \tag{32}$$

where β is defined in Eq. (A.9), and consequently

$$S_{\psi K_S} = \sin 2\beta, \quad C_{\psi K_S} = 0\ . \tag{33}$$

(Below the per cent level, several effects modify this equation [6–9].)

Exercise 3: *Show that, if the $B \to \pi\pi$ decays were dominated by tree diagrams, then $S_{\pi\pi} = \sin 2\alpha$.*

Exercise 4: *Estimate the accuracy of the predictions $S_{\phi K_S} = \sin 2\beta$ and $C_{\phi K_S} = 0$.*

When we consider extensions of the SM, we still do not expect any significant new contribution to the tree level decay, $b \to c\bar{c}s$, beyond the SM W-mediated diagram. Thus the expression $\bar{A}_{\psi K_S}/A_{\psi K_S} = (V_{cb}V_{cd}^*)/(V_{cb}^*V_{cd})$ remains valid, though the approximation of neglecting sub-dominant phases can be somewhat less accurate than Eq. (31). On the other hand, M_{12}, the B^0–$\overline{B}^0$ mixing amplitude, can in principle get large and even dominant contributions from new physics. We can parametrize the modification to the SM in terms of two parameters, r_d^2 signifying the change in magnitude, and $2\theta_d$ signifying the change in phase:

$$M_{12} = r_d^2\, e^{2i\theta_d}\, M_{12}^{\rm SM}(\rho,\eta). \tag{34}$$

This leads to the following generalization of Eq. (33):

$$S_{\psi K_S} = \sin(2\beta + 2\theta_d), \quad C_{\psi K_S} = 0\ . \tag{35}$$

The experimental measurements give the following ranges [10]:

$$S_{\psi K_S} = 0.671 \pm 0.024, \quad C_{\psi K_S} = 0.005 \pm 0.019\ . \tag{36}$$

4.2 Self-consistency of the CKM assumption

The three-generation Standard Model has room for CP violation, through the KM phase in the quark mixing matrix. Yet, one would like to make sure that CP is indeed violated by the SM interactions, namely that $\sin\delta_{\rm KM} \neq 0$. If we establish that this is the case, we would further like to know whether the SM contributions to CP violating observables are dominant. More quantitatively, we would like to put an upper bound on the ratio between the new physics and the SM contributions.

As a first step, one can assume that flavour-changing processes are fully described by the SM, and check the consistency of the various measurements with this assumption. There are four relevant mixing parameters, which can be taken to be the Wolfenstein parameters λ, A, ρ, and η defined in Eq. (A.4). The values of λ and A are known rather accurately [11] from, respectively, $K \to \pi\ell\nu$ and $b \to c\ell\nu$ decays:

$$\lambda = 0.2257 \pm 0.0010, \quad A = 0.814 \pm 0.022. \tag{37}$$

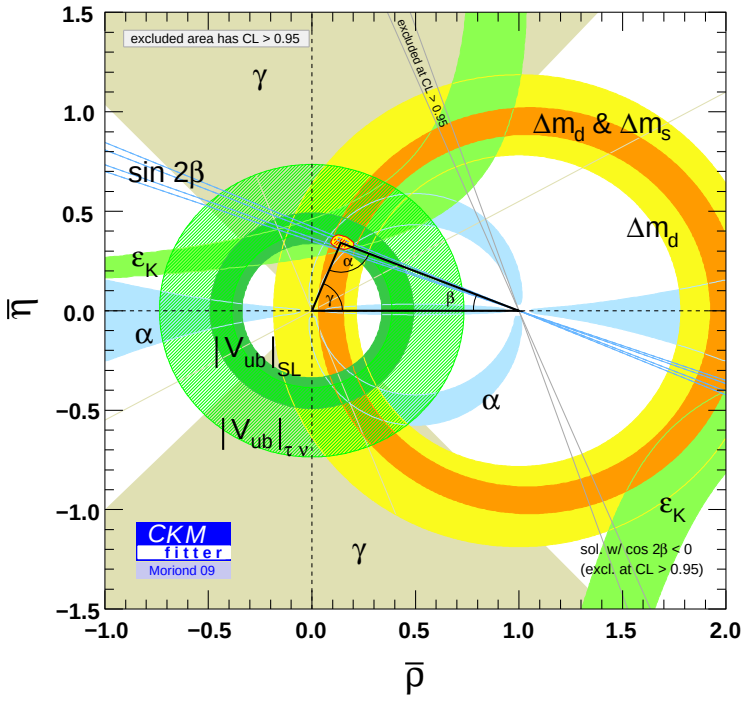


Fig. 2: Allowed region in the ρ–η plane. Superimposed are the individual constraints from charmless semileptonic B decays ($|V_{ub}/V_{cb}|$), mass differences in the B^0 (Δm_d) and B_s (Δm_s) neutral meson systems, and CP violation in $K \to \pi\pi$ (ε_K), $B \to \psi K$ ($\sin 2\beta$), $B \to \pi\pi, \rho\pi, \rho\rho$ (α), and $B \to DK$ (γ). Taken from Ref. [12].

Then, one can express all the relevant observables as a function of the two remaining parameters, ρ and η, and check whether there is a range in the ρ–η plane that is consistent with all measurements. The list of observables includes the following:

- The rates of inclusive and exclusive charmless semileptonic B decays depend on $|V_{ub}|^2 \propto \rho^2 + \eta^2$.
- The CP asymmetry in $B \to \psi K_S$, $S_{\psi K_S} = \sin 2\beta = \frac{2\eta(1-\rho)}{(1-\rho)^2+\eta^2}$.
- The rates of various $B \to DK$ decays depend on the phase γ, where $e^{i\gamma} = \frac{\rho+i\eta}{\rho^2+\eta^2}$.
- The rates of various $B \to \pi\pi, \rho\pi, \rho\rho$ decays depend on the phase $\alpha = \pi - \beta - \gamma$.
- The ratio between the mass splittings in the neutral B and B_s systems is sensitive to $|V_{td}/V_{ts}|^2 = \lambda^2[(1-\rho)^2 + \eta^2]$.
- The CP violation in $K \to \pi\pi$ decays, ϵ_K, depends in a complicated way on ρ and η.

The resulting constraints are shown in Fig. 2.

The consistency of the various constraints is impressive. In particular, the following ranges for ρ and η can account for all the measurements [11]:

$$\rho = 0.135^{+0.031}_{-0.016}, \quad \eta = 0.349 \pm 0.017. \tag{38}$$

One can then make the following statement [13]:
Very likely, CP violation in flavour-changing processes is dominated by the Kobayashi–Maskawa phase.

In the next two subsections, we explain how we can remove the phrase ‘very likely’ from this statement, and how we can quantify the KM dominance.

4.3 Is the Kobayashi–Maskawa mechanism at work?

In proving that the KM mechanism is at work, we assume that charged-current tree-level processes are dominated by the W-mediated SM diagrams (see, for example, Ref. [14]). This is a very plausible assumption. I am not aware of any viable well-motivated model where this assumption is not valid. Thus we can use all tree-level processes and fit them to ρ and η, as we did before. The list of such processes includes the following:

1. Charmless semileptonic B-decays, $b \to u\ell\nu$, measure R_u [see Eq. (A.8)].
2. $B \to DK$ decays, which go through the quark transitions $b \to c\bar{u}s$ and $b \to u\bar{c}s$, measure the angle γ [see Eq. (A.9)].
3. $B \to \rho\rho$ decays (and, similarly, $B \to \pi\pi$ and $B \to \rho\pi$ decays) go through the quark transition $b \to u\bar{u}d$. With an isospin analysis, one can determine the relative phase between the tree decay amplitude and the mixing amplitude. By incorporating the measurement of $S_{\psi K_S}$, one can subtract the phase from the mixing amplitude, finally providing a measurement of the angle γ [see Eq. (A.9)].

In addition, we can use loop processes, but then we must allow for new physics contributions, in addition to the (ρ,η)-dependent SM contributions. Of course, if each such measurement adds a separate mode-dependent parameter, then we do not gain anything by using this information. However, there are a number of observables where the only relevant loop process is B^0–$\overline{B}^0$ mixing. The list includes $S_{\psi K_S}$, Δm_B, and the CP asymmetry in semileptonic B decays:

$$\begin{aligned} S_{\psi K_S} &= \sin(2\beta + 2\theta_d), \\ \Delta m_B &= r_d^2 (\Delta m_B)^{\text{SM}}, \\ \mathcal{A}_{\text{SL}} &= -\mathcal{R}e\left(\frac{\Gamma_{12}}{M_{12}}\right)^{\text{SM}} \frac{\sin 2\theta_d}{r_d^2} + \mathcal{I}m\left(\frac{\Gamma_{12}}{M_{12}}\right)^{\text{SM}} \frac{\cos 2\theta_d}{r_d^2}. \end{aligned} \tag{39}$$

As explained above, such processes involve two new parameters [see Eq. (34)]. Since there are three relevant observables, we can further tighten the constraints in the (ρ,η) plane. Similarly, one can use measurements related to B_s–$\overline{B}_s$ mixing. One gains three new observables at the cost of two new parameters (see, for example, Ref. [15]).

The results of such a fit, projected on the ρ–η plane, can be seen in Fig. 3. It gives [12]

$$\eta = 0.44^{+0.05}_{-0.23} \ \ (3\sigma). \tag{40}$$

[A similar analysis in Ref. [16] obtains the 3σ range $(0.31\text{–}0.46)$.] It is clear that $\eta \neq 0$ is well established:

The Kobayashi–Maskawa mechanism of CP violation is at work.

Another way to establish that CP is violated by the CKM matrix is to find, within the same procedure, the allowed range for $\sin 2\beta$ [16]:

$$\sin 2\beta^{\text{tree}} = 0.76 \pm 0.04. \tag{41}$$

(Reference [12] finds $0.82^{+0.02}_{-0.13}$.) Thus, $\beta \neq 0$ is well established.

The consistency of the experimental results (36) with the SM predictions (33,41) means that the KM mechanism of CP violation dominates the observed CP violation. In the next subsection, we make this statement more quantitative.

4.4 How much can new physics contribute to B^0–$\overline{B}^0$ mixing?

All that we need to do in order to establish whether the SM dominates the observed CP violation, and to put an upper bound on the new physics contribution to B^0–$\overline{B}^0$ mixing, is to project the results of

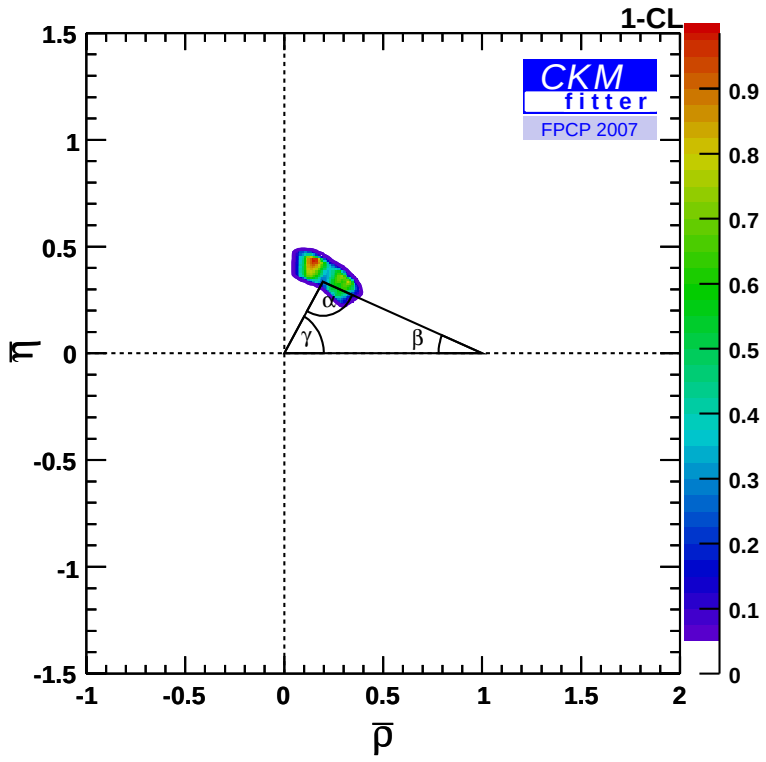


Fig. 3: The allowed region in the ρ–η plane, assuming that tree diagrams are dominated by the Standard Model [12]

the fit performed in the previous subsection on the r_d^2–$2\theta_d$ plane. If we find that $\theta_d \ll \beta$, then the SM dominance in the observed CP violation will be established. The constraints are shown in Fig. 4(a). Indeed, $\theta_d \ll \beta$.

An alternative way to present the data is to use the h_d, σ_d parametrization,

$$r_d^2 e^{2i\theta_d} = 1 + h_d e^{2i\sigma_d}. \tag{42}$$

While the r_d, θ_d parameters give the relation between the full mixing amplitude and the SM one, and are convenient to apply to the measurements, the h_d, σ_d parameters give the relation between the new physics and SM contributions, and are more convenient in testing theoretical models:

$$h_d e^{i\sigma_d} = \frac{M_{12}^{\rm NP}}{M_{12}^{\rm SM}}. \tag{43}$$

The constraints in the h_d–σ_d plane are shown in Fig. 4(b). We can make the following two statements:

1. A new physics contribution to the B^0–$\overline{B}^0$ mixing amplitude that carries a phase that is significantly different from the KM phase is constrained to lie below the 20–30% level.
2. A new physics contribution to the B^0–$\overline{B}^0$ mixing amplitude which is aligned with the KM phase is constrained to be at most comparable to the CKM contribution.

One can reformulate these statements as follows:

1. The KM mechanism dominates CP violation in B^0–$\overline{B}^0$ mixing.
2. The CKM mechanism is a major player in B^0–$\overline{B}^0$ mixing.

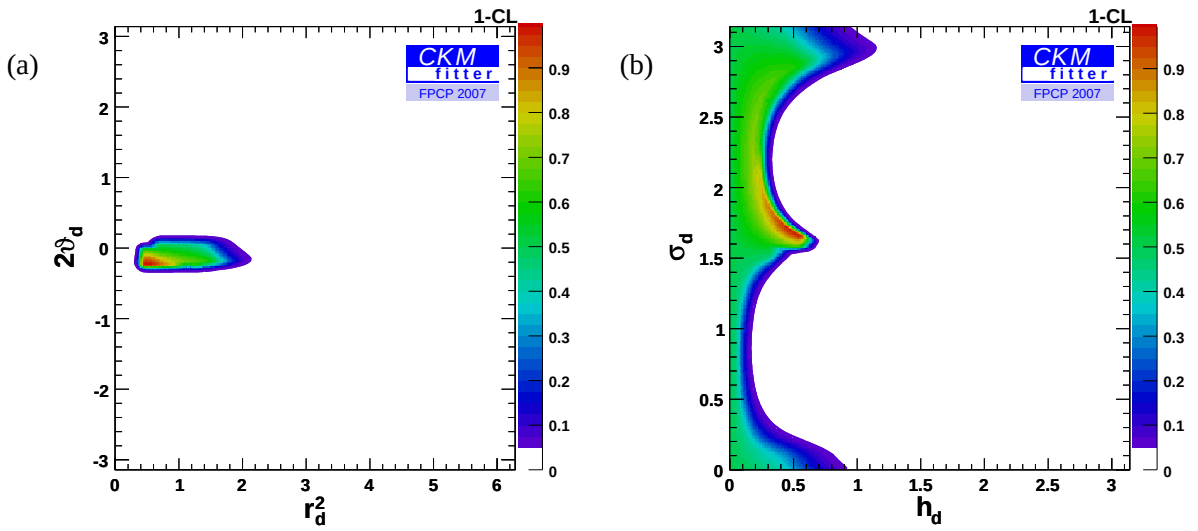


Fig. 4: Constraints in the (a) r_d^2–$2\theta_d$ plane, and (b) h_d–σ_d plane, assuming that new physics contributions to tree-level processes are negligible [12]

5 The new physics flavour puzzle

It is clear that the Standard Model is not a complete theory of Nature:

1. It does not include gravity, and therefore it cannot be valid at energy scales above $m_{\text{Planck}} \sim 10^{19}$ GeV.
2. It does not allow for neutrino masses, and therefore it cannot be valid at energy scales above $m_{\text{seesaw}} \sim 10^{15}$ GeV.
3. The fine-tuning problem of the Higgs mass and the puzzle of dark matter suggest that the scale where the SM is replaced with a more fundamental theory is actually much lower, $\Lambda_{\text{NP}} \lesssim 1$ TeV.

Given that the SM is only an effective low-energy theory, non-renormalizable terms must be added to $\mathcal{L}_{\text{SM}}$ of Eq. (4). These are terms of dimension higher than four in the fields which, therefore, have couplings that are inversely proportional to the scale of new physics Λ_{NP}. For example, the lowest-dimension non-renormalizable terms are dimension five:

$$-\mathcal{L}_{\text{Yukawa}}^{\text{dim}-5} = \frac{Z_{ij}^{\nu}}{\Lambda_{\text{NP}}} L_{Li}^I L_{Lj}^I \phi\phi + \text{h.c.} \tag{44}$$

These are the seesaw terms, leading to neutrino masses. We shall return to the topic of neutrino masses in Section 8.

Exercise 5: *How does the global symmetry breaking pattern (14) change when (44) is taken into account?*

Exercise 6: *What is the number of physical lepton flavour parameters in this case? Identify these parameters in the mass basis.*

As concerns quark flavour physics, consider, for example, the following dimension-six, four-fermion, flavour-changing operators:

$$\mathcal{L}_{\Delta F=2} = \frac{z_{sd}}{\Lambda_{\text{NP}}^2}(\overline{d_L}\gamma_\mu s_L)^2 + \frac{z_{cu}}{\Lambda_{\text{NP}}^2}(\overline{c_L}\gamma_\mu u_L)^2 + \frac{z_{bd}}{\Lambda_{\text{NP}}^2}(\overline{d_L}\gamma_\mu b_L)^2 + \frac{z_{bs}}{\Lambda_{\text{NP}}^2}(\overline{s_L}\gamma_\mu b_L)^2. \tag{45}$$

Each of these terms contributes to the mass splitting between the corresponding two neutral mesons. For example, the term $\mathcal{L}_{\Delta B=2} \propto (\overline{d_L}\gamma_\mu b_L)^2$ contributes to Δm_B, the mass difference between the two

neutral B-mesons. We use $M_{12}^B = \frac{1}{2m_B}\langle B^0|\mathcal{L}_{\Delta F=2}|\overline{B}^0\rangle$ and

$$\langle B^0|(\overline{d_{La}}\gamma^\mu b_{La})(\overline{d_{Lb}}\gamma_\mu b_{Lb})|\overline{B}^0\rangle = -\frac{1}{3}m_B^2 f_B^2 B_B. \tag{46}$$

Analogous expressions hold for the other neutral mesons[2]. This leads to $\Delta m_B/m_B = 2|M_{12}^B|/m_B \sim (|z_{bd}|/3)(f_B/\Lambda_{\rm NP})^2$. Experiments give, for CP conserving observables (the experimental evidence for Δm_D is at the 3σ level):

$$\begin{aligned}
\Delta m_K/m_K &\sim 7.0\times 10^{-15},\\
\Delta m_D/m_D &\sim 8.7\times 10^{-15},\\
\Delta m_B/m_B &\sim 6.3\times 10^{-14},\\
\Delta m_{B_s}/m_{B_s} &\sim 2.1\times 10^{-12},
\end{aligned} \tag{47}$$

and for CP violating ones

$$\begin{aligned}
\epsilon_K &\sim 2.3\times 10^{-3},\\
A_\Gamma/y_{\rm CP} &\lesssim 0.2,\\
S_{\psi K_S} &= 0.67\pm 0.02,\\
S_{\psi\phi} &\lesssim 1.
\end{aligned} \tag{48}$$

These measurements give then the following constraints:

$$\Lambda_{\rm NP} \gtrsim \begin{cases}
\sqrt{z_{sd}}\, 1\times 10^3 \text{ TeV} & \Delta m_K\\
\sqrt{z_{cu}}\, 1\times 10^3 \text{ TeV} & \Delta m_D\\
\sqrt{z_{bd}}\, 4\times 10^2 \text{ TeV} & \Delta m_B\\
\sqrt{z_{bs}}\, 7\times 10^1 \text{ TeV} & \Delta m_{B_s}
\end{cases} \tag{49}$$

and, for maximal phases,

$$\Lambda_{\rm NP} \gtrsim \begin{cases}
\sqrt{z_{sd}}\, 2\times 10^4 \text{ TeV} & \epsilon_K\\
\sqrt{z_{cu}}\, 3\times 10^3 \text{ TeV} & A_\Gamma\\
\sqrt{z_{bd}}\, 8\times 10^2 \text{ TeV} & S_{\psi K}\\
\sqrt{z_{bs}}\, 7\times 10^1 \text{ TeV} & S_{\psi\phi}
\end{cases} \tag{50}$$

If the new physics has a generic flavour structure, that is $z_{ij} = \mathcal{O}(1)$, then its scale must be above 10^3–10^4 TeV (or, if the leading contributions involve electroweak loops, above 10^2–10^3 TeV).[3]

If indeed $\Lambda_{\rm NP} \gg$ TeV, it means that we have misinterpreted the hints from the fine-tuning problem and the dark matter puzzle. There is, however, another way to look at these constraints:

$$\begin{aligned}
z_{sd} &\lesssim 8\times 10^{-7}\,(\Lambda_{\rm NP}/\text{TeV})^2,\\
z_{cu} &\lesssim 5\times 10^{-7}\,(\Lambda_{\rm NP}/\text{TeV})^2,\\
z_{bd} &\lesssim 5\times 10^{-6}\,(\Lambda_{\rm NP}/\text{TeV})^2,\\
z_{bs} &\lesssim 2\times 10^{-4}\,(\Lambda_{\rm NP}/\text{TeV})^2,
\end{aligned} \tag{51}$$

$$z_{sd}^I \lesssim 6\times 10^{-9}\,(\Lambda_{\rm NP}/\text{TeV})^2,$$

[2]The PDG [11] quotes the following values, extracted from leptonic charged meson decays: $f_K \approx 0.16$ GeV, $f_D \approx 0.23$ GeV, $f_B \approx 0.18$ GeV. We further use $f_{B_s} \approx 0.20$ GeV.

[3]The bounds from the corresponding four-fermi terms with LR structure, instead of the LL structure of Eq. (45), are even stronger.

$$\begin{aligned} z^I_{cu} &\lesssim 1\times10^{-7}\,(\Lambda_{\rm NP}/{\rm TeV})^2,\\ z^I_{bd} &\lesssim 1\times10^{-6}\,(\Lambda_{\rm NP}/{\rm TeV})^2,\\ z^I_{bs} &\lesssim 2\times10^{-4}\,(\Lambda_{\rm NP}/{\rm TeV})^2. \end{aligned} \tag{52}$$

It could be that the scale of new physics is of order TeV, but its flavour structure is far from generic.

One can use that language of effective operators also for the SM, integrating out all particles significantly heavier than the neutral mesons (that is, the top, the Higgs, and the weak gauge bosons). Thus the scale is $\Lambda_{\rm SM}\sim m_W$. Since the leading contributions to neutral meson mixings come from box diagrams, the z_{ij} coefficients are suppressed by α_2^2. To identify the relevant flavour suppression factor, one can employ the spurion formalism. For example, the flavour transition that is relevant to B^0–$\overline{B}^0$ mixing involves $\overline{d_L}b_L$ which transforms as $(8,1,1)_{SU(3)^3_q}$. The leading contribution must then be proportional to $(Y^uY^{u\dagger})_{13}\propto y_t^2V_{tb}V_{td}^*$. Indeed, an explicit calculation (using VIA for the matrix element and neglecting QCD corrections) gives[4]

$$\frac{2M^B_{12}}{m_B}\approx-\frac{\alpha_2^2}{12}\frac{f_B^2}{m_W^2}S_0(x_t)(V_{tb}V_{td}^*)^2, \tag{53}$$

where $x_i=m_i^2/m_W^2$ and

$$S_0(x)=\frac{x}{(1-x)^2}\left[1-\frac{11x}{4}+\frac{x^2}{4}-\frac{3x^2\ln x}{2(1-x)}\right]. \tag{54}$$

Similar spurion analyses, or explicit calculations, allow us to extract the weak and flavour suppression factors that apply in the SM:

$$\begin{aligned} \mathcal{I}m(z^{\rm SM}_{sd}) &\sim \alpha_2^2y_t^2|V_{td}V_{ts}|^2\sim1\times10^{-10},\\ z^{\rm SM}_{sd} &\sim \alpha_2^2y_c^2|V_{cd}V_{cs}|^2\sim5\times10^{-9},\\ z^{\rm SM}_{bd} &\sim \alpha_2^2y_t^2|V_{td}V_{tb}|^2\sim7\times10^{-8},\\ z^{\rm SM}_{bs} &\sim \alpha_2^2y_t^2|V_{ts}V_{tb}|^2\sim2\times10^{-6}. \end{aligned} \tag{55}$$

(We did not include $z^{\rm SM}_{cu}$ in the list because it requires a more detailed consideration. The naively leading short distance contribution is $\propto\alpha_2^2(y_s^4/y_c^2)|V_{cs}V_{us}|^2\sim5\times10^{-13}$. However, higher dimension terms can replace a y_s^2 factor with $(\Lambda/m_D)^2$ [18]. Moreover, long distance contributions are expected to dominate. In particular, peculiar phase space effects [19, 20] have been identified which are expected to enhance Δm_D to within an order of magnitude of its measured value.)

It is clear then that contributions from new physics at $\Lambda_{\rm NP}\sim1$ TeV should be suppressed by factors that are comparable to or smaller than the SM ones. Why does that happen? This is the new physics flavour puzzle.

The fact that the flavour structure of new physics at the TeV scale must be non-generic means that flavour measurements are a good probe of the new physics. Perhaps the best-studied example is that of supersymmetry. Here, the spectrum of the superpartners and the structure of their couplings to the SM fermions will allow us to probe the mechanism of dynamical supersymmetry breaking.

6 Lessons for supersymmetry from D^0–$\overline{D}^0$ mixing

Interesting experimental results concerning D^0–$\overline{D}^0$ mixing have recently been achieved by the BELLE and BaBar experiments. For the first time, there is evidence for width splitting [21,22] and mass splitting

[4] A detailed derivation can be found in Appendix B of Ref. [17].

(of order one per cent) between the two neutral D-mesons. Allowing for indirect CP violation, the world averages of the mixing parameters are [10]

$$x = (1.00 \pm 0.25) \times 10^{-2},$$
$$y = (0.77 \pm 0.18) \times 10^{-2}. \tag{56}$$

It is important to note, however, that there is no evidence for CP violation in this mixing [10]:

$$1 - |q/p| = +0.06 \pm 0.14,$$
$$\phi_D = -0.04 \pm 0.09. \tag{57}$$

We use this recent experimental information to draw important lessons on supersymmetry. This demonstrates how flavour physics—at the GeV scale—provides a significant probe of supersymmetry—at the TeV scale.

6.1 Neutral meson mixing with supersymmetry

We consider the contributions from the box diagrams involving the squark doublets of the first two generations, $\tilde{Q}_{L1,2}$, to the D^0–$\overline{D}^0$ and K^0–$\overline{K}^0$ mixing amplitudes. The contributions that are relevant to the neutral D system are proportional to $K^u_{2i}K^{u*}_{1i}K^u_{2j}K^{u*}_{1j}$, where K^u is the mixing matrix of the gluino couplings to a left-handed up quark and their supersymmetric squark partners. (In the language of the mass insertion approximation, we calculate here the contribution that is $\propto [(\delta^u_{LL})_{12}]^2$.) The contributions that are relevant to the neutral K system are proportional to $K^{d*}_{2i}K^d_{1i}K^{d*}_{2j}K^d_{1j}$, where K^d is the mixing matrix of the gluino couplings to a left-handed down quark and their supersymmetric squark partners ($\propto [(\delta^d_{LL})_{12}]^2$ in the mass insertion approximation). We work in the mass basis for both quarks and squarks. A detailed derivation [23] is given in Appendix C. It gives

$$M^D_{12} = \frac{\alpha_s^2 m_D f_D^2 B_D \eta_{\rm QCD}}{108 m^2_{\tilde{u}}}[11\tilde{f}_6(x_u) + 4x_u f_6(x_u)]\frac{(\Delta m^2_{\tilde{u}})^2}{m^4_{\tilde{u}}}(K^u_{21}K^{u*}_{11})^2, \tag{58}$$

$$M^K_{12} = \frac{\alpha_s^2 m_K f_K^2 B_K \eta_{\rm QCD}}{108 m^2_{\tilde{d}}}[11\tilde{f}_6(x_d) + 4x_d f_6(x_d)]\frac{(\Delta \tilde{m}^2_{\tilde{d}})^2}{\tilde{m}^4_{\tilde{d}}}(K^{d*}_{21}K^d_{11})^2. \tag{59}$$

Here $m_{\tilde{u},\tilde{d}}$ is the average mass of the corresponding two squark generations, $\Delta m^2_{\tilde{u},\tilde{d}}$ is the mass-squared difference, and $x_{u,d} = m^2_{\tilde{g}}/m^2_{\tilde{u},\tilde{d}}$.

One can immediately identify three generic ways in which supersymmetric contributions to neutral meson mixing can be suppressed:

1. Heaviness: $m_{\tilde{q}} \gg 1$ TeV.
2. Degeneracy: $\Delta m^2_{\tilde{q}} \ll m^2_{\tilde{q}}$.
3. Alignment: $K^{d,u}_{21} \ll 1$.

When heaviness is the only suppression mechanism, as in split supersymmetry [24], the squarks are very heavy and supersymmetry no longer solves the fine tuning problem[5]. If we want to maintain supersymmetry as a solution to the fine tuning problem, either degeneracy, or alignment, or a combination of both is needed. This means that the flavour structure of supersymmetry is not generic, as argued in the previous section.

The 2×2 mass-squared matrices for the relevant squarks have the following form:

$$\tilde{M}^2_{U_L} = \tilde{m}^2_{Q_L} + \left(\frac{1}{2} - \frac{2}{3}s^2_W\right) m^2_Z \cos 2\beta + M_u M_u^\dagger,$$

[5]When the first two squark generations are mildly heavy and the third generation is light, as in effective supersymmetry [25], the fine tuning problem is still solved, but additional suppression mechanisms are needed.

$$\tilde{M}^2_{D_L} = \tilde{m}^2_{Q_L} - \left(\frac{1}{2} - \frac{1}{3}s^2_W\right) m^2_Z \cos 2\beta + M_d M_d^\dagger. \tag{60}$$

We note the following features of the various terms:

- $\tilde{m}^2_{Q_L}$ is a 2×2 Hermitian matrix of soft supersymmetry breaking terms. It does not break $SU(2)_{\rm L}$ and consequently it is common to $\tilde{M}^2_{U_L}$ and $\tilde{M}^2_{D_L}$. On the other hand, it breaks in general the $SU(2)_Q$ flavour symmetry.
- The terms proportional to m^2_Z are the D terms. They break supersymmetry (since they involve $D_{T_3} \neq 0$ and $D_Y \neq 0$) and $SU(2)_{\rm L}$ but conserve $SU(2)_Q$.
- The terms proportional to M^2_q come from the F_{U_R} and F_{D_R} terms. They break the gauge $SU(2)_{\rm L}$ and the global $SU(2)_Q$ but, since $F_{U_R} = F_{D_R} = 0$, conserve supersymmetry.

Given that we are interested in squark masses close to the TeV scale (and the experimental lower bounds are of order 300 GeV), the scale of the eigenvalues of $\tilde{m}^2_{Q_L}$ is much higher than m^2_Z which, in turn, is much higher than m^2_c, the largest eigenvalue in $M_q M_q^\dagger$ (in the two-generation framework). We can draw the following conclusions:

1. $m^2_{\tilde{u}} = m^2_{\tilde{d}} \equiv m^2_{\tilde{q}}$ up to effects of order m^2_Z, namely to an accuracy of $\mathcal{O}(10^{-2})$.
2. $\Delta m^2_{\tilde{u}} = \Delta m^2_{\tilde{d}} \equiv \Delta m^2_{\tilde{q}}$ up to effects of order m^2_c, namely to an accuracy of $\mathcal{O}(10^{-5})$.
3. Since $K_u \simeq V_{uL}\tilde{V}_L^\dagger$ and $K_d \simeq V_{dL}\tilde{V}_L^\dagger$ [the matrices V_{qL} are defined in Eq. (21), while $\tilde{V}_L$ diagonalizes $\tilde{m}^2_{Q_L}$], the mixing matrices K^u and K^d are different from each other, but the following relation to the CKM matrix holds to an accuracy of $\mathcal{O}(10^{-5})$:

$$K^u K^{d\dagger} = V. \tag{61}$$

6.2 Non-degenerate squarks at the LHC?

Equations (58) and (59) can be translated into our generic language:

$$\Lambda_{\rm NP} = m_{\tilde{q}}, \tag{62}$$
$$z_{cu} = z_{12}\sin^2\theta_u,$$
$$z_{sd} = z_{12}\sin^2\theta_d,$$
$$z_{12} = \frac{11\tilde{f}_6(x) + 4x f_6(x)}{18}\alpha_s^2\left(\frac{\Delta\tilde{m}^2_{\tilde{q}}}{m^2_{\tilde{q}}}\right)^2, \tag{63}$$

with Eq. (61) giving

$$\sin\theta_u - \sin\theta_d \approx \sin\theta_c = 0.23. \tag{64}$$

We now ask the following question: Is it possible that the first two-generation squarks, $\tilde{Q}_{L1,2}$, are accessible to the LHC ($m_{\tilde{q}} \lesssim 1$ TeV), and are not degenerate ($\Delta m^2_{\tilde{q}}/m^2_{\tilde{q}} = \mathcal{O}(1)$)?

To answer this question, we use Eqs. (51) and (52). For $\Lambda_{\rm NP} \lesssim 1$ TeV, we have $z_{cu} \lesssim 5\times 10^{-7}$ and, for a phase that is $\not\ll 0.1$, $z_{sd} \lesssim 6\times 10^{-8}$. On the other hand, for non-degenerate squarks, and, for example, $11\tilde{f}_6(1) + 4f_6(1) = 1/6$, we have $z_{12} = 8\times 10^{-5}$. Then we need, simultaneously, $\sin\theta_u \lesssim 0.08$ and $\sin\theta_d \lesssim 0.03$, but this is inconsistent with Eq. (64).

There are three ways out of this situation:

1. The first two generation squarks are quasi-degenerate. The minimal level of degeneracy is $(\tilde{m}_2 - \tilde{m}_1)/(\tilde{m}_2 + \tilde{m}_1) \lesssim 0.1$. It could be the result of RGE [26]. However, for maximal phases, the bound is even stronger, of order 0.04 [27], which is difficult to achieve with just RGE effects.

2. The first two generation squarks are heavy. Putting $\sin\theta_u = 0.23$ and $\sin\theta_d \approx 0$, as in models of alignment [28, 29], Eq. (50) leads to
$$m_{\tilde{q}} \gtrsim 3 \text{ TeV} . \tag{65}$$
3. The ratio $x = \tilde{m}_g^2/\tilde{m}_q^2$ is in a fine-tuned region of parameter space where there are accidental cancellations in $11\tilde{f}_6(x) + 4x f_6(x)$. For example, for $x = 2.33$, this combination is ~ 0.003 and the bound (65) is relaxed by a factor of 7.

Barring accidental cancellations, the *model-independent* conclusion is that, if the first two generations of squark doublets are within the reach of the LHC, they must be quasi-degenerate [30, 31]. Analogous conclusions can be drawn for many TeV-scale new physics scenarios: a strong level of degeneracy is required (for definitions and detailed analysis, see Ref. [27]).

Exercise 7: *Does $K_{31}^d \sim |V_{ub}|$ suffice to satisfy the Δm_B constraint with neither degeneracy nor heaviness? (Use the two-generation approximation and ignore the second generation.)*

Is there a natural way to make the squarks degenerate? Examining Eqs. (60) we learn that degeneracy requires $\tilde{m}^2_{Q_L} \simeq \tilde{m}^2_{\tilde{q}}\mathbf{1}$. We have mentioned already that flavour universality is a generic feature of gauge interactions. Thus the requirement of degeneracy is perhaps a hint that supersymmetry breaking is *gauge mediated* to the MSSM fields.

7 Flavour at the LHC

The LHC will study the physics of electroweak symmetry breaking. There are high hopes that it will discover not only the Higgs, but also shed light on the fine-tuning problem that is related to the Higgs mass. Here, we focus on the issue of how, through the study of new physics, the LHC can shed light on the new physics flavour puzzle.

7.1 Minimal flavour violation (MFV)

If supersymmetry breaking is gauge mediated, the squark mass matrices of Eq. (60), and those for the SU(2)-singlet squarks, have the following form at the scale of mediation m_M:

$$\begin{aligned}
\tilde{M}^2_{U_L}(m_M) &= \left(m^2_{\tilde{Q}_L} + D_{U_L}\right)\mathbf{1} + M_u M_u^\dagger, \\
\tilde{M}^2_{D_L}(m_M) &= \left(m^2_{\tilde{Q}_L} + D_{D_L}\right)\mathbf{1} + M_d M_d^\dagger, \\
\tilde{M}^2_{U_R}(m_M) &= \left(m^2_{\tilde{U}_R} + D_{U_R}\right)\mathbf{1} + M_u^\dagger M_u, \\
\tilde{M}^2_{D_R}(m_M) &= \left(m^2_{\tilde{D}_R} + D_{D_R}\right)\mathbf{1} + M_d^\dagger M_d,
\end{aligned} \tag{66}$$

where $D_{q_A} = (T_3)_{q_A} - (Q_{\rm EM})_{q_A} s_W^2 m_Z^2 \cos 2\beta$ are the D-term contributions. Here, the only source of the $SU(3)^3_q$ breaking are the SM Yukawa matrices.

This statement holds also when the renormalization group evolution is applied to find the form of these matrices at the weak scale. Taking the scale of the soft breaking terms $m_{\tilde{q}_A}$ to be somewhat higher than the electroweak breaking scale m_Z allows us to neglect the D_{q_A} and M_q terms in (66). Then we obtain

$$\begin{aligned}
\tilde{M}^2_{Q_L}(m_Z) &\sim m^2_{\tilde{Q}_L}\left(r_3\mathbf{1} + c_u Y_u Y_u^\dagger + c_d Y_d Y_d^\dagger\right), \\
\tilde{M}^2_{U_R}(m_Z) &\sim m^2_{\tilde{U}_R}\left(r_3\mathbf{1} + c_{uR} Y_u^\dagger Y_u\right), \\
\tilde{M}^2_{D_R}(m_Z) &\sim m^2_{\tilde{D}_R}\left(r_3\mathbf{1} + c_{dR} Y_d^\dagger Y_d\right).
\end{aligned} \tag{67}$$

Here r_3 represent the universal RGE contribution that is proportional to the gluino mass ($r_3 = \mathcal{O}(6) \times (M_3(m_M)/m_{\tilde{q}}(m_M))$) and the c-coefficients depend logarithmically on m_M/m_Z and can be of $\mathcal{O}(1)$ when m_M is not far below the GUT scale.

Models of gauge mediated supersymmetry breaking (GMSB) provide a concrete example of a large class of models that obey a simple principle called *minimal flavour violation* (MFV) [32]. This principle guarantees that low-energy flavour-changing processes deviate only very little from the SM predictions. The basic idea can be described as follows. The gauge interactions of the SM are universal in flavour space. The only breaking of this flavour universality comes from the three Yukawa matrices, Y_U, Y_D, and Y_E. If this remains true in the presence of the new physics, namely Y_U, Y_D, and Y_E are the only flavour non-universal parameters, then the model belongs to the MFV class.

Let us now formulate this principle in a more formal way, using the language of spurions that we presented in Section 3.2. The Standard Model with vanishing Yukawa couplings has a large global symmetry of Eqs. (11) and (12). In this section we concentrate only on the quarks. The non-Abelian part of the flavour symmetry for the quarks is $SU(3)_q^3$ of Eq. (12) with the three generations of quark fields transforming as follows:

$$Q_L(3,1,1), \;\; U_R(1,3,1), \;\; D_R(1,1,3). \tag{68}$$

The Yukawa interactions,

$$\mathcal{L}_Y = \overline{Q_L} Y_D D_R H + \overline{Q_L} Y_U U_R H_c, \tag{69}$$

($H_c = i\tau_2 H^*$) break this symmetry. The Yukawa couplings can thus be thought of as spurions with the following transformation properties under $SU(3)_q^3$ [see Eq. (15)]:

$$Y_U \sim (3, \bar{3}, 1), \qquad Y_D \sim (3, 1, \bar{3}). \tag{70}$$

When we say 'spurions', we mean that we pretend that the Yukawa matrices are fields which transform under the flavour symmetry, and then require that all the Lagrangian terms, constructed from the SM fields, Y_D and Y_U, must be (formally) invariant under the flavour group $SU(3)_q^3$. Of course, in reality, $\mathcal{L}_Y$ breaks $SU(3)_q^3$ precisely because $Y_{D,U}$ are *not* fields and do not transform under the symmetry.

The idea of minimal flavour violation is relevant to extensions of the SM, and can be applied in two ways:

1. If we consider the SM as a low-energy effective theory, then all higher-dimension operators, constructed from SM fields and Y spurions, are formally invariant under G_{global}.
2. If we consider a full high-energy theory that extends the SM, then all operators, constructed from SM and the new fields, and from Y spurions, are formally invariant under G_{global}.

Exercise 8: *Use the spurion formalism to argue that, in MFV models, the $K_L \to \pi^0 \nu\bar{\nu}$ decay amplitude is proportional to $y_t^2 V_{td} V_{ts}^*$.*

Examples of MFV models include models of supersymmetry with gauge- or anomaly-mediation of its breaking. If the LHC discovers new particles that couple to the SM fermions, then it will be able to test solutions to the new physics flavour puzzle such as MFV [33]. Much of its power to test such frameworks is based on identifying top and bottom quarks.

To understand this statement, we note that the spurions Y_U and Y_D can always be written in terms of the two diagonal Yukawa matrices λ_u and λ_d and the CKM matrix V, see Eqs. (17) and (18). Thus, the only source of quark flavour-changing transitions in MFV models is the CKM matrix. Next, note that to an accuracy that is better than $\mathcal{O}(0.05)$, we can write the CKM matrix as follows:

$$V = \begin{pmatrix} 1 & 0.23 & 0 \\ -0.23 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} . \tag{71}$$

Exercise 9: *The approximation (71) should be intuitively obvious to top-physicists, but definitely counter-intuitive to bottom-physicists. (Some of them have dedicated a large part of their careers to experimental or theoretical efforts to determine V_{cb} and V_{ub}.) What does the approximation imply for the bottom quark? When we take into account that it is only good to $\mathcal{O}(0.05)$, what would the implications be?*

We learn that the third generation of quarks is decoupled, to a good approximation, from the first two. This, in turn, means that any new particle that couples to the SM quarks (think, for example, of heavy quarks in vector-like representations of $G_{\rm SM}$), decays into either a third-generation quark, or into a non-third-generation quark, but not to both. For example, in Ref. [33], MFV models with additional charge $-1/3$, $SU(2)_{\rm L}$-singlet quarks, B', were considered. A concrete test of MFV was proposed, based on the fact that the largest mixing effect involving the third generation is of order $|V_{cb}|^2 \sim 0.002$: Is the following prediction, concerning events of B' pair production, fulfilled?

$$\frac{\Gamma(B'\overline{B'} \to Xq_{1,2}q_3)}{\Gamma(B'\overline{B'} \to Xq_{1,2}q_{1,2}) + \Gamma(B'\overline{B'} \to Xq_3q_3)} \lesssim 10^{-3}. \tag{72}$$

If not, then MFV is excluded.

7.2 Supersymmetric flavour at the LHC

One can think of analogous tests in the supersymmetric framework [34–39]. Here, there is also a generic prediction that, in each of the three sectors (Q_L, U_R, D_R), squarks of the first two generations are quasi-degenerate, and do not decay into third-generation quarks. Squarks of the third generation can be separated in mass (though, for small $\tan\beta$, the degeneracy in the $\tilde{D}_R$ sector is threefold), and decay only to third-generation quarks.

It is not necessary, however, that the mediation of supersymmetry breaking be MFV. Examples of natural and viable solutions to the supersymmetric flavour problem that are not MFV include the following:

1. The leading contribution to the soft supersymmetry breaking terms is gauge mediated, and therefore MFV, but there are subleading contributions that are gravity mediated and provide new sources of flavour and CP violation [34, 39]. The gravity mediated contributions could either have some structure (dictated, for example, by a Froggatt–Nielsen symmetry [34] or by localization in extra dimensions [40]) or be anarchical [41].
2. The first two sfermion generations are heavy, and their mixing with the third generation is suppressed (for a recent analysis, see Ref. [42]). These features can come, for example, from conformal dynamics [43].

Such frameworks have different predictions concerning the mass splitting between sfermion generations and the flavour decomposition of the sfermion mass eigenstates. Note that measurements of flavour-changing neutral current processes are only sensitive to the products of the form

$$\delta_{ij} = \frac{\Delta\tilde{m}^2_{ij}}{\tilde{m}^2}\, K_{ij}K^*_{jj}, \tag{73}$$

where $\Delta\tilde{m}^2_{ij}$ is the mass-squared splitting between the sfermion generations i and j, $\tilde{m}^2$ is their average mass-squared, and K is the mixing matrix of gaugino couplings to these sfermions. On the other hand, the LHC experiments—ATLAS and CMS—can, at least in principle, measure the mass splitting and the mixing separately [37].

The present situation is depicted schematically in Fig. 5(a). Flavour factories have provided only upper bounds on deviations of FCNC processes, such as $\mu \to e\gamma$ or D^0–$\overline{D}^0$ mixing, from the Standard

Model predictions. In the supersymmetric framework, such bounds translate into an upper bound on a δ_{ij} parameter of Eq. (73), corresponding to the blue region in the figure. The supersymmetric flavour puzzle can be stated as the question of why the region in the upper right corner—where the flavour parameters are of order one—is excluded. MFV often puts us in the lower left corner of the plot, far from the experimental constraints (this is particularly true for δ_{12} parameters).

The optimal future situation is depicted schematically in Fig. 5(b). Imagine that a flavour factory does provide evidence for new physics, such as observation of $\Gamma(\mu \to e\gamma) \neq 0$ or CP violation in D^0–$\overline{D}^0$ mixing. This will constrain the corresponding δ parameter, which is shown as the blue region in the figure. If ATLAS/CMS measure the corresponding sfermion mass splitting and/or mixing, we shall get a small allowed region in this flavour plane.

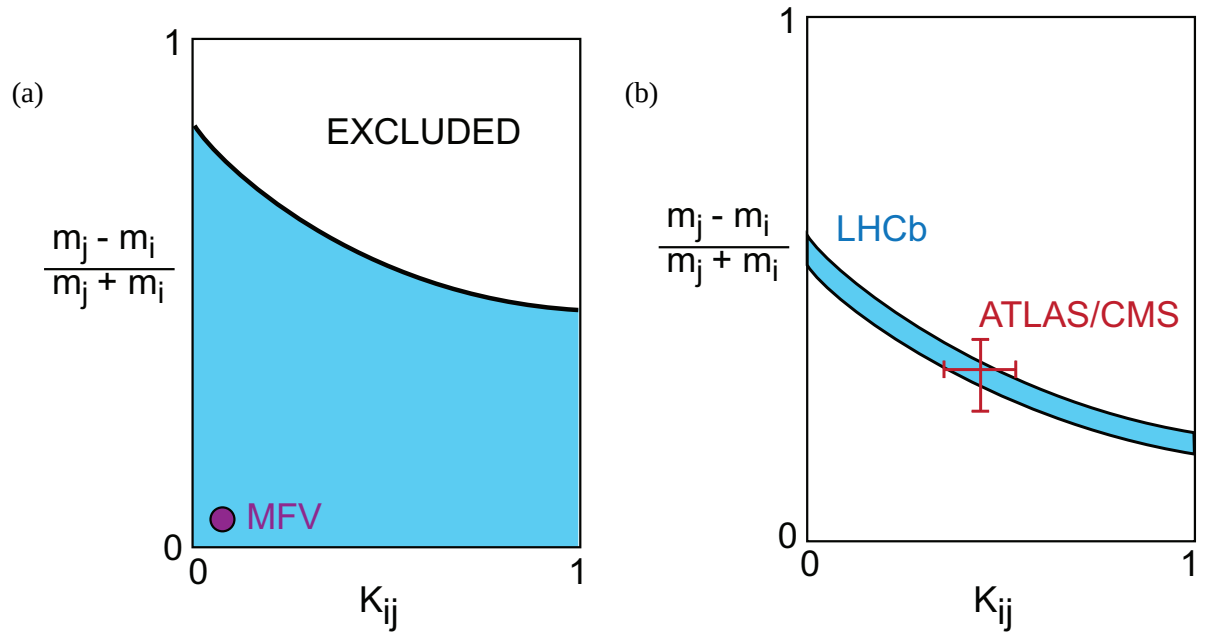


Fig. 5: Schematic description of the constraints in the plane of sfermion mass-squared splitting, $\Delta\tilde{m}^2_{ij}/\tilde{m}^2$, and mixing, $K_{ij}K^*_{jj}$: (a) Upper bounds from not observing any deviation from the SM predictions in present experiments; (b) Hypothetical future situation, where deviations have been observed in flavour factories (such as LHCb, a super-B factory, a $\mu \to e\gamma$ measurement, etc.) and the mass splitting and flavour decomposition have been measured by ATLAS/CMS.

If we have at our disposal three such consistent measurements (rate of FCNC process, spectrum and splitting), then we shall understand the mechanism by which supersymmetry has its flavour violation suppressed. This will provide strong hints about the mechanism of supersymmetry breaking mediation.

If the sfermions are quasi-degenerate, then the mixing is determined by the small corrections to the unit mass-squared matrix. As mentioned above, the structure of such corrections may be dictated by the same symmetry or dynamics that gives the structure of the Yukawa couplings. If that is the case, then the measurement of the flavour decomposition might shed light on the Standard Model flavour puzzle.

We conclude that measurements at the LHC related to new particles that couple to the SM fermions are likely to teach us much more about flavour physics.

8 Neutrino anarchy versus quark hierarchy

A detailed presentation of the physics and the formalism of neutrino flavour transitions is given in Appendix D for both vacuum oscillations (D.1) and the matter transitions (D.2). It follows Ref. [44].

Exercise 10: *For atmospheric ν_μ's with $E \sim 1$ GeV, the flux coming from above has $P_{\mu\mu}(L \sim 10$ km$) \approx 1$, while the flux from below has $P_{\mu\mu}(L \sim 10^4$ km$) \approx 0.5$. Assuming that for the flux coming from below the oscillations are averaged out, estimate Δm^2 and $\sin^2 2\theta$.*

Exercise 11: *For solar ν_e's, the transition between matter ($\beta_{MSW} > 1$) and vacuum ($\beta_{MSW} < \cos 2\theta$) flavour transitions occurs around $E \sim 2$ MeV. The transition probability is measured to be roughly $P_{ee} \sim 0.30$ for $\beta_{MSW} > 1$. Estimate Δm^2 and θ and predict P_{ee} for $\beta_{MSW} \ll 1$.*

The derived ranges for the three mixing angles and two mass-squared differences at 1σ are [45]

$$\begin{aligned} \Delta m^2_{21} &= (7.9 \pm 0.3) \times 10^{-5}\ \text{eV}^2, \quad |\Delta m^2_{32}| = (2.6 \pm 0.2) \times 10^{-3}\ \text{eV}^2, \\ \sin^2\theta_{12} &= 0.31 \pm 0.02, \quad \sin^2\theta_{23} = 0.47 \pm 0.07, \quad \sin^2\theta_{13} = 0^{+0.008}_{-0.0}. \end{aligned} \tag{74}$$

The 3σ range for the matrix elements of U are the following [45]:

$$|U| = \begin{pmatrix} 0.79 \to 0.86 & 0.50 \to 0.61 & 0.00 \to 0.20 \\ 0.25 \to 0.53 & 0.47 \to 0.73 & 0.56 \to 0.79 \\ 0.21 \to 0.51 & 0.42 \to 0.69 & 0.61 \to 0.83 \end{pmatrix} . \tag{75}$$

8.1 New physics

The simplest and most straightforward lesson of the evidence for neutrino masses is also the most striking one: there is new physics beyond the Standard Model. This is the first experimental result that is inconsistent with the SM.

Most likely, the new physics is related to the existence of $G_{\rm SM}$-singlet fermions at some high energy scale that induce, at low energies, the effective terms of Eq. (44) through the seesaw mechanism. The existence of heavy singlet fermions is predicted by many extensions of the SM, especially by GUTs [beyond $SU(5)$] and left–right-symmetric theories. The seesaw mechanism could also be driven by an $SU(2)_{\rm L}$-triplet fermion.

There are other possibilities. In particular, neutrino masses can be generated without introducing any new fermions beyond those of the SM. Instead, the existence of a scalar $\Delta_L(1,3)_{+1}$, that is, an $SU(2)_{\rm L}$-triplet, is required. The smallness of the neutrino masses is related here to the smallness of the vacuum expectation value $\langle\Delta^0_L\rangle$ (required also by the success of the $\rho = 1$ relation) and does not have a generic natural explanation.

In left–right-symmetric models, however, where the breaking of $SU(2)_{\rm R} \times U(1)_{\text{B-L}} \to U(1)_{\rm Y}$ is induced by the VEV of an $SU(2)_{\rm R}$-triplet, Δ_R, there must exist also an $SU(2)_{\rm L}$-triplet scalar. Furthermore, the Higgs potential leads to an order of magnitude relation between the various VEVs, $\langle\Delta^0_L\rangle\langle\Delta^0_R\rangle \sim v^2$, and the smallness of $\langle\Delta^0_L\rangle$ is correlated with the high scale of $SU(2)_{\rm R}$ breaking. This situation can be thought of as a seesaw of VEVs. In this model there are, however, also SM-singlet fermions. The light neutrino masses arise from both the seesaw mechanism ('type I') and the triplet VEV ('type II').

Neutrino masses could also be of the Dirac type. Here, again, singlet fermions are introduced, but lepton number is imposed by hand. This possibility is disfavoured by theorists since it is likely that global symmetries are violated by gravitational effects. Furthermore, the lightness of the neutrinos (compared to charged fermions) is unexplained.

Another possibility is that neutrino masses are generated by mixing with singlet fermions but the mass scale of these fermions is not high. Here again the lightness of neutrino masses remains a puzzle. The best known example of such a scenario is the framework of supersymmetry without R parity.

Let us emphasize that the seesaw mechanism or, more generally, the extension of the SM with non-renormalizable terms, is the simplest explanation of neutrino masses. Models in which neutrino masses are generated by new physics at low energy imply a much more dramatic departure from the SM. Furthermore, the existence of seesaw masses is an unavoidable prediction of various extensions of the

SM. In contrast, many (but not all) of the low-energy mechanisms are introduced for the specific purpose of generating neutrino masses.

8.2 The scale of new physics

Equation (44) gives a light neutrino mass matrix:

$$(M_\nu)_{ij} = Z^\nu_{ij}\frac{v^2}{\Lambda_{\rm NP}}. \tag{76}$$

It is straightforward to use the measured neutrino masses of Eq. (74) in combination with Eq. (76) to estimate the scale of new physics that is relevant to their generation. In particular, if there is no quasi-degeneracy in the neutrino masses, the heaviest of the active neutrino masses can be estimated:

$$m_h = m_3 \sim \sqrt{\Delta m^2_{32}} \approx 0.05 \text{ eV}. \tag{77}$$

(In the case of inverted hierarchy, the implied scale is $m_h = m_2 \sim \sqrt{\Delta m^2_{32}} \approx 0.05$ eV.) It follows that the scale in the non-renormalizable terms (44) is given by

$$\Lambda_{\rm NP} \sim v^2/m_h \approx 10^{15} \text{ GeV}. \tag{78}$$

We should clarify two points regarding Eq. (78):

1. There could be some level of degeneracy between the neutrino masses. In such a case, Eq. (77) is modified into a lower bound on m_3 and, consequently, Eq. (78) becomes an upper bound on $\Lambda_{\rm NP}$.
2. It could be that the Z_{ij} of Eq. (44) are much smaller than 1. In such a case, again, Eq. (78) becomes an upper bound on the scale of new physics.

On the other hand, in models of approximate flavour symmetries, there are relations between the structures of the charged lepton and neutrino mass matrices that give, quite generically, $Z_{33} \gtrsim m^2_\tau/v^2 \sim 10^{-4}$. We conclude that the likely range for $\Lambda_{\rm NP}$ is given by

$$10^{11} \text{ GeV} \lesssim \Lambda_{\rm NP} \lesssim 10^{15} \text{ GeV} . \tag{79}$$

The estimates (78) and (79) are very exciting. First, the upper bound on the scale of new physics is well below the Planck scale. This means that there is new physics in Nature which is intermediate between the two known scales, the Planck scale, $m_{\rm Pl} \sim 10^{19}$ GeV, and the electroweak breaking scale, $v \sim 10^2$ GeV.

Second, the scale $\Lambda_{\rm NP} \sim 10^{15}$ GeV is intriguingly close to the scale of gauge coupling unification.

Third, the range (79) for the scale of lepton number breaking is optimal for leptogenesis [46] (for a recent review, see Ref. [47]). If (i) leptogenesis is generated by the decays of the lightest singlet neutrino N_1, and (ii) the masses of the singlet neutrinos are hierarchical, $M_1/M_{2,3...} \ll 1$, and (iii) the temperature when leptogenesis occurs is high enough, $T_{\rm LG} > 10^{12}$ GeV, so that flavour effects are unimportant, then there is an upper bound on the CP asymmetry in N_1 decays [48]:

$$|\epsilon_{N_1}| \leq \frac{3}{16\pi}\frac{M_1(m_3 - m_2)}{v^2}. \tag{80}$$

Given that $Y_B^{\rm obs} \sim 9\times10^{-11}$, and that $Y_B \sim 10^{-3}\eta\epsilon_{N_1}$, where $\eta \lesssim 1$ is a washout factor, we must require $|\epsilon_{N_1}| \gtrsim 10^{-7}$. Moreover, we have $m_3 - m_2 \leq \sqrt{\Delta m^2_{32}} \sim 0.05$ eV and therefore obtain $M_1 \gtrsim 10^9$ GeV. Violating any of the three conditions will relax this bound, but typically not by more than about an order of magnitude.

8.3 The flavour puzzle

In the absence of neutrino masses, there are 13 flavour parameters in the SM:

$$\begin{aligned} y_t &\sim 1,\ y_c \sim 10^{-2},\ y_u \sim 10^{-5}, \\ y_b &\sim 10^{-2},\ y_s \sim 10^{-3},\ y_d \sim 10^{-4}, \\ y_\tau &\sim 10^{-2},\ y_\mu \sim 10^{-3},\ y_e \sim 10^{-6}, \\ |V_{us}| &\sim 0.2,\ |V_{cb}| \sim 0.04,\ |V_{ub}| \sim 0.004,\ \sin\delta_{\rm KM} \sim 1. \end{aligned} \tag{81}$$

These flavour parameters are hierarchical (their magnitudes span six orders of magnitude), and all but two or three (the top Yukawa, the CP violating phase, and perhaps the Cabibbo angle) are small. The unexplained smallness and hierarchy pose the SM *flavour puzzle*. Its solution may direct us to physics beyond the Standard Model.

Several mechanisms have been proposed in response to this puzzle. For example, approximate horizontal symmetries, broken by a small parameter, can lead to selection rules that explain the hierarchy of the Yukawa couplings.

In the extension of the SM with three active neutrinos that have Majorana masses, there are nine new flavour parameters in addition to those of Eq. (81). These are three neutrino masses, three lepton mixing angles, and three phases in the mixing matrix. Of the nine new parameters, four have been measured: two mass-squared differences and two mixing angles [see Eq. (74)]. This adds significantly to the input data on flavour physics and provides an opportunity to test and refine flavour models.

If neutrino masses arise from effective terms of the form of Eq. (44), then the overall scale of neutrino masses is related to the scale $\Lambda_{\rm NP}$ and, in most cases, does not tell us anything about flavour physics. More significant information for flavour models can be written in terms of three dimensionless parameters whose values can be read from Eq. (74), that is $\sin\theta_{12}$, $\sin\theta_{23}$ and

$$\Delta m^2_{21}/|\Delta m^2_{32}| = 0.030 \pm 0.003. \tag{82}$$

In addition, the upper bound on $\sin\theta_{13}$ often plays a significant role in flavour model building.

There are several features in the numerical estimates (74) and (82) that have drawn much attention and have driven numerous investigations:

(i) *Large mixing and strong hierarchy:* The mixing angle that is relevant to the 2–3 sector is large, $\sin\theta_{23} \sim 0.7$. On the other hand, if there is no quasi-degeneracy in the neutrino masses, the corresponding mass ratio is small, $m_2/m_3 \sim 0.17$. It is difficult to explain in a natural way a situation where there is an $\mathcal{O}(1)$ mixing but the corresponding masses are hierarchical.

(ii) *Two large and one small mixing angles:* The mixing angles relevant to the 2–3 sector ($\sin\theta_{23} \sim 0.7$) and 1–2 sector ($\sin\theta_{12} \sim 0.55$) are large, yet the 1–3 mixing angle is small ($\sin\theta_{13} \lesssim 0.20$). Such a situation is, again, difficult—though not impossible—to explain from approximate symmetries. An example of a symmetry that does predict such a pattern is that of L_e–L_μ–L_τ. This symmetry predicts, however, $\theta_{12} \simeq \pi/4$, which is experimentally excluded.

(iii) *Maximal mixing:* The value of θ_{23} is intriguingly close to maximal mixing ($\sin^2 2\theta_{23} = 1$). It is interesting to understand whether a symmetry could explain this special value.

(iv) *Tribimaximal mixing:* The mixing matrix (75) has a structure that is consistent with the following unitary matrix [49]:

$$U = \begin{pmatrix} \sqrt{\frac{2}{3}} & \sqrt{\frac{1}{3}} & 0 \\ -\sqrt{\frac{1}{6}} & \sqrt{\frac{1}{3}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{3}} & \sqrt{\frac{1}{2}} \end{pmatrix}. \tag{83}$$

It is interesting to understand whether a symmetry could explain this special structure.

All four features enumerated above are difficult to explain in a large class of flavour models that do very well in explaining the flavour features of the quark sector. In particular, models with Abelian horizontal symmetries (Froggatt–Nielsen type [50]) predict that, in general, $|V_{ub}| \sim |V_{us}V_{cb}|$, $|V_{ij}| \gtrsim m_i/m_j$ ($i < j$) and $V \sim \mathbf{1}$ [29, 51]. All of these are successful predictions. At the same time, however, these models predict [52] that for the neutrinos, in general, $|U_{ij}|^2 \sim m_i/m_j$ and $|U_{e3}| \sim |U_{e2}U_{\mu 3}|$, in contradiction to, respectively, points (i) and (ii) above (and there is no way to make θ_{23} parametrically close to $\pi/4$). On the other hand, there exist very specific models where these features are related to a symmetry.

It is possible, however, that the above interpretation of the results is wrong. Indeed, the data can be interpreted in a very different way:

(v) *No small parameters:* The two measured mixing angles are larger than any of the quark mixing angles. Indeed, they are both of order one. The measured mass ratio, $m_2/m_3 \gtrsim 0.16$ is larger than any of the quark and charged lepton mass ratios, and could be interpreted as an $\mathcal{O}(1)$ parameter (namely, it is accidentally small, without any parametric suppression). If this is the correct way of reading the data, the measured neutrino parameters may actually reflect the absence of any hierarchical structure in the neutrino mass matrices [53]. The possibility that there is no structure—neither hierarchy, nor degeneracy—in the neutrino sector has been called 'neutrino mass anarchy'. An important test of this idea will be provided by the measurement of $|U_{e3}|$. If indeed the entries in M_ν have random values of the same order, all three mixing angles are expected to be of order one. If experiments measure $|U_{e3}| \sim 0.1$, that is, close to the present bound, it can be argued that its smallness is accidental. The stronger the upper bound on this angle becomes, the more difficult it will be to maintain this view.

Neutrino mass anarchy can be accommodated within models of Abelian flavour symmetries, if the three lepton doublets carry the same charge. Indeed, consider a supersymmetric model with a $U(1)_H$ symmetry that is broken by a single small spurion ϵ_H of charge -1. Let us assume that the three fermion generations contained in the 10-representation of $SU(5)$ carry charges $(2, 1, 0)$, while the three $\bar{5}$-representations carry charges $(0, 0, 0)$. (The Higgs fields carry no H charges.) Such a model predicts ϵ_H^2 hierarchy in the up sector, ϵ_H hierarchy in the down and charged lepton sectors, and anarchy in the neutrino sector.

Exercise 12: *The selection rule for this model is that a term in the superpotential that carries H charge $n \geq 0$ is suppressed by ϵ_H^n. Find the parametric suppression of the various entries in M_u, M_d, M_ℓ, and M_ν. Find the parametric suppression of the mixing angles.*

It would be nice if the features of quark mass hierarchy and neutrino mass anarchy can be traced back to some fundamental principle or to a stringy origin (see, for example, Ref. [54]).

9 Conclusions

(i) Measurements of CP violating B-meson decays have established that the Kobayashi–Maskawa mechanism is the dominant source of the observed CP violation.

(ii) Measurements of flavour-changing B-meson decays have established that the Cabibbo–Kobayashi–Maskawa mechanism is a major player in flavour violation.

(iii) The consistency of all these measurements with the CKM predictions sharpens the new physics flavour puzzle: If there is new physics at, or below, the TeV scale, then its flavour structure must be highly non-generic.

(iv) Measurements of D^0–$\overline{D}^0$ mixing imply that alignment by itself cannot solve the supersymmetric flavour problem. The first two squark generations must be quasi-degenerate.

(v) Measurements of neutrino flavour parameters have not only not clarified the Standard Model flavour puzzle, but actually deepened it. Whether they imply an anarchical structure, or a tribimaximal mixing, it seems that the neutrino flavour structure is very different from that of quarks.

(vi) If the LHC experiments, ATLAS and CMS, discover new particles that couple to the Standard Model fermions, then, in principle, they will be able to measure new flavour parameters. Consequently, the new physics flavour puzzle is likely to be understood.

(vii) If the flavour structure of such new particles is affected by the same physics that sets the flavour structure of the Yukawa couplings, then the LHC experiments (and future flavour factories) may be able to shed light also on the Standard Model flavour puzzle.

The huge progress in flavour physics in recent years has provided answers to many questions. At the same time, new questions arise. We look forward to the LHC era for more answers and more questions.

Acknowledgements

The research of Y. Nir is supported by the Israel Science Foundation; the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel; the German–Israeli Foundation for Scientific Research and Development (GIF); and the Minerva Foundation.

Appendices

A The CKM matrix

The CKM matrix V is a 3×3 unitary matrix. Its form, however, is not unique:

(i) There is freedom in defining V in that we can permute between the various generations. This freedom is fixed by ordering the up quarks and the down quarks by their masses, i.e., $(u_1, u_2, u_3) \to (u, c, t)$ and $(d_1, d_2, d_3) \to (d, s, b)$. The elements of V are written as follows:

$$V = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} . \tag{A.1}$$

(ii) There is further freedom in the phase structure of V. This means that the number of physical parameters in V is smaller than the number of parameters in a general unitary 3×3 matrix which is nine (three real angles and six phases). Let us define P_q ($q = u, d$) to be diagonal unitary (phase) matrices. Then, if instead of using V_{qL} and V_{qR} for the rotation (21) to the mass basis we use $\tilde{V}_{qL}$ and $\tilde{V}_{qR}$, defined by $\tilde{V}_{qL} = P_q V_{qL}$ and $\tilde{V}_{qR} = P_q V_{qR}$, we still maintain a legitimate mass basis since $M_q^{\rm diag}$ remains unchanged by such transformations. However, V does change:

$$V \to P_u V P_d^* . \tag{A.2}$$

This freedom is fixed by demanding that V has the minimal number of phases. In the three-generation case V has a single phase. (There are five phase differences between the elements of P_u and P_d and, therefore, five of the six phases in the CKM matrix can be removed.) This is the Kobayashi–Maskawa phase $\delta_{\rm KM}$ which is the single source of CP violation in the quark sector of the Standard Model [1].

The fact that V is unitary and depends on only four independent physical parameters can be made manifest by choosing a specific parametrization. The standard choice is [55]

$$V = \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta} & c_{23}c_{13} \end{pmatrix} , \tag{A.3}$$

where $c_{ij} \equiv \cos\theta_{ij}$ and $s_{ij} \equiv \sin\theta_{ij}$. The θ_{ij}'s are the three real mixing parameters while δ is the Kobayashi–Maskawa phase. It is known experimentally that $s_{13} \ll s_{23} \ll s_{12} \ll 1$. It is convenient to

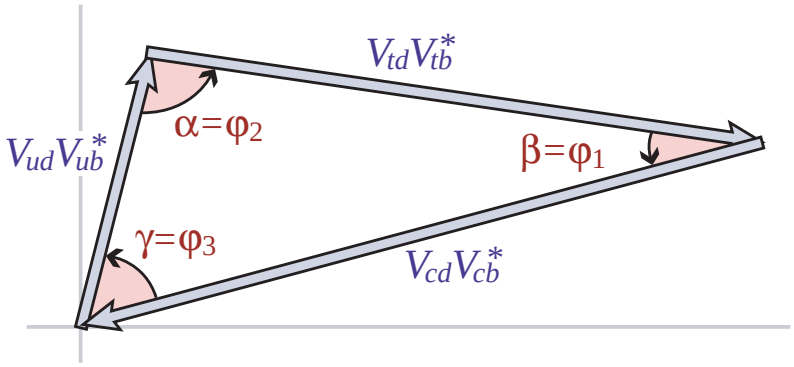


Fig. A.1: Graphical representation of the unitarity constraint $V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0$ as a triangle in the complex plane

choose an approximate expression where this hierarchy is manifest. This is the Wolfenstein parametrization, where the four mixing parameters are (λ, A, ρ, η) with $\lambda = |V_{us}| = 0.23$ playing the role of an expansion parameter and η representing the CP violating phase [56, 57]:

$$V = \begin{pmatrix} 1-\frac{1}{2}\lambda^2-\frac{1}{8}\lambda^4 & \lambda & A\lambda^3(\rho-i\eta) \\ -\lambda+\frac{1}{2}A^2\lambda^5[1-2(\rho+i\eta)] & 1-\frac{1}{2}\lambda^2-\frac{1}{8}\lambda^4(1+4A^2) & A\lambda^2 \\ A\lambda^3[1-(1-\frac{1}{2}\lambda^2)(\rho+i\eta)] & -A\lambda^2+\frac{1}{2}A\lambda^4[1-2(\rho+i\eta)] & 1-\frac{1}{2}A^2\lambda^4 \end{pmatrix} . \tag{A.4}$$

A very useful concept is that of the *unitarity triangles*. The unitarity of the CKM matrix leads to various relations among the matrix elements, e.g.,

$$V_{ud}V_{us}^* + V_{cd}V_{cs}^* + V_{td}V_{ts}^* = 0, \tag{A.5}$$
$$V_{us}V_{ub}^* + V_{cs}V_{cb}^* + V_{ts}V_{tb}^* = 0, \tag{A.6}$$
$$V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0. \tag{A.7}$$

Each of these three relations requires the sum of three complex quantities to vanish and so can be geometrically represented in the complex plane as a triangle. These are 'the unitarity triangles', though the term 'unitarity triangle' is usually reserved for the relation (A.7) only. The unitarity triangle related to Eq. (A.7) is depicted in Fig. A.1.

The rescaled unitarity triangle is derived from (A.7) by (a) choosing a phase convention such that $(V_{cd}V_{cb}^*)$ is real, and (b) dividing the lengths of all sides by $|V_{cd}V_{cb}^*|$. Step (a) aligns one side of the triangle with the real axis, and step (b) makes the length of this side 1. The form of the triangle is unchanged. Two vertices of the rescaled unitarity triangle are thus fixed at (0,0) and (1,0). The coordinates of the remaining vertex correspond to the Wolfenstein parameters (ρ, η). The area of the rescaled unitarity triangle is $|\eta|/2$.

Depicting the rescaled unitarity triangle in the (ρ, η) plane, the lengths of the two complex sides are

$$R_u \equiv \left|\frac{V_{ud}V_{ub}}{V_{cd}V_{cb}}\right| = \sqrt{\rho^2+\eta^2}, \quad R_t \equiv \left|\frac{V_{td}V_{tb}}{V_{cd}V_{cb}}\right| = \sqrt{(1-\rho)^2+\eta^2}. \tag{A.8}$$

The three angles of the unitarity triangle are defined as follows [58, 59]:

$$\alpha \equiv \arg\left[-\frac{V_{td}V_{tb}^*}{V_{ud}V_{ub}^*}\right], \quad \beta \equiv \arg\left[-\frac{V_{cd}V_{cb}^*}{V_{td}V_{tb}^*}\right], \quad \gamma \equiv \arg\left[-\frac{V_{ud}V_{ub}^*}{V_{cd}V_{cb}^*}\right] . \tag{A.9}$$

They are physical quantities and can be independently measured by CP asymmetries in B decays. It is also useful to define the two small angles of the unitarity triangles (A.5), (A.6):

$$\beta_s \equiv \arg\left[-\frac{V_{ts}V_{tb}^*}{V_{cs}V_{cb}^*}\right], \quad \beta_K \equiv \arg\left[-\frac{V_{cs}V_{cd}^*}{V_{us}V_{ud}^*}\right] . \tag{A.10}$$

The λ and A parameters are very well determined at present, see Eq. (37). The main effort in CKM measurements is thus aimed at improving our knowledge of ρ and η:

$$\rho = 0.14^{+0.03}_{-0.02}, \quad \eta = 0.35 \pm 0.02. \tag{A.11}$$

The present status of our knowledge is best seen in a plot of the various constraints and the final allowed region in the ρ–η plane. This is shown in Fig. 2.

B CP violation in neutral B decays to final CP eigenstates

We define decay amplitudes of B (which could be charged or neutral) and its CP conjugate $\overline{B}$ to a multiparticle final state f and its CP conjugate $\overline{f}$ as

$$A_f = \langle f|\mathcal{H}|B\rangle \quad , \quad \overline{A}_f = \langle f|\mathcal{H}|\overline{B}\rangle \quad , \quad A_{\overline{f}} = \langle \overline{f}|\mathcal{H}|B\rangle \quad , \quad \overline{A}_{\overline{f}} = \langle \overline{f}|\mathcal{H}|\overline{B}\rangle \ , \tag{B.1}$$

where $\mathcal{H}$ is the Hamiltonian governing weak interactions. The action of CP on these states introduces phases ξ_B and ξ_f according to

$$\begin{aligned} CP\,|B\rangle &= e^{+i\xi_B}\,|\overline{B}\rangle \quad , \quad CP\,|f\rangle = e^{+i\xi_f}\,|\overline{f}\rangle \ , \\ CP\,|\overline{B}\rangle &= e^{-i\xi_B}\,|B\rangle \quad , \quad CP\,|\overline{f}\rangle = e^{-i\xi_f}\,|f\rangle \ , \end{aligned} \tag{B.2}$$

so that $(CP\,)^2 = 1$. The phases ξ_B and ξ_f are arbitrary and unphysical because of the flavour symmetry of the strong interaction. If CP is conserved by the dynamics, $[CP\,,\mathcal{H}] = 0$, then A_f and $\overline{A}_{\overline{f}}$ have the same magnitude and an arbitrary unphysical relative phase

$$\overline{A}_{\overline{f}} = e^{i(\xi_f - \xi_B)}\,A_f \ . \tag{B.3}$$

A state that is initially a superposition of B^0 and $\overline{B}^0$, say

$$|\psi(0)\rangle = a(0)|B^0\rangle + b(0)|\overline{B}^0\rangle \ , \tag{B.4}$$

will evolve in time acquiring components that describe all possible decay final states $\{f_1, f_2, \ldots\}$, that is,

$$|\psi(t)\rangle = a(t)|B^0\rangle + b(t)|\overline{B}^0\rangle + c_1(t)|f_1\rangle + c_2(t)|f_2\rangle + \cdots \ . \tag{B.5}$$

If we are interested in computing only the values of $a(t)$ and $b(t)$ (and not the values of all $c_i(t)$), and if the times t in which we are interested are much larger than the typical strong interaction scale, then we can use a much simplified formalism [60]. The simplified time evolution is determined by a 2×2 effective Hamiltonian $\mathcal{H}$ that is not Hermitian, since otherwise the mesons would only oscillate and not decay. Any complex matrix, such as $\mathcal{H}$, can be written in terms of Hermitian matrices M and Γ as

$$\mathcal{H} = M - \frac{i}{2}\,\Gamma \ . \tag{B.6}$$

M and Γ are associated with $(B^0, \overline{B}^0) \leftrightarrow (B^0, \overline{B}^0)$ transitions via off-shell (dispersive) and on-shell (absorptive) intermediate states, respectively. Diagonal elements of M and Γ are associated with the flavour-conserving transitions $B^0 \to B^0$ and $\overline{B}^0 \to \overline{B}^0$ while off-diagonal elements are associated with flavour-changing transitions $B^0 \leftrightarrow \overline{B}^0$.

The eigenvectors of $\mathcal{H}$ have well-defined masses and decay widths. We introduce complex parameters $p_{L,H}$ and $q_{L,H}$ to specify the components of the strong interaction eigenstates, B^0 and $\overline{B}^0$, in the light (B_L) and heavy (B_H) mass eigenstates:

$$|B_{L,H}\rangle = p_{L,H}|B^0\rangle \pm q_{L,H}|\overline{B}^0\rangle \tag{B.7}$$

with the normalization $|p_{L,H}|^2 + |q_{L,H}|^2 = 1$. If either CP or CPT is a symmetry of $\mathcal{H}$ (independently of whether T is conserved or violated) then $M_{11} = M_{22}$ and $\Gamma_{11} = \Gamma_{22}$, and solving the eigenvalue problem for $\mathcal{H}$ yields $p_L = p_H \equiv p$ and $q_L = q_H \equiv q$ with

$$\left(\frac{q}{p}\right)^2 = \frac{M_{12}^* - (i/2)\Gamma_{12}^*}{M_{12} - (i/2)\Gamma_{12}} \,. \tag{B.8}$$

From now on we assume that CPT is conserved. If either CP or T is a symmetry of $\mathcal{H}$ (independently of whether CPT is conserved or violated), then M_{12} and Γ_{12} are relatively real, leading to

$$\left(\frac{q}{p}\right)^2 = e^{2i\xi_B} \quad \Rightarrow \quad \left|\frac{q}{p}\right| = 1 \,, \tag{B.9}$$

where ξ_B is the arbitrary unphysical phase introduced in Eq. (B.2).

The real and imaginary parts of the eigenvalues of $\mathcal{H}$ corresponding to $|B_{L,H}\rangle$ represent their masses and decay-widths, respectively. The mass difference Δm_B and the width difference $\Delta\Gamma_B$ are defined as follows:

$$\Delta m_B \equiv M_H - M_L, \quad \Delta\Gamma_B \equiv \Gamma_H - \Gamma_L \,. \tag{B.10}$$

Note that here Δm_B is positive by definition, while the sign of $\Delta\Gamma_B$ is to be experimentally determined. The average mass and width are given by

$$m_B \equiv \frac{M_H + M_L}{2}, \quad \Gamma_B \equiv \frac{\Gamma_H + \Gamma_L}{2} \,. \tag{B.11}$$

It is useful to define dimensionless ratios x and y:

$$x \equiv \frac{\Delta m_B}{\Gamma_B}, \quad y \equiv \frac{\Delta\Gamma_B}{2\Gamma_B}. \tag{B.12}$$

Solving the eigenvalue equation gives

$$(\Delta m_B)^2 - \frac{1}{4}(\Delta\Gamma_B)^2 = (4|M_{12}|^2 - |\Gamma_{12}|^2), \quad \Delta m_B \Delta\Gamma_B = 4\mathcal{R}e(M_{12}\Gamma_{12}^*). \tag{B.13}$$

All CP-violating observables in B and $\overline{B}$ decays to final states f and $\overline{f}$ can be expressed in terms of phase-convention-independent combinations of A_f, $\overline{A}_f$, $A_{\overline{f}}$, and $\overline{A}_{\overline{f}}$, together with, for neutral-meson decays only, q/p. CP violation in charged-meson decays depends only on the combination $|\overline{A}_{\overline{f}}/A_f|$, while CP violation in neutral-meson decays is complicated by $B^0 \leftrightarrow \overline{B}^0$ oscillations and depends, additionally, on $|q/p|$ and on $\lambda_f \equiv (q/p)(\overline{A}_f/A_f)$.

For neutral D, B, and B_s mesons, $\Delta\Gamma/\Gamma \ll 1$ and so both mass eigenstates must be considered in their evolution. We denote the state of an initially pure $|B^0\rangle$ or $|\overline{B}^0\rangle$ after an elapsed proper time t as $|B^0{}_{\rm phys}(t)\rangle$ or $|\overline{B}^0_{\rm phys}(t)\rangle$, respectively. Using the effective Hamiltonian approximation, we obtain

$$\begin{aligned} |B^0{}_{\rm phys}(t)\rangle &= g_+(t)\,|B^0\rangle - \frac{q}{p}\,g_-(t)|\overline{B}^0\rangle, \\ |\overline{B}^0_{\rm phys}(t)\rangle &= g_+(t)\,|\overline{B}^0\rangle - \frac{p}{q}\,g_-(t)|B^0\rangle \,, \end{aligned} \tag{B.14}$$

where

$$g_\pm(t) \equiv \frac{1}{2}\left(e^{-im_H t - \frac{1}{2}\Gamma_H t} \pm e^{-im_L t - \frac{1}{2}\Gamma_L t}\right). \tag{B.15}$$

One obtains the following time-dependent decay rates:

$$\frac{d\Gamma[B^0{}_{\rm phys}(t) \to f]/dt}{e^{-\Gamma t}\mathcal{N}_f} = \left(|A_f|^2 + |(q/p)\overline{A}_f|^2\right)\cosh(y\Gamma t) + \left(|A_f|^2 - |(q/p)\overline{A}_f|^2\right)\cos(x\Gamma t)$$

$$+ \quad 2\mathcal{R}e((q/p)A_f^*\overline{A}_f)\sinh(y\Gamma t) - 2\mathcal{I}m((q/p)A_f^*\overline{A}_f)\sin(x\Gamma t)\ , \quad \text{(B.16)}$$

$$\frac{d\Gamma[\overline{B}^0_{\rm phys}(t)\to f]/dt}{e^{-\Gamma t}\mathcal{N}_f} = \left(|(p/q)A_f|^2 + |\overline{A}_f|^2\right)\cosh(y\Gamma t) - \left(|(p/q)A_f|^2 - |\overline{A}_f|^2\right)\cos(x\Gamma t)$$

$$+ \quad 2\mathcal{R}e((p/q)A_f\overline{A}_f^*)\sinh(y\Gamma t) - 2\mathcal{I}m((p/q)A_f\overline{A}_f^*)\sin(x\Gamma t)\ , \quad \text{(B.17)}$$

where $\mathcal{N}_f$ is a common normalization factor. Decay rates to the CP-conjugate final state $\overline{f}$ are obtained analogously, with $\mathcal{N}_f = \mathcal{N}_{\overline{f}}$ and the substitutions $A_f \to A_{\overline{f}}$ and $\overline{A}_f \to \overline{A}_{\overline{f}}$ in Eqs. (B.16) and (B.17). Terms proportional to $|A_f|^2$ or $|\overline{A}_f|^2$ are associated with decays that occur without any net $B \leftrightarrow \overline{B}$ oscillation, while terms proportional to $|(q/p)\overline{A}_f|^2$ or $|(p/q)A_f|^2$ are associated with decays following a net oscillation. The $\sinh(y\Gamma t)$ and $\sin(x\Gamma t)$ terms of Eqs. (B.16) and (B.17) are associated with the interference between these two cases. Note that, in multi-body decays, amplitudes are functions of phase-space variables. Interference may be present in some regions but not in others, and is strongly influenced by resonant substructure.

One possible manifestation of CP-violating effects in meson decays [61] is in the interference between a decay without mixing, $B^0 \to f$, and a decay with mixing, $B^0 \to \overline{B}^0 \to f$ (such an effect occurs only in decays to final states that are common to B^0 and $\overline{B}^0$, including all CP eigenstates). It is defined by

$$\mathcal{I}m(\lambda_f) \neq 0\ , \quad \text{(B.18)}$$

with

$$\lambda_f \equiv \frac{q}{p}\frac{\overline{A}_f}{A_f}\ . \quad \text{(B.19)}$$

This form of CP violation can be observed, for example, using the asymmetry of neutral meson decays into final CP eigenstates f_{CP}

$$\mathcal{A}_{f_{CP}}(t) \equiv \frac{d\Gamma/dt[\overline{B}^0_{\rm phys}(t)\to f_{CP}\,] - d\Gamma/dt[B^0{}_{\rm phys}(t)\to f_{CP}\,]}{d\Gamma/dt[\overline{B}^0_{\rm phys}(t)\to f_{CP}\,] + d\Gamma/dt[B^0{}_{\rm phys}(t)\to f_{CP}\,]}\ . \quad \text{(B.20)}$$

For $\Delta\Gamma = 0$ and $|q/p| = 1$ (which is a good approximation for B mesons), $\mathcal{A}_{f_{CP}}$ has a particularly simple form [62–64]:

$$\begin{aligned}\mathcal{A}_f(t) &= S_f\sin(\Delta m t) - C_f\cos(\Delta m t),\\ S_f &\equiv \frac{2\mathcal{I}m(\lambda_f)}{1+|\lambda_f|^2},\quad C_f \equiv \frac{1-|\lambda_f|^2}{1+|\lambda_f|^2}\ .\end{aligned} \quad \text{(B.21)}$$

Consider the $B \to f$ decay amplitude A_f, and the CP conjugate process $\overline{B} \to \overline{f}$ with decay amplitude $\overline{A}_{\overline{f}}$. There are two types of phases that may appear in these decay amplitudes. Complex parameters in any Lagrangian term that contributes to the amplitude will appear in complex conjugate form in the CP-conjugate amplitude. Thus their phases appear in A_f and $\overline{A}_{\overline{f}}$ with opposite signs. In the Standard Model, these phases occur only in the couplings of the $W^\pm$ bosons and hence are often called 'weak phases'. The weak phase of any single term is convention dependent. However, the difference between the weak phases in two different terms in A_f is convention independent. A second type of phase can appear in scattering or decay amplitudes even when the Lagrangian is real. Their origin is the possible contribution from intermediate on-shell states in the decay process. Since these phases are generated by CP-invariant interactions, they are the same in A_f and $\overline{A}_{\overline{f}}$. Usually the dominant rescattering is due to strong interactions and hence the designation 'strong phases' for the phase shifts so induced. Again, only the relative strong phases between different terms in the amplitude are physically meaningful.

The 'weak' and 'strong' phases discussed here appear in addition to the 'spurious' CP transformation phases of Eq. (B.3). Those spurious phases are due to an arbitrary choice of phase convention, and

do not originate from any dynamics or induce any CP violation. For simplicity, we set them to zero from here on.

It is useful to write each contribution a_i to A_f in three parts: its magnitude $|a_i|$, its weak phase ϕ_i, and its strong phase δ_i. If, for example, there are two such contributions, $A_f = a_1 + a_2$, we have

$$\begin{aligned} A_f &= |a_1|e^{i(\delta_1+\phi_1)} + |a_2|e^{i(\delta_2+\phi_2)}, \\ \overline{A}_{\overline{f}} &= |a_1|e^{i(\delta_1-\phi_1)} + |a_2|e^{i(\delta_2-\phi_2)}. \end{aligned} \tag{B.22}$$

Similarly, for neutral meson decays, it is useful to write

$$M_{12} = |M_{12}|e^{i\phi_M} \quad , \quad \Gamma_{12} = |\Gamma_{12}|e^{i\phi_\Gamma} \ . \tag{B.23}$$

Each of the phases appearing in Eqs. (B.22) and (B.23) is convention dependent, but combinations such as $\delta_1 - \delta_2$, $\phi_1 - \phi_2$, $\phi_M - \phi_\Gamma$ and $\phi_M + \phi_1 - \overline{\phi}_1$ (where $\overline{\phi}_1$ is a weak phase contributing to $\overline{A}_f$) are physical.

In the approximations that only a single weak phase contributes to decay, $A_f = |a_f|e^{i(\delta_f+\phi_f)}$, and that $|\Gamma_{12}/M_{12}| = 0$, we obtain $|\lambda_f| = 1$ and the CP asymmetries in decays to a final CP eigenstate f [Eq. (B.20)] with eigenvalue $\eta_f = \pm 1$ are given by

$$\mathcal{A}_{f_{CP}}(t) = \mathcal{I}m(\lambda_f)\ \sin(\Delta m t)\ \text{ with }\ \mathcal{I}m(\lambda_f) = \eta_f \sin(\phi_M + 2\phi_f). \tag{B.24}$$

Note that the phase so measured is purely a weak phase, and no hadronic parameters are involved in the extraction of its value from $\mathcal{I}m(\lambda_f)$.

C Supersymmetric contributions to neutral meson mixing

We consider the squark–gluino box diagram contribution to D^0–$\overline{D}^0$ mixing amplitude that is proportional to $K^u_{2i}K^{u*}_{1i}K^u_{2j}K^{u*}_{1j}$, where K^u is the mixing matrix of the gluino couplings to left-handed up quarks and their up squark partners. (In the language of the mass insertion approximation, we calculate here the contribution that is $\propto [(\delta^u_{LL})_{12}]^2$.) We work in the mass basis for both quarks and squarks.

The contribution is given by

$$M^D_{12} = -i\frac{4\pi^2}{27}\alpha_s^2 m_D f_D^2 B_D \eta_{\rm QCD} \sum_{i,j}(K^u_{2i}K^{u*}_{1i}K^u_{2j}K^{u*}_{1j})(11\tilde{I}_{4ij} + 4\tilde{m}_g^2 I_{4ij}) \ , \tag{C.1}$$

where

$$\begin{aligned} \tilde{I}_{4ij} &\equiv \int \frac{d^4p}{(2\pi)^4}\frac{p^2}{(p^2-\tilde{m}_g^2)^2(p^2-\tilde{m}_i^2)(p^2-\tilde{m}_j^2)} \\ &= \frac{i}{(4\pi)^2}\left[\frac{\tilde{m}_g^2}{(\tilde{m}_i^2-\tilde{m}_g^2)(\tilde{m}_j^2-\tilde{m}_g^2)}\right. \\ &\quad \left. + \frac{\tilde{m}_i^4}{(\tilde{m}_i^2-\tilde{m}_j^2)(\tilde{m}_i^2-\tilde{m}_g^2)^2}\ln\frac{\tilde{m}_i^2}{\tilde{m}_g^2} + \frac{\tilde{m}_j^4}{(\tilde{m}_j^2-\tilde{m}_i^2)(\tilde{m}_j^2-\tilde{m}_g^2)^2}\ln\frac{\tilde{m}_j^2}{\tilde{m}_g^2}\right], \end{aligned} \tag{C.2}$$

$$\begin{aligned} I_{4ij} &\equiv \int \frac{d^4p}{(2\pi)^4}\frac{1}{(p^2-\tilde{m}_g^2)^2(p^2-\tilde{m}_i^2)(p^2-\tilde{m}_j^2)} \\ &= \frac{i}{(4\pi)^2}\left[\frac{1}{(\tilde{m}_i^2-\tilde{m}_g^2)(\tilde{m}_j^2-\tilde{m}_g^2)}\right. \end{aligned}$$

$$+\frac{\tilde{m}_i^2}{(\tilde{m}_i^2-\tilde{m}_j^2)(\tilde{m}_i^2-\tilde{m}_g^2)^2}\ln\frac{\tilde{m}_i^2}{\tilde{m}_g^2}+\frac{\tilde{m}_j^2}{(\tilde{m}_j^2-\tilde{m}_i^2)(\tilde{m}_j^2-\tilde{m}_g^2)^2}\ln\frac{\tilde{m}_j^2}{\tilde{m}_g^2}\Bigg]. \qquad \text{(C.3)}$$

We now follow the discussion in Refs. [23, 26]. To see the consequences of the super-GIM mechanism, let us expand the expression for the box integral around some value $\tilde{m}_q^2$ for the squark masses-squared:

$$\begin{aligned} I_4(\tilde{m}_g^2,\tilde{m}_i^2,\tilde{m}_j^2) &= I_4(\tilde{m}_g^2,\tilde{m}_q^2+\delta\tilde{m}_i^2,\tilde{m}_q^2+\delta\tilde{m}_j^2) \\ &= I_4(\tilde{m}_g^2,\tilde{m}_q^2,\tilde{m}_q^2)+(\delta\tilde{m}_i^2+\delta\tilde{m}_j^2)I_5(\tilde{m}_g^2,\tilde{m}_q^2,\tilde{m}_q^2,\tilde{m}_q^2) \\ &+ \frac{1}{2}\left[(\delta\tilde{m}_i^2)^2+(\delta\tilde{m}_j^2)^2+2(\delta\tilde{m}_i^2)(\delta\tilde{m}_j^2)\right]I_6(\tilde{m}_g^2,\tilde{m}_q^2,\tilde{m}_q^2,\tilde{m}_q^2,\tilde{m}_q^2)+\cdots \qquad \text{(C.4)} \end{aligned}$$

where

$$I_n(\tilde{m}_g^2,\tilde{m}_q^2,\ldots,\tilde{m}_q^2)\equiv\int\frac{d^4p}{(2\pi)^4}\frac{1}{(p^2-\tilde{m}_g^2)^2(p^2-\tilde{m}_q^2)^{n-2}}, \qquad \text{(C.5)}$$

and similarly for $\tilde{I}_{4ij}$. Note that $I_n\propto(\tilde{m}_q^2)^{n-2}$ and $\tilde{I}_n\propto(\tilde{m}_q^2)^{n-3}$. Thus, using $x\equiv\tilde{m}_g^2/\tilde{m}_q^2$, it is customary to define

$$I_n\equiv\frac{i}{(4\pi)^2(\tilde{m}_q^2)^{n-2}}f_n(x),\quad \tilde{I}_n\equiv\frac{i}{(4\pi)^2(\tilde{m}_q^2)^{n-3}}\tilde{f}_n(x). \qquad \text{(C.6)}$$

The unitarity of the mixing matrix implies that

$$\sum_i(K^u_{2i}K^{u*}_{1i}K^u_{2j}K^{u*}_{1j})=\sum_j(K^u_{2i}K^{u*}_{1i}K^u_{2j}K^{u*}_{1j})=0. \qquad \text{(C.7)}$$

We learn that the terms that are proportional f_4, $\tilde{f}_4$, f_5, and $\tilde{f}_5$ vanish in their contribution to M_{12}. When $\delta\tilde{m}_i^2\ll\tilde{m}_q^2$ for all i, the leading contributions to M_{12} come from f_6 and $\tilde{f}_6$. We learn that for quasi-degenerate squarks, the leading contribution is quadratic in the small mass-squared difference. The functions $f_6(x)$ and $\tilde{f}_6(x)$ are given by

$$\begin{aligned} f_6(x) &= \frac{6(1+3x)\ln x+x^3-9x^2-9x+17}{6(1-x)^5}, \\ \tilde{f}_6(x) &= \frac{6x(1+x)\ln x-x^3-9x^2+9x+1}{3(1-x)^5}. \end{aligned} \qquad \text{(C.8)}$$

For example, with $x=1$, $f_6(1)=-1/20$ and $\tilde{f}_6=+1/30$; with $x=2.33$, $f_6(2.33)=-0.015$ and $\tilde{f}_6=+0.013$.

To further simplify things, let us consider a two-generation case. Then

$$\begin{aligned} M_{12}^D &\propto 2(K^u_{21}K^{u*}_{11})^2(\delta\tilde{m}_1^2)^2+2(K^u_{22}K^{u*}_{12})^2(\delta\tilde{m}_2^2)^2+(K^u_{21}K^{u*}_{11}K^u_{22}K^{u*}_{12})(\delta\tilde{m}_1^2+\delta\tilde{m}_2^2)^2 \\ &= (K^u_{21}K^{u*}_{11})^2(\tilde{m}_2^2-\tilde{m}_1^2)^2. \end{aligned} \qquad \text{(C.9)}$$

We thus rewrite Eq. (C.1) for the case of quasi-degenerate squarks:

$$M_{12}^D=\frac{\alpha_s^2 m_D f_D^2 B_D\eta_{\rm QCD}}{108\tilde{m}_q^2}[11\tilde{f}_6(x)+4xf_6(x)]\frac{(\Delta\tilde{m}_{21}^2)^2}{\tilde{m}_q^4}(K^u_{21}K^{u*}_{11})^2. \qquad \text{(C.10)}$$

For example, for $x=1$, $11\tilde{f}_6(x)+4xf_6(x)=+0.17$. For $x=2.33$, $11\tilde{f}_6(x)+4xf_6(x)=+0.003$.

D Neutrino flavour transitions

D.1 Neutrinos in vacuum

Neutrino oscillations in vacuum [65] arise since neutrinos are massive and mix. In other words, the neutrino state that is produced by electroweak interactions is not a mass eigenstate. The weak eigenstates ν_α ($\alpha = e, \mu, \tau$ denotes the charged lepton mass eigenstates and their neutrino doublet-partners) are linear combinations of the mass eigenstates ν_i ($i = 1, 2, 3$):

$$|\nu_\alpha\rangle = U^*_{\alpha i}|\nu_i\rangle. \tag{D.1}$$

After travelling a distance L (or, equivalently for relativistic neutrinos, time t), a neutrino originally produced with a flavour α evolves as follows:

$$|\nu_\alpha(t)\rangle = U^*_{\alpha i}|\nu_i(t)\rangle. \tag{D.2}$$

It can be detected in the charged-current interaction $\nu_\alpha(t)N' \to \ell_\beta N$ with a probability

$$P_{\alpha\beta} = |\langle\nu_\beta|\nu_\alpha(t)\rangle|^2 = \left|\sum_{i=1}^{3}\sum_{j=1}^{3} U^*_{\alpha i}U_{\beta j}\langle\nu_j(0)|\nu_i(t)\rangle\right|^2. \tag{D.3}$$

We follow the analysis of Ref. [44]. We use the standard approximation that $|\nu\rangle$ is a plane wave, $|\nu_i(t)\rangle = e^{-iE_it}|\nu_i(0)\rangle$. In all cases of interest to us, the neutrinos are relativistic:

$$E_i = \sqrt{p_i^2 + m_i^2} \simeq p_i + \frac{m_i^2}{2E_i}, \tag{D.4}$$

where E_i and m_i are, respectively, the energy and the mass of the neutrino mass eigenstate. Furthermore, we can assume that $p_i \simeq p_j \equiv p \simeq E$. Then, we obtain the following transition probability:

$$P_{\alpha\beta} = \delta_{\alpha\beta} - 4\sum_{i=1}^{2}\sum_{j=i+1}^{3} \mathcal{R}e\left(U_{\alpha i}U^*_{\beta i}U^*_{\alpha j}U_{\beta j}\right)\sin^2 x_{ij}, \tag{D.5}$$

where $x_{ij} \equiv \Delta m^2_{ij}L/(4E)$, $\Delta m^2_{ij} = m^2_i - m^2_j$, and $L = t$ is the distance between the source (that is, the production point of ν_α) and the detector (that is, the detection point of ν_β). In deriving Eq. (D.5) we used the orthogonality relation $\langle\nu_j(0)|\nu_i(0)\rangle = \delta_{ij}$. It is convenient to use the following units:

$$x_{ij} = 1.27\,\frac{\Delta m^2_{ij}}{\mathrm{eV}^2}\,\frac{L/E}{\mathrm{m/MeV}}. \tag{D.6}$$

The transition probability [Eq. (D.5)] has an oscillatory behaviour, with oscillation lengths

$$L^{\mathrm{osc}}_{0,ij} = \frac{4\pi E}{\Delta m^2_{ij}} \tag{D.7}$$

and amplitude that is proportional to elements of the mixing matrix. Thus, in order to have oscillations, neutrinos must have different masses ($\Delta m^2_{ij} \neq 0$) and they must mix ($U_{\alpha i}U_{\beta i} \neq 0$).

An experiment is characterized by the typical neutrino energy E and by the source-detector distance L. In order to be sensitive to a given value of Δm^2_{ij}, the experiment has to be set up with $E/L \approx \Delta m^2_{ij}$ ($L \sim L^{\mathrm{osc}}_{0,ij}$). The typical values of L/E for different types of neutrino sources and experiments are summarized in Table D.1.

If $(E/L) \gg \Delta m^2_{ij}$ ($L \ll L^{\mathrm{osc}}_{0,ij}$), the oscillation does not have time to give an appreciable effect because $\sin^2 x_{ij} \ll 1$. The case of $(E/L) \ll \Delta m^2_{ij}$ ($L \gg L^{\mathrm{osc}}_{0,ij}$) requires more careful consideration.

Table D.1: Characteristic values of L and E for various neutrino sources and experiments.

Experiment	L (m)	E (MeV)	Δm^2 (eV^2)
Solar	10^{10}	1	10^{-10}
Atmospheric	10^4–10^7	10^2–10^5	10^{-1}–10^{-4}
Reactor	10^2–10^3	1	10^{-2}–10^{-3}
KamLAND	10^5	1	10^{-5}
Accelerator	10^2	10^3–10^4	$\gtrsim 10^{-1}$
Long-baseline accelerator	10^5–10^6	10^4	10^{-2}–10^{-3}

One must take into account that, in general, neutrino beams are not monochromatic. Thus, rather than measuring $P_{\alpha\beta}$, the experiments are sensitive to the average probability

$$\langle P_{\alpha\beta}\rangle = \delta_{\alpha\beta} - 4\sum_{i=1}^{2}\sum_{j=i+1}^{3} \mathcal{R}e\left(U_{\alpha i}U^*_{\beta i}U^*_{\alpha j}U_{\beta j}\right)\langle\sin^2 x_{ij}\rangle. \tag{D.8}$$

For $L \gg L^{\rm osc}_{0,ij}$, the oscillation phase goes through many cycles before the detection and is averaged to $\langle\sin^2 x_{ij}\rangle = 1/2$.

For a two-neutrino case,

$$P_{\alpha\beta} = \delta_{\alpha\beta} - (2\delta_{\alpha\beta} - 1)\sin^2 2\theta \sin^2 x. \tag{D.9}$$

For averaged oscillations we get, for example,

$$P_{ee} = 1 - \frac{1}{2}\sin^2 2\theta. \tag{D.10}$$

For a recent careful derivation of the oscillation formulae, see Ref. [66].

D.2 Neutrinos in matter

When neutrinos propagate in dense matter, the interactions with the medium affect their properties. These effects are either coherent or incoherent. For purely incoherent ν–p scattering, the characteristic cross-section is very small,

$$\sigma \sim \frac{G_F^2 s}{\pi} \sim 10^{-43}\ \text{cm}^2 \left(\frac{E}{1\ \text{MeV}}\right)^2 . \tag{D.11}$$

The smallness of this cross-section is demonstrated by the fact that if a beam of 10^{10} neutrinos with $E \sim 1$ MeV was aimed at Earth, only one would be deflected by the Earth's matter. It may seem then that for neutrinos matter is irrelevant. However, one must take into account that Eq. (D.11) does not contain the contribution from forward elastic coherent interactions. In coherent interactions, the medium remains unchanged and it is possible to have interference of scattered and unscattered neutrino waves which enhances the effect. Coherence further allows one to decouple the evolution equation of neutrinos from the equations of the medium. In this approximation, the effect of the medium is described by an effective potential which depends on the density and composition of the matter [67].

Consider, for example, the effective potential for ν_e induced by its charged-current interactions with electrons in matter:

$$V_C = \langle\nu_e| \int d^3x H_C^{(e)} |\nu_e\rangle = \sqrt{2} G_F N_e. \tag{D.12}$$

For $\overline{\nu_e}$ the sign of V is reversed. The potential can also be expressed in terms of the matter density ρ:

$$V_C = 7.6 \, \frac{N_e}{N_p + N_n} \, \frac{\rho}{10^{14} \text{ g/cm}^3} \text{ eV} \, . \tag{D.13}$$

Two examples that are relevant to observations are the following:

- At the Earth's core $\rho \sim 10$ g/cm^3 and $V \sim 10^{-13}$ eV.
- At the solar core $\rho \sim 100$ g/cm^3 and $V \sim 10^{-12}$ eV.

Consider a state that is an admixture of two neutrino species, $|\nu_e\rangle$ and $|\nu_a\rangle$ or, equivalently, $|\nu_1\rangle$ and $|\nu_2\rangle$. With some approximations, the time evolution can be written in the following matrix form [67]:

$$-i\frac{\partial}{\partial x}\begin{pmatrix}\nu_e \\ \nu_a\end{pmatrix} = -\frac{1}{2E} M_w^2 \begin{pmatrix}\nu_e \\ \nu_a\end{pmatrix} \, , \tag{D.14}$$

where we have defined an effective mass matrix in matter,

$$M_w^2 = \frac{1}{2}\begin{pmatrix} m_1^2 + m_2^2 + 4EV_e - \Delta m^2 \cos 2\theta & \Delta m^2 \sin 2\theta \\ \Delta m^2 \sin 2\theta & m_1^2 + m_2^2 + 4EV_a + \Delta m^2 \cos 2\theta \end{pmatrix} \, , \tag{D.15}$$

with $\Delta m^2 = m_2^2 - m_1^2$.

We define the instantaneous mass eigenstates in matter, ν_i^m, as the eigenstates of M_w for a fixed value of x. They are related to the interaction eigenstates by a unitary transformation,

$$\begin{pmatrix}\nu_e \\ \nu_a\end{pmatrix} = U(\theta_m)\begin{pmatrix}\nu_1^m \\ \nu_2^m\end{pmatrix} = \begin{pmatrix}\cos\theta_m & \sin\theta_m \\ -\sin\theta_m & \cos\theta_m\end{pmatrix}\begin{pmatrix}\nu_1^m \\ \nu_2^m\end{pmatrix} \, . \tag{D.16}$$

The eigenvalues of M_w, that is, the effective masses in matter, are given by [67, 68]

$$\mu_{1,2}^2 = \frac{m_1^2 + m_2^2}{2} + E(V_e + V_a) \mp \frac{1}{2}\sqrt{(\Delta m^2 \cos 2\theta - A)^2 + (\Delta m^2 \sin 2\theta)^2}, \tag{D.17}$$

while the mixing angle in matter is given by

$$\tan 2\theta_m = \frac{\Delta m^2 \sin 2\theta}{\Delta m^2 \cos 2\theta - A}, \tag{D.18}$$

where

$$A \equiv 2E(V_e - V_a). \tag{D.19}$$

The instantaneous mass eigenstates ν_i^m are, in general, not energy eigenstates: they mix in the evolution. The importance of this effect is controlled by the relative size of $4E\dot{\theta}_m(t)$ with respect to $\mu_2^2(t) - \mu_1^2(t)$. When the latter is much larger than the first, ν_i^m behave approximately as energy eigenstates and do not mix during the evolution. This is the adiabatic transition approximation. The adiabaticity condition reads

$$\mu_2^2(t) - \mu_1^2(t) \gg 2EA\Delta m^2 \sin 2\theta \left|\dot{A}/A\right| . \tag{D.20}$$

The transition probability for the adiabatic case is given by

$$P_{ee}(t) = \left| \sum_i U_{ei}(\theta) U_{ei}^*(\theta_p) \exp\left(-\frac{i}{2E}\int_{t_0}^t \mu_i^2(t')dt'\right)\right|^2 , \tag{D.21}$$

where θ_p is the mixing angle at the production point. For the case of two-neutrino mixing, Eq. (D.21) takes the form

$$P_{ee}(t) = \cos^2\theta_p \cos^2\theta + \sin^2\theta_p \sin^2\theta + \frac{1}{2}\sin 2\theta_p \sin 2\theta \cos\left(\frac{\delta(t)}{2E}\right), \tag{D.22}$$

where

$$\delta(t) = \int_{t_p}^{t} [\mu_2^2(t') - \mu_1^2(t')]dt'. \tag{D.23}$$

For $\mu_2^2(t) - \mu_1^2(t) \gg E$, the last term in Eq. (D.22) is averaged out and the survival probability takes the form

$$P_{ee} = \frac{1}{2}[1 + \cos 2\theta_p \cos 2\theta]. \tag{D.24}$$

The relative importance of the MSW matter term [A of Eq. (D.19)] and the kinematic vacuum oscillation term in the Hamiltonian [the off-diagonal term in Eq. (D.15)] can be parametrized by the quantity $\beta_{\rm MSW}$, which represents the ratio of matter to vacuum effects (see, for example, Ref. [69]). From Eq. (D.15) we see that the appropriate ratio is

$$\beta_{\rm MSW} = \frac{2\sqrt{2}G_F n_e E_\nu}{\Delta m^2}. \tag{D.25}$$

The quantity $\beta_{\rm MSW}$ is the ratio between the oscillation length in matter and the oscillation length in vacuum. In convenient units, $\beta_{\rm MSW}$ can be written as

$$\beta_{\rm MSW} = 0.19\left(\frac{E_\nu}{1\ {\rm MeV}}\right)\left(\frac{\mu_e\rho}{100\ {\rm g\ cm^{-3}}}\right)\left(\frac{8\times 10^{-5}\ {\rm eV}^2}{\Delta m^2}\right). \tag{D.26}$$

Here μ_e is the electron mean molecular weight ($\mu_e \approx 0.5(1+X)$, where X is the mass fraction of hydrogen) and ρ is the total density. If $\beta_{\rm MSW} \lesssim \cos 2\theta$, the survival probability corresponds to vacuum averaged oscillations [see Eq. (D.9)],

$$P_{ee} = \left(1 - \frac{1}{2}\sin^2 2\theta\right) \quad (\beta_{\rm MSW} < \cos 2\theta,\ \text{vacuum}). \tag{D.27}$$

If $\beta_{\rm MSW} > 1$, the survival probability corresponds to matter-dominated oscillations [see Eq. (D.24)],

$$P_{ee} = \sin^2\theta \quad (\beta_{\rm MSW} > 1,\ \text{MSW}). \tag{D.28}$$

The survival probability is approximately constant in either of the two limiting regimes, $\beta_{\rm MSW} < \cos 2\theta$ and $\beta_{\rm MSW} > 1$. There is a strong energy dependence only in the transition region between the limiting regimes.

For the Sun, $N_e(R) = N_e(0)\exp(-R/r_0)$, with $r_0 \equiv R_\odot/10.54 = 6.6\times 10^7$ m $= 3.3\times 10^{14}$ eV^{-1}. Then, the adiabaticity condition for the Sun reads

$$\frac{(\Delta m^2/{\rm eV}^2)\sin^2 2\theta}{(E/{\rm MeV})\cos 2\theta} \gg 3\times 10^{-9}. \tag{D.29}$$

References

[1] M. Kobayashi and T. Maskawa, Prog. Theor. Phys. **49**, 652 (1973).

[2] N. Cabibbo, Phys. Rev. Lett. **10**, 531 (1963).

[3] A. B. Carter and A. I. Sanda, Phys. Rev. Lett. **45**, 952 (1980); Phys. Rev. D **23**, 1567 (1981).

[4] I. I. Y. Bigi and A. I. Sanda, Nucl. Phys. B **193**, 85 (1981).

[5] G. Buchalla, A. J. Buras, and M. E. Lautenbacher, Rev. Mod. Phys. **68**, 1125 (1996) [arXiv:hep-ph/9512380].

[6] Y. Grossman, A. L. Kagan, and Z. Ligeti, Phys. Lett. B **538**, 327 (2002) [arXiv:hep-ph/0204212].

[7] H. Boos, T. Mannel, and J. Reuter, Phys. Rev. D **70**, 036006 (2004) [arXiv:hep-ph/0403085].

[8] H. n. Li and S. Mishima, JHEP **0703**, 009 (2007) [arXiv:hep-ph/0610120].

[9] M. Gronau and J. L. Rosner, Phys. Lett. B **672**, 349 (2009) [arXiv:0812.4796 [hep-ph]].

[10] E. Barberio *et al.* [Heavy Flavor Averaging Group], arXiv:0808.1297 [hep-ex], online update at http://www.slac.stanford.edu/xorg/hfag

[11] C. Amsler *et al.* [Particle Data Group], Phys. Lett. B **667**, 1 (2008).

[12] CKMfitter Group (J. Charles *et al.*), Eur. Phys. J. C **41**, 1–131 (2005), [hep-ph/0406184], updated results and plots available at: http://ckmfitter.in2p3.fr

[13] Y. Nir, Nucl. Phys. Proc. Suppl. **117**, 111 (2003) [arXiv:hep-ph/0208080].

[14] Y. Grossman, Y. Nir, and M. P. Worah, Phys. Lett. B **407**, 307 (1997) [hep-ph/9704287].

[15] Y. Grossman, Y. Nir, and G. Raz, Phys. Rev. Lett. **97**, 151801 (2006) [arXiv:hep-ph/0605028].

[16] M. Bona *et al.* [UTfit Collaboration], JHEP **0803**, 049 (2008) [arXiv:0707.0636 [hep-ph]].

[17] G. C. Branco, L. Lavoura, and J. P. Silva, *CP Violation* (Clarendon Press, Oxford, 1999).

[18] I. I. Y. Bigi and N. G. Uraltsev, Nucl. Phys. B **592**, 92 (2001) [arXiv:hep-ph/0005089].

[19] A. F. Falk, Y. Grossman, Z. Ligeti, and A. A. Petrov, Phys. Rev. D **65**, 054034 (2002) [arXiv:hep-ph/0110317].

[20] A. F. Falk, Y. Grossman, Z. Ligeti, Y. Nir, and A. A. Petrov, Phys. Rev. D **69**, 114021 (2004) [arXiv:hep-ph/0402204].

[21] B. Aubert *et al.* [BaBar Collaboration], Phys. Rev. Lett. **98**, 211802 (2007) [arXiv:hep-ex/0703020].

[22] M. Staric *et al.* [Belle Collaboration], Phys. Rev. Lett. **98**, 211803 (2007) [arXiv:hep-ex/0703036].

[23] G. Raz, Phys. Rev. D **66**, 037701 (2002) [arXiv:hep-ph/0205310].

[24] N. Arkani-Hamed and S. Dimopoulos, JHEP **0506**, 073 (2005) [arXiv:hep-th/0405159].

[25] A. G. Cohen, D. B. Kaplan, and A. E. Nelson, Phys. Lett. B **388**, 588 (1996) [arXiv:hep-ph/9607394].

[26] Y. Nir and G. Raz, Phys. Rev. D **66**, 035007 (2002) [arXiv:hep-ph/0206064].

[27] K. Blum, Y. Grossman, Y. Nir and G. Perez, Phys. Rev. Lett. **102**, 211802 (2009) [arXiv:0903.2118 [hep-ph]].

[28] Y. Nir and N. Seiberg, Phys. Lett. B **309**, 337 (1993) [arXiv:hep-ph/9304307].

[29] M. Leurer, Y. Nir, and N. Seiberg, Nucl. Phys. B **420**, 468 (1994) [arXiv:hep-ph/9310320].

[30] M. Ciuchini, E. Franco, D. Guadagnoli, V. Lubicz, M. Pierini, V. Porretti, and L. Silvestrini, Phys. Lett. B **655**, 162 (2007) [arXiv:hep-ph/0703204].

[31] Y. Nir, JHEP **0705**, 102 (2007) [arXiv:hep-ph/0703235].

[32] G. D'Ambrosio, G. F. Giudice, G. Isidori, and A. Strumia, Nucl. Phys. B **645**, 155 (2002) [arXiv:hep-ph/0207036].

[33] Y. Grossman, Y. Nir, J. Thaler, T. Volansky, and J. Zupan, Phys. Rev. D **76**, 096006 (2007) [arXiv:0706.1845 [hep-ph]].

[34] J. L. Feng, C. G. Lester, Y. Nir, and Y. Shadmi, Phys. Rev. D **77**, 076002 (2008) [arXiv:0712.0674 [hep-ph]].

[35] G. Engelhard, J. L. Feng, I. Galon, D. Sanford and F. Yu, arXiv:0904.1415 [hep-ph].

[36] J. L. Feng, I. Galon, D. Sanford, Y. Shadmi and F. Yu, Phys. Rev. D **79**, 116009 (2009) [arXiv:0904.1416 [hep-ph]].

[37] J. L. Feng, S. T. French, C. G. Lester, Y. Nir and Y. Shadmi, arXiv:0906.4215 [hep-ph].

[38] G. Hiller and Y. Nir, JHEP **0803**, 046 (2008) [arXiv:0802.0916 [hep-ph]].

[39] G. Hiller, Y. Hochberg, and Y. Nir, arXiv:0812.0511 [hep-ph].

[40] Y. Nomura, M. Papucci, and D. Stolarski, Phys. Rev. D **77**, 075006 (2008) [arXiv:0712.2074 [hep-ph]]; JHEP **0807**, 055 (2008) [arXiv:0802.2582 [hep-ph]].

[41] G. Hiller, Y. Hochberg, and Y. Nir, work in progress.

[42] G. F. Giudice, M. Nardecchia, and A. Romanino, arXiv:0812.3610 [hep-ph].

[43] A. E. Nelson and M. J. Strassler, JHEP **0009**, 030 (2000) [arXiv:hep-ph/0006251]; JHEP **0207**, 021 (2002) [arXiv:hep-ph/0104051].

[44] M. C. Gonzalez-Garcia and Y. Nir, Rev. Mod. Phys. **75**, 345 (2003) [arXiv:hep-ph/0202058].

[45] M. C. Gonzalez-Garcia and M. Maltoni, Phys. Rep. **460**, 1 (2008) [arXiv:0704.1800 [hep-ph]].

[46] M. Fukugita and T. Yanagida, Phys. Lett. B **174**, 45 (1986).

[47] S. Davidson, E. Nardi, and Y. Nir, Phys. Rep. **466**, 105 (2008) [arXiv:0802.2962 [hep-ph]].

[48] S. Davidson and A. Ibarra, Phys. Lett. B **535**, 25 (2002) [arXiv:hep-ph/0202239].

[49] P. F. Harrison, D. H. Perkins, and W. G. Scott, Phys. Lett. B **530**, 167 (2002) [arXiv:hep-ph/0202074].

[50] C. D. Froggatt and H. B. Nielsen, Nucl. Phys. B **147**, 277 (1979).

[51] M. Leurer, Y. Nir, and N. Seiberg, Nucl. Phys. B **398**, 319 (1993) [arXiv:hep-ph/9212278].

[52] Y. Grossman and Y. Nir, Nucl. Phys. B **448**, 30 (1995) [arXiv:hep-ph/9502418].

[53] L. J. Hall, H. Murayama, and N. Weiner, Phys. Rev. Lett. **84**, 2572 (2000) [arXiv:hep-ph/9911341].

[54] Y. E. Antebi, Y. Nir, and T. Volansky, Phys. Rev. D **73**, 075009 (2006) [arXiv:hep-ph/0512211].

[55] L. Chau and W. Keung, Phys. Rev. Lett. **53**, 1802 (1984).

[56] L. Wolfenstein, Phys. Rev. Lett. **51**, 1945 (1983).

[57] A. J. Buras, M. E. Lautenbacher, and G. Ostermaier, Phys. Rev. D **50**, 3433 (1994) [arXiv:hep-ph/9403384].

[58] C. Dib, I. Dunietz, F. J. Gilman, and Y. Nir, Phys. Rev. D **41**, 1522 (1990).

[59] J. L. Rosner, A. I. Sanda, and M. P. Schmidt, EFI-88-12-CHICAGO [Presented at Workshop on High Sensitivity Beauty Physics, Batavia, IL, Nov 11–14, 1987].

[60] V. Weisskopf and E. P. Wigner, Z. Phys. **63**, 54 (1930); Z. Phys. **65**, 18 (1930). [See Appendix A of P. K. Kabir, *The CP Puzzle: Strange Decays of the Neutral Kaon* (Academic Press, London, 1968).]

[61] Y. Nir, SLAC-PUB-5874 [Lectures given at *20th Summer Institute on Particle Physics: The Third Family and the Physics of Flavor*, Stanford, CA, 1992, ed. L. Vassilian (SLAC, Stanford, 1993)].

[62] I. Dunietz and J. L. Rosner, Phys. Rev. D **34**, 1404 (1986).

[63] Ya. I. Azimov, N. G. Uraltsev, and V. A. Khoze, Sov. J. Nucl. Phys. **45**, 878 (1987) [Yad. Fiz. **45**, 1412 (1987)].

[64] I. I. Bigi and A. I. Sanda, Nucl. Phys. B **281**, 41 (1987).

[65] B. Pontecorvo, Sov. Phys. JETP **6**, 429 (1957) [Zh. Eksp. Teor. Fiz. **33**, 549 (1957)].

[66] A. G. Cohen, S. L. Glashow, and Z. Ligeti, arXiv:0810.4602 [hep-ph].

[67] L. Wolfenstein, Phys. Rev. D **17**, 2369 (1978).

[68] S.P. Mikheyev and A. Yu. Smirnov, Sov. J. Nucl. Phys. **42**, 913 (1985) [Yad. Fiz. **42**, 1441 (1985)].

[69] J. N. Bahcall and C. Pena-Garay, New J. Phys. **6**, 63 (2004) [arXiv:hep-ph/0404061].

Particle cosmology

A. Riotto
CERN, Geneva, Switzerland

Abstract
In these lectures the present status of the so-called standard cosmological model, based on the hot Big Bang theory and the inflationary paradigm is reviewed. Special emphasis is given to the origin of the cosmological perturbations we see today under the form of the cosmic microwave background anisotropies and the large scale structure and to the dark matter and dark energy puzzles.

1 Introduction

The evolution of the universe is determined to a large extent by the same microphysics laws of physics that govern high-energy physics phenomena. Hence, any progress in particle physics has a large impact on the cosmological model(s) and, conversely, any new step taken towards the understanding of the past, present and future of our universe might provide a hint of high-energy physics beyond the one we currently know. This is the reason why these lectures are entitled Particle Cosmology. If the reader takes only one lesson home from them it is that particle physics and cosmology are nowadays intimately connected.

There are fundamental questions we are on the edge of answering: what is the origin of our universe? Why is the universe so homogeneous and isotropic on large scales? What are the origins of dark matter and dark energy? What is the fate of our universe? While these lectures will certainly not be able to give definite answers to them, we shall try to provide the students with some tools they might find useful in order to solve these overwhelming mysteries themselves.

These lectures will contain a short review of the standard Big Bang model; a rather long discussion of the inflation paradigm with particular emphasis on the possibility that the cosmological seeds originated from a period of primordial acceleration; the physics of the Cosmic Microwave Background (CMB) anisotropies, and a discussion of the dark matter and dark energy puzzles.

Since these lectures were delivered at a school, we shall not provide an exhaustive list of references to original material, but refer to several basic cosmology books and reviews where students can find the references to the original material [1–8].

2 Basics of the Big Bang model

We know two basic facts about our local universe (the universe we may observe). First, it is homogeneous and isotropic on sufficiently large cosmological scales [2]. Once this experimental evidence is accepted, one can promote it to a principle, dubbed "the cosmological principle". Secondly, it expands. The next question would then be: how can we describe such a universe?

The standard cosmology is based upon the maximally spatially symmetric Friedmann–Robertson–Walker (FRW) line element

$$ds^2 = -dt^2 + a(t)^2 \left[\frac{dr^2}{1-kr^2} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \right] ; \qquad (1)$$

where $a(t)$ is the cosmic-scale factor, $R_{\rm curv} \equiv a(t)|k|^{-1/2}$ is the curvature radius, and $k = -1, 0, 1$ is the curvature signature. All three models are without boundary: the positively curved model is finite and curves back on itself; the negatively curved and flat models are infinite in extent. The Robertson–Walker metric embodies the observed isotropy and homogeneity of the universe. It is interesting to note that

this form of the line element was originally introduced for the sake of mathematical simplicity; we now know that it is well justified at early times or today on large scales ($\gg 10\,\mathrm{Mpc}$), at least within our visible patch.

The coordinates, r, θ, and ϕ, are referred to as *co-moving* coordinates: A particle at rest in these coordinates remains at rest, i.e., constant r, θ, and ϕ. A freely moving particle eventually comes to rest in these coordinates, as its momentum is redshifted by the expansion, $p \propto a^{-1}$. Motion with respect to the co-moving coordinates (or cosmic rest frame) is referred to as peculiar velocity; unless supported by the inhomogeneous distribution of matter, peculiar velocities decay away as a^{-1}. Thus the measurement of peculiar velocities, which is not easy as it requires independent measures of both the distance and velocity of an object, can be used to probe the distribution of mass in the universe.

Physical separations between freely moving particles scale as $a(t)$; or said another way the physical separation between two points is simply $a(t)$ times the coordinate separation. The momenta of freely propagating particles decrease, or redshift, as $a(t)^{-1}$, and thus the wavelength of a photon stretches as $a(t)$, which is the origin of the cosmological redshift. The redshift suffered by a photon emitted from a distant galaxy $1 + z = a_0/a(t)$; that is, a galaxy whose light is redshifted by $1 + z$, emitted that light when the universe was a factor of $(1 + z)^{-1}$ smaller. When the light from the most distant quasar yet seen ($z = 4.9$) was emitted, the universe was a factor of almost six smaller; when CMB photons last scattered, the universe was about 1100 times smaller.

2.1 Friedmann equations

The evolution of the scale factor $a(t)$ is governed by Einstein equations

$$R_{\mu\nu} - \frac{1}{2}\, R\, g_{\mu\nu} \equiv G_{\mu\nu} = 8\pi G\,, T_{\mu\nu} \tag{2}$$

where $R_{\mu\nu}$ $(\mu, \nu = 0, \cdots 3)$ is the Riemann tensor and R is the Ricci scalar constructed via the metric (1) [2], and $T_{\mu\nu}$ is the energy-momentum tensor. $G = m_{\rm Pl}^{-2}$ is the Newton constant. Under the hypothesis of homogeneity and isotropy, we can always write the energy-momentum tensor under the form $T_{\mu\nu} = \mathrm{diag}\,(\rho, P, P, P)$ where ρ is the energy density of the system and P its pressure. They are functions of time.

The evolution of the cosmic-scale factor is governed by the Friedmann equation

$$H^2 \equiv \left(\frac{\dot a}{a}\right)^2 = \frac{8\pi G\rho}{3} - \frac{k}{a^2}\,, \tag{3}$$

where ρ is the total energy density of the universe, matter, radiation, vacuum energy, and so on.

Differentiating wrt to time both members of Eq. (3) and using the the mass conservation equation

$$\dot\rho + 3H(\rho + P) = 0\,, \tag{4}$$

we find the equation for the acceleration of the scale factor

$$\frac{\ddot a}{a} = -\frac{4\pi G}{3}(\rho + 3P). \tag{5}$$

Combining Eqs. (3) and (5) we find

$$\dot H = -4\pi G\,(\rho + P)\,. \tag{6}$$

The evolution of the energy density of the universe is governed by

$$d(\rho a^3) = -P d\left(a^3\right); \tag{7}$$

which is the first law of thermodynamics for a fluid in the expanding universe. (In the case that the stress energy of the universe is comprised of several, non-interacting components, this relation applies to each

separately; *e.g.*, to the matter and radiation separately today.) For $P = \rho/3$, ultra-relativistic matter, $\rho \propto a^{-4}$ and $a \sim t^{\frac{1}{2}}$; for $P = 0$, very nonrelativistic matter, $\rho \propto a^{-3}$ and $a \sim t^{\frac{2}{3}}$; and for $P = -\rho$, vacuum energy, $\rho =$ const. If the rhs of the Friedmann equation is dominated by a fluid with equation of state $P = w\rho$, it follows that $\rho \propto a^{-3(1+w)}$ and $a \propto t^{2/3(1+w)}$.

We can use the Friedmann equation to relate the curvature of the universe to the energy density and expansion rate:

$$\Omega - 1 = \frac{k}{a^2 H^2}\,; \qquad \Omega = \frac{\rho}{\rho_{\rm crit}}\,; \tag{8}$$

and the critical density today $\rho_{\rm crit} = 3H^2/8\pi G = 1.88 h^2\,{\rm g\,cm^{-3}} \simeq 1.05 \times 10^4\,{\rm eV\,cm^{-3}}$. There is a one-to-one correspondence between Ω and the spatial curvature of the universe: positively curved, $\Omega_0 > 1$; negatively curved, $\Omega_0 < 1$; and flat ($\Omega_0 = 1$). Further, the fate of the universe is determined by the curvature: model universes with $k \leq 0$ expand forever, while those with $k > 0$ necessarily recollapse. The curvature radius of the universe is related to the Hubble radius and Ω by

$$R_{\rm curv} = \frac{H^{-1}}{|\Omega - 1|^{1/2}}\,. \tag{9}$$

In physical terms, the curvature radius sets the scale for the size of spatial separations where the effects of curved space become pronounced. And in the case of the positively curved model it is just the radius of the 3-sphere.

The energy content of the universe consists of matter and radiation (today, photons and neutrinos). Since the photon temperature is accurately known, $T_0 = 2.73 \pm 0.01\,{\rm K}$, the fraction of critical density contributed by radiation is also accurately known: $\Omega_R h^2 = 4.2 \times 10^{-5}$, where $h = 0.72 \pm 0.07$ is the present Hubble rate in units of 100 km ${\rm s}^{-1}$ ${\rm Mpc}^{-1}$ [9]. The remaining content of the universe is another matter. Rapid progress has been made recently toward the measurement of cosmological parameters [10]. Over the past years the basic features of our universe have been determined. The universe is spatially flat; accelerating; comprised of one third dark matter and two thirds a new form of dark energy. The measurements of the cosmic microwave background anisotropies at different angular scales performed by the WMAP Collaboration [9] have recently significantly increased the case for accelerated expansion in the early universe (the inflationary paradigm) and at the current epoch (dark energy dominance), especially when combined with data on high-redshift supernovae (SN1) and large-scale structure (LSS) [10]. The CMB+LSS+SN1 data give [9]

$$\Omega_0 = 1.00^{+0.07}_{-0.03}\,,$$

meaning that the present universe is spatially flat (or at least very close to being flat). Restricting to $\Omega_0 = 1$, the dark matter density is given by [9]

$$\Omega_{\rm DM} h^2 = 0.11^{+0.0034}_{-0.059}\,,$$

and a baryon density

$$\Omega_B = 0.045 \pm 0.0015,$$

while the Big Bang nucleosynthesis estimate is $\Omega_B h^2 = 0.019 \pm 0.002$. Substantial dark (unclustered) energy is inferred:

$$\Omega_{\rm DE} \approx 0.72 \pm 0.015\,.$$

What is most relevant for us is that this universe was apparently born from a burst of rapid expansion, inflation, during which quantum noise was stretched to astrophysical size seeding cosmic structure. This is exactly the phenomenon we want to address in part of these lectures.

2.2 The early, radiation-dominated universe

In any case, at present, matter outweighs radiation by a wide margin. However, since the energy density in matter decreases as a^{-3}, and that in radiation as a^{-4} (the extra factor due to the redshifting of the energy of relativistic particles), at early times the universe was radiation dominated—indeed the calculations of primordial nucleosynthesis provide excellent evidence for this. Denoting the epoch of matter and radiation equality by subscript 'EQ,' and using $T_0 = 2.73\,\mathrm{K}$, it follows that

$$a_{\rm EQ} = 4.18 \times 10^{-5}\,(\Omega_0 h^2)^{-1}\,; \qquad T_{\rm EQ} = 5.62(\Omega_0 h^2)\,\mathrm{eV}\,; \tag{10}$$

$$t_{\rm EQ} = 4.17 \times 10^{10}(\Omega_0 h^2)^{-2}\mathrm{s}\,. \tag{11}$$

At early times the expansion rate and age of the universe were determined by the temperature of the universe and the number of relativistic degrees of freedom:

$$\rho_{\rm rad} = g_*(T)\frac{\pi^2 T^4}{30}; \qquad H \simeq 1.67 g_*^{1/2} T^2/m_{\rm Pl}; \tag{12}$$

$$\Rightarrow a \propto t^{1/2}; \qquad t \simeq 2.42 \times 10^{-6} g_*^{-1/2}(T/\,\mathrm{GeV})^{-2}\,\mathrm{s}\,; \tag{13}$$

where $g_*(T)$ counts the number of ultra-relativistic degrees of freedom ($\approx$ the sum of the internal degrees of freedom of particle species much less massive than the temperature) and $m_{\rm Pl} \equiv G^{-1/2} = 1.22 \times 10^{19}\,\mathrm{GeV}$ is the Planck mass. For example, at the epoch of nucleosynthesis, $g_* = 10.75$ assuming three, light ($\ll$ MeV) neutrino species; taking into account all the species in the Standard Model, $g_* = 106.75$ at temperatures much greater than $300\,\mathrm{GeV}$.

A quantity of importance related to g_* is the entropy density in relativistic particles,

$$s = \frac{\rho + p}{T} = \frac{2\pi^2}{45} g_* T^3,$$

and the entropy per co-moving volume,

$$S \;\propto\; a^3 s \;\propto\; g_* a^3 T^3.$$

By a wide margin most of the entropy in the universe exists in the radiation bath. The entropy density is proportional to the number density of relativistic particles. At present, the relativistic particle species are the photons and neutrinos, and the entropy density is a factor of 7.04 times the photon-number density: $n_\gamma = 413\,\mathrm{cm}^{-3}$ and $s = 2905\,\mathrm{cm}^{-3}$.

In thermal equilibrium—which provides a good description of most of the history of the universe—the entropy per co-moving volume S remains constant. This fact is very useful. First, it implies that the temperature and scale factor are related by

$$T \propto g_*^{-1/3} a^{-1}, \tag{14}$$

which for $g_* =$ const leads to the familiar $T \propto a^{-1}$.

Second, it provides a way of quantifying the net baryon number (or any other particle number) per co-moving volume:

$$N_B \equiv R^3 n_B = \frac{n_B}{s} \simeq (4-7) \times 10^{-11}. \tag{15}$$

The baryon number of the universe tells us two things: (1) the entropy per particle in the universe is extremely high, about 10^{10} or so compared to about 10^{-2} in the Sun and a few in the core of a newly formed neutron star. (2) The asymmetry between matter and antimatter is very small, about 10^{-10}, since at early times quarks and antiquarks were roughly as abundant as photons. One of the great successes of particle cosmology is baryogenesis, the idea that B, C, and CP violating interactions occurring out-of-equilibrium early on allow the universe to develop a net baryon number of this magnitude. Finally, the

constancy of the entropy per co-moving volume allows us to characterize the size of co-moving volume corresponding to our present Hubble volume in a very physical way: by the entropy it contains,

$$S_U = \frac{4\pi}{3} H_0^{-3} s \simeq 10^{90}. \tag{16}$$

The standard cosmology is tested back to times as early as about 0.01 s; it is only natural to ask how far back one can sensibly extrapolate. Since the fundamental particles of Nature are point-like quarks and leptons whose interactions are perturbatively weak at energies much greater than 1 GeV, one can imagine extrapolating as far back as the epoch where general relativity becomes suspect, i.e., where quantum gravitational effects are likely to be important: the Planck epoch, $t \sim 10^{-43}$s and $T \sim 10^{19}$ GeV. Of course, at present, our firm understanding of the elementary particles and their interactions only extends to energies of the order of 100 GeV, which corresponds to a time of the order of 10^{-11}s or so. We can be relatively certain that at a temperature of 100–200 MeV ($t \sim 10^{-5}$s) there was a transition (likely a second-order phase transition) from quark/gluon plasma to very hot hadronic matter, and that some kind of phase transition associated with the symmetry breakdown of the electroweak theory took place at a temperature of the order of 300 GeV ($t \sim 10^{-11}$s).

2.3 The concept of particle horizon

In spite of the fact that the universe was vanishingly small at early times, the rapid expansion precluded causal contact from being established throughout. Photons travel on null paths characterized by $dr = dt/a(t)$; the physical distance that a photon could have travelled since the bang until time t, the distance to the particle horizon, is

$$\begin{aligned} R_H(t) &= a(t) \int_0^t \frac{dt'}{a(t')} \\ &= \frac{t}{(1-n)} = n \frac{H^{-1}}{(1-n)} \sim H^{-1} \qquad \text{for } a(t) \propto t^n, \;\; n < 1. \end{aligned} \tag{17}$$

Using the conformal time $d\tau = dt/a$, the particle horizon becomes

$$R_H(t) = a(\tau) \int_{\tau_0}^{\tau} d\tau, \tag{18}$$

where τ_0 indicates the conformal time corresponding to $t = 0$. Note, in the standard cosmology the distance to the horizon is finite, and up to numerical factors, equal to the age of the universe or the Hubble radius, H^{-1}. For this reason, we shall use horizon and Hubble radius interchangeably[1].

Note also that a physical length scale λ is within the horizon if $\lambda < R_H \sim H^{-1}$. Since we can identify the length scale λ with its wavenumber k, $\lambda = 2\pi a/k$, we shall have the following rule

$$\begin{aligned} \frac{k}{aH} &\ll 1 \Longrightarrow \text{SCALE } \lambda \text{ OUTSIDE THE HORIZON} \\ \frac{k}{aH} &\gg 1 \Longrightarrow \text{SCALE } \lambda \text{ WITHIN THE HORIZON} \end{aligned}$$

[1] As we shall see, in inflationary models the horizon and Hubble radius are not roughly equal as the horizon distance grows exponentially relative to the Hubble radius; in fact, at the end of inflation they differ by e^N, where N is the number of e-folds of inflation. However, we shall slip and use "horizon" and "Hubble radius" interchangeably, though we shall always mean Hubble radius.

3 The shortcomings of the standard Big Bang theory

By now the shortcomings of standard cosmology are well appreciated: the horizon or large-scale smoothness problem; the small-scale inhomogeneity problem (origin of density perturbations); and the flatness or oldness problem. we shall briefly review only the horizon problem here here.

3.1 The horizon problem

According to standard cosmology, photons decoupled from the rest of the components (electrons and baryons) at a temperature of the order of 0.3 eV. This corresponds to the so-called surface of 'last scattering' at a redshift of about 1100 and an age of about $180000\,(\Omega_0 h^2)^{-1/2}\mathrm{yr}$. From the epoch of last scattering onwards, photons free-stream and reach us basically untouched. Detecting primordial photons is therefore equivalent to take a picture of the universe when the latter was about 300 000 years old. The spectrum of the cosmic background radiation (CBR) is consistent with that of a black body at temperature 2.73 K over more than three decades in wavelength.

The most accurate measurement of the temperature and spectrum is that by the WMAP5 instrument on the COBE satellite which determined its temperature to be $2.726 \pm 0.01\,\mathrm{K}$ [9]. The length corresponding to our present Hubble radius (which is approximately the radius of our observable universe) at the time of last scatteringwas

$$\lambda_H(t_{\rm LS}) = R_H(t_0)\left(\frac{a_{\rm LS}}{a_0}\right) = R_H(t_0)\left(\frac{T_0}{T_{\rm LS}}\right).$$

On the other hand, during the matter-dominated period, the Hubble length decreased with a different law

$$H^2 \propto \rho_M \propto a^{-3} \propto T^3.$$

At last-scattering

$$H_{LS}^{-1} = R_H(t_0)\left(\frac{T_{LS}}{T_0}\right)^{-3/2} \ll R_H(t_0).$$

The length corresponding to our present Hubble radius was much larger that the horizon at that time. This can be by shown comparing the volumes corresponding to these two scales

$$\frac{\lambda_H^3(T_{LS})}{H_{LS}^{-3}} = \left(\frac{T_0}{T_{LS}}\right)^{-\frac{3}{2}} \approx 10^6. \tag{19}$$

There were $\sim 10^6$ casually disconnected regions within the volume that now corresponds to our horizon! It is difficult to come up with a process other than an early hot and dense phase in the history of the universe that would lead to a precise black body for a bath of photons which were causally disconnected the last time they interacted with the surrounding plasma.

The horizon problem is well represented by Fig. 1 where the solid line indicates the horizon scale and the dashed line any generic physical length scale λ. Suppose, indeed, that λ indicates the distance between two photons we detect today. From Eq. (19) we discover that at the time of emission (last-scattering) the two photons could not talk to each other, the dashed line is above the solid line. There is another aspect of the horizon problem which is related to the problem of initial conditions for the cosmological perturbations. We have every indication that the universe at early times, say $t \ll 300\,000\,\mathrm{yr}$, was very homogeneous; however, today inhomogeneity (or structure) is ubiquitous: stars ($\delta\rho/\rho \sim 10^{30}$), galaxies ($\delta\rho/\rho \sim 10^5$), clusters of galaxies ($\delta\rho/\rho \sim 10$—10^3), superclusters, or "clusters of clusters" ($\delta\rho/\rho \sim 1$), voids ($\delta\rho/\rho \sim -1$), great walls, and so on. For some twenty-five years standard cosmology has provided a general framework for understanding this picture. Once the universe becomes matter dominated (around 1000 yr after the bang) primeval density inhomogeneities ($\delta\rho/\rho \sim 10^{-5}$) are amplified by gravity and grow into the structure we see today [2]. The existence of density inhomogeneities

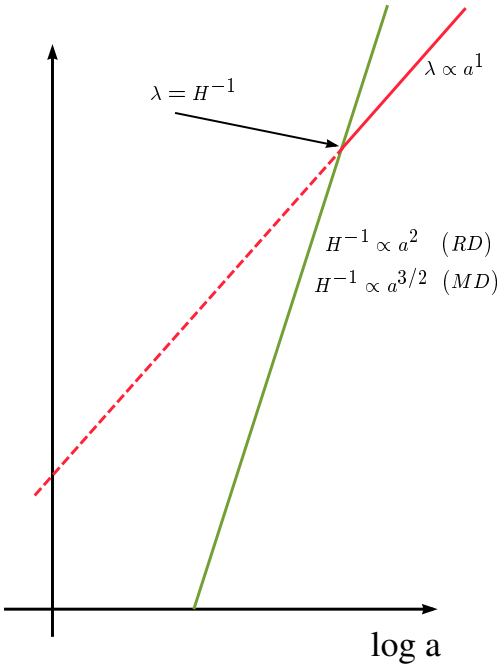


Fig. 1: The horizon scale (solid line) and a physical scale λ (dashed line) as function of the scale factor a

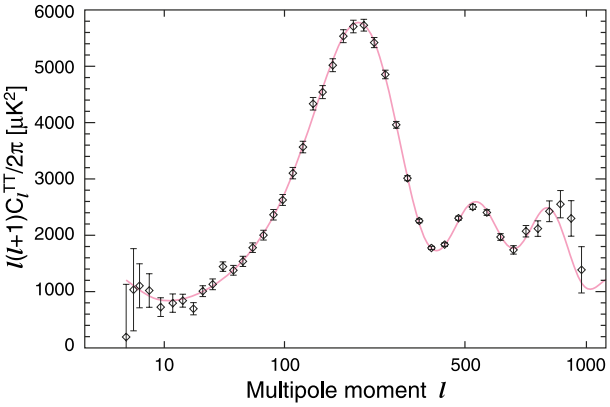


Fig. 2: The CMBR anisotropy as function of ℓ (from Ref. [9])

has another important consequence: fluctuations in the temperature of the CMB radiation of a similar amplitude. The temperature difference measured between two points separated by a large angle ($\gtrsim 1^\circ$) arises due to a very simple physical effect: the difference in the gravitational potential between the two points on the last scatteringsurface, which in turn is related to the density perturbation, determines the temperature anisotropy on the angular scale subtended by that length scale,

$$\left(\frac{\delta T}{T}\right)_\theta \approx \left(\frac{\delta \rho}{\rho}\right)_\lambda , \tag{20}$$

where the scale $\lambda \sim 100 h^{-1}\,\mathrm{Mpc}(\theta/\mathrm{deg})$ subtends an angle θ on the last-scattering surface. This is known as the Sachs–Wolfe effect [11, 12]. We shall come back to this piece of physics. The temperature

anisotropy is commonly expanded in spherical harmonics

$$\frac{\Delta T}{T}(x_0,\tau_0,\mathbf{n}) = \sum_{\ell m} a_{\ell,m}(x_0) Y_{\ell m}(\mathbf{n}), \tag{21}$$

where x_0 and τ_0 are our position and the preset time, respectively, $\mathbf{n}$ is the direction of observation, ℓ's are the different multipoles and[2]

$$\langle a_{\ell m} a^*_{\ell' m'}\rangle = \delta_{\ell,\ell'}\delta_{m,m'} C_\ell, \tag{22}$$

where the deltas are due to the fact that the process that created the anisotropy is statistically isotropic. The C_ℓ's are the so-called CMB power spectrum. For homogeneity and isotropy, the C_ℓ's are neither a function of x_0, nor of m. The two-point correlation function is related to the C_l's in the following way

$$\begin{aligned}\left\langle \frac{\delta T(\mathbf{n})}{T}\frac{\delta T(\mathbf{n}')}{T}\right\rangle &= \sum_{\ell\ell' mm'} \langle a_{\ell m} a^*_{\ell' m'}\rangle Y_{\ell m}(\mathbf{n}) Y^*_{\ell' m'}(\mathbf{n}') \\ &= \sum_\ell C_\ell \sum_m Y_{\ell m}(\mathbf{n}) Y^*_{\ell m}(\mathbf{n}') = \frac{1}{4\pi}\sum_\ell (2\ell+1) C_\ell P_\ell(\mu = \mathbf{n}\cdot\mathbf{n}')\end{aligned} \tag{23}$$

where we have used the addition theorem for the spherical harmonics, and P_ℓ is the Legendre polynom of order ℓ. In expression (23) the expectation value is an ensemble average. It can be regarded as an average over the possible observer positions, but not in general as an average over the single sky we observe, because of the cosmic variance[3]. WMAP5 data are given in Fig. 2.

Let us now consider the last scatteringsurface. In co-moving coordinates the latter is 'far' from us a distance equal to

$$\int_{t_{\rm LS}}^{t_0} \frac{dt}{a} = \int_{\tau_{\rm LS}}^{\tau_0} d\tau = (\tau_0 - \tau_{\rm LS})\,. \tag{24}$$

A given co-moving scale λ is therefore projected on the last scatteringsurface sky on an angular scale

$$\theta \simeq \frac{\lambda}{(\tau_0 - \tau_{\rm LS})}, \tag{25}$$

where we have neglected tiny curvature effects. Consider now that the scale λ is of the order of the co-moving sound horizon at the time of last-scattering, $\lambda \sim c_s \tau_{\rm LS}$, where $c_s \simeq 1/\sqrt{3}$ is the sound velocity at which photons propagate in the plasma at the last-scattering. This corresponds to an angle

$$\theta \simeq c_s \frac{\tau_{\rm LS}}{(\tau_0 - \tau_{\rm LS})} \simeq c_s \frac{\tau_{\rm LS}}{\tau_0}, \tag{26}$$

where the last passage has been performed knowing that $\tau_0 \gg \tau_{\rm LS}$. Since the universe is matter-dominated from the time of last scatteringonwards, the scale factor has the following behaviour: $a \sim T^{-1} \sim t^{2/3} \sim \tau^2$. The angle $\theta_{\rm HOR}$ subtended by the sound horizon on the last-scattering surface then becomes

$$\theta_{\rm HOR} \simeq c_s \left(\frac{T_0}{T_{\rm LS}}\right)^{1/2} \sim 1^\circ, \tag{27}$$

where we have used $T_{\rm LS} \simeq 0.3$ eV and $T_0 \sim 10^{-13}$ GeV. This corresponds to a multipole $\ell_{\rm HOR}$

[2] An alternative definition is $C_\ell = \langle |a_{\ell m}|^2\rangle = \frac{1}{2\ell+1}\sum_{m=-\ell}^{\ell} |a_{\ell m}|^2$.

[3] The usual hypothesis is that we observe a typical realization of the ensemble. This means that we expect the difference between the observed values $|a_{\ell m}|^2$ and the ensemble averages C_ℓ to be of the order of the mean-square deviation of $|a_{\ell m}|^2$ from C_ℓ. The latter is called cosmic variance and, because we are dealing with a Gaussian distribution, it is equal to $2C_\ell$ for each multipole ℓ. For a single ℓ, averaging over the $(2\ell+1)$ values of m reduces the cosmic variance by a factor $(2\ell+1)$, but it remains a serious limitation for low multipoles.

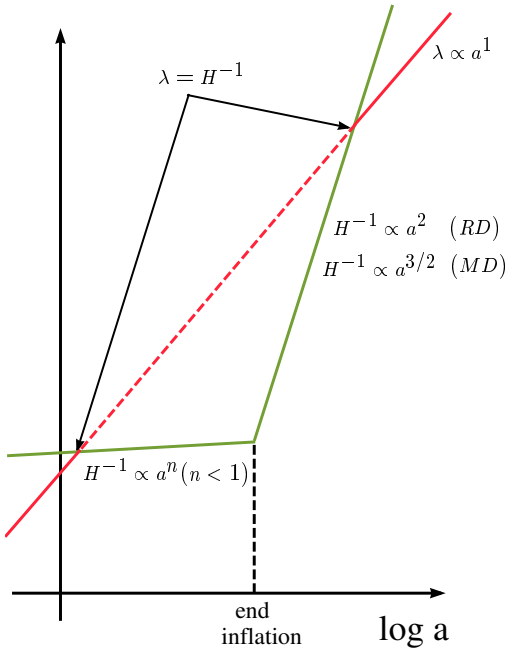


Fig. 3: The behaviour of a generic scale λ and the horizon scale H^{-1} in the standard inflationary model

$$\ell_{\rm HOR} = \frac{\pi}{\theta_{\rm HOR}} \simeq 200\,. \tag{28}$$

From these estimates we conclude that two photons which on the last scatteringsurface were separated by an angle larger than $\theta_{\rm HOR}$, corresponding to multipoles smaller than $\ell_{\rm HOR} \sim 200$, were not in causal contact. On the other hand, from Fig. 2 it is clear that small anisotropies, of the *same* order of magnitude $\delta T/T \sim 10^{-5}$ are present at $\ell \ll 200$. We conclude that one of the striking features of the CMB fluctuations is that they appear to be non-causal. Photons at the last scatteringsurface which were causally disconnected have the same small anisotropies! The existence of particle horizons in the standard cosmology precludes explaining the smoothness as a result of microphysical events: the horizon at decoupling, the last time one could imagine temperature fluctuations being smoothed by particle interactions, corresponds to an angular scale on the sky of about 1°, which precludes temperature variations on larger scales from being erased. To account for the small-scale lumpiness of the universe today, density perturbations with horizon-crossing amplitudes of 10^{-5} on scales of $1\,{\rm Mpc}$ to $10^4\,{\rm Mpc}$ or so are required.

As can be seen in Fig. 1, in the standard cosmology the physical size of a perturbation, which grows as the scale factor, begins larger than the horizon and, relatively late in the history of the universe, crosses inside the horizon. This precludes a causal microphysical explanation for the origin of the required density perturbations.

From the considerations made so far, it appears that solving the horizon problem of the standard Big Bang theory requires that the universe go through a primordial period during which the physical scales λ evolve faster than the horizon scale H^{-1}.

If there is period during which physical length scales grow faster than H^{-1}, length scales λ which are within the horizon today, $\lambda < H^{-1}$ (such as the distance between two detected photons) and were outside the horizon for some period, $\lambda > H^{-1}$ (for instance at the time of last scatteringwhen the two

photons were emitted), had a chance to be within the horizon at some primordial epoch, $\lambda < H^{-1}$ again, see Fig. 3. If this happens, the homogeneity and the isotropy of the CMB can easily be explained: photons that we receive today and were emitted from the last scattering surface from causally disconnected regions have the same temperature because they had a chance to 'talk' to each other at some primordial stage of the evolution of the universe.

The second condition can easily be expressed as a condition on the scale factor a. Since a given scale λ scales like $\lambda \sim a$ and $H^{-1} = a/\dot{a}$, we need to impose that there is a period during which

$$\left(\frac{\lambda}{H^{-1}}\right)^{\cdot} = \ddot{a} > 0\,.$$

We can therefore introduce the following rigorous definition: an inflationary stage is a period of the universe during which the latter accelerates

$$\text{INFLATION} \quad \Longleftrightarrow \quad \ddot{a} > 0.$$

Comment: Let us stress that during such an accelerating phase the universe expands *adiabatically*. This means that during inflation one can exploit the usual FRW equations (3) and (5). It must be clear therefore that the non-adiabaticity condition is satisfied not during inflation, but during the phase transition between the end of inflation and the beginning of the radiation-dominated phase. At this transition phase a large entropy is generated under the form of relativistic degrees of freedom: the Big Bang has taken place.

4 The standard inflationary universe

From the previous section we have learned that an accelerating stage during the primordial phases of the evolution of the universe might be able to solve the horizon problem. From Eq. (5) we learn that

$$\ddot{a} > 0 \Longleftrightarrow (\rho + 3P) < 0\,.$$

An accelerating period is obtainable only if the overall pressure p of the universe is negative: $p < -\rho/3$. Neither a radiation-dominated phase nor a matter-dominated phase (for which $p = \rho/3$ and $p = 0$, respectively) satisfy such a condition. Let us postpone for the time being the problem of finding a 'candidate' able to provide the condition $P < -\rho/3$. For sure, inflation is a phase of the history of the universe occurring before the era of nucleosynthesis ($t \approx 1$ s, $T \approx 1$ MeV) during which the light elements abundances were formed. This is because nucleosynthesis is the earliest epoch from which we have experimental data and they are in agreement with the predictions of the standard Big Bang theory. However, the thermal history of the universe before the epoch of nucleosynthesis is unknown.

In order to study the properties of the period of inflation, we assume the extreme condition $p = -\rho$ which considerably simplifies the analysis. A period of the universe during which $P = -\rho$ is called the de Sitter stage. By inspecting Eqs. (3) and (4), we learn that during the de Sitter phase

$$\begin{aligned} \rho &= \text{constant}\,, \\ H_I &= \text{constant}\,, \end{aligned}$$

where we have indicated by H_I the value of the Hubble rate during inflation. Correspondingly, solving Eq. (3) gives

$$a = a_i\, e^{H_I(t-t_i)}, \tag{29}$$

where t_i denotes the time at which inflation starts. Let us now see how such a period of exponential expansion takes care of the shortcomings of the standard Big Bang Theory[4].

4.1 Inflation and the horizon problem

During the inflationary (de Sitter) epoch the horizon scale H^{-1} is constant. If inflation lasts long enough, all the physical scales that have left the horizon during the radiation-dominated or matter-dominated phase can re-enter the horizon in the past: this is because such scales are exponentially reduced. As we have seen in the previous section, this explains both the problem of the homogeneity of CMB and the initial condition problem of small cosmological perturbations. Once the physical length is within the horizon, microphysics can act, the universe can be made approximately homogeneous and the primeval inhomogeneities can be created.

Let us see how long inflation must be sustained in order to solve the horizon problem. Let t_i and t_f be, respectively, the time of beginning and end of inflation. We can define the corresponding number of e-foldings N

$$N = \ln\left[H_I(t_e - t_i)\right]. \tag{30}$$

A necessary condition to solve the horizon problem is that the largest scale we observe today, the present horizon H_0^{-1}, was reduced during inflation to a value $\lambda_{H_0}(t_i)$ smaller than the value of horizon length H_I^{-1} during inflation. This gives

$$\lambda_{H_0}(t_i) = H_0^{-1}\left(\frac{a_{t_f}}{a_{t_0}}\right)\left(\frac{a_{t_i}}{a_{t_f}}\right) = H_0^{-1}\left(\frac{T_0}{T_f}\right)e^{-N} \lesssim H_I^{-1},$$

where we have neglected for simplicity the short period of matter-domination and we have called T_f the temperature at the end of inflation (to be indentified with the reheating temperature T_{RH} at the beginning of the radiation-dominated phase after inflation, see later). We get

$$N \gtrsim \ln\left(\frac{T_0}{H_0}\right) - \ln\left(\frac{T_f}{H_I}\right) \approx 67 + \ln\left(\frac{T_f}{H_I}\right).$$

Apart from the logarithmic dependence, we obtain $N \gtrsim 70$.

4.2 A prediction of inflation

Since during inflation the Hubble rate is constant

$$\Omega - 1 = \frac{k}{a^2H^2} \propto \frac{1}{a^2}\,.$$

On the other hand it is easy to show that to reproduce a value of $(\Omega_0 - 1)$ of order of unity today, the initial value of $(\Omega - 1)$ at the beginning of the radiation-dominated phase must be $|\Omega - 1| \sim 10^{-60}$. Since we identify the beginning of the radiation-dominated phase with the beginning of inflation, we require

$$|\Omega - 1|_{t=t_f} \sim 10^{-60}.$$

During inflation

$$\frac{|\Omega - 1|_{t=t_f}}{|\Omega - 1|_{t=t_i}} = \left(\frac{a_i}{a_f}\right)^2 = e^{-2N}. \tag{31}$$

Taking $|\Omega - 1|_{t=t_i}$ of order unity, it is enough to require that $N \approx 70$. However, IF the period of inflation lasts longer than 70 e-foldings the present-day value of Ω_0 will be equal to unity with great precision. One can say that a generic prediction of inflation is that

[4]Despite the fact that the growth of the scale factor is exponential and the expansion is *superluminal*, this is not in contradiction with what is dictated by relativity. Indeed, it is the spacetime itself which is progating so fast and not a light signal in it.

$$\text{INFLATION} \implies \Omega_0 = 1.$$

This statement, however, must be taken *cum grano salis* and properly specified. Inflation does not change the global geometric properties of the space-time. If the universe is open or closed, it will always remain flat or closed, independently from inflation. What inflation does is to magnify the radius of curvature $R_{\rm curv}$ defined in Eq. (9) so that locally the universe is flat with a great precision. As we shall see, the current data on the CMB anisotropies confirm this prediction.

4.3 Inflation and the inflaton

In the previous subsections we have described the various advantages of having a period of accelerating phase. The latter required $P < -\rho/3$. Now, we would like to show that this condition can be attained by means of a simple scalar field. We shall call this field the *inflaton* ϕ.

The action of the inflaton field reads

$$S = \int d^4x\,\sqrt{-g}\,\mathcal{L} = \int d^4x\,\sqrt{-g}\,\left[\frac{1}{2}\partial_\mu\phi\partial^\mu\phi + V(\phi)\right], \tag{32}$$

where $\sqrt{-g} = a^3$ for the FRW metric (1). From the Euler–Lagrange equations

$$\partial^\mu \frac{\delta(\sqrt{-g}\mathcal{L})}{\delta\,\partial^\mu\phi} - \frac{\delta(\sqrt{-g}\mathcal{L})}{\delta\phi} = 0\,, \tag{33}$$

we obtain

$$\ddot{\phi} + 3H\dot{\phi} - \frac{\nabla^2\phi}{a^2} + V'(\phi) = 0\,, \tag{34}$$

where $V'(\phi) = (dV(\phi)/d\phi)$. Note, in particular, the appearance of the friction term $3H\dot{\phi}$: a scalar field rolling down its potential suffers a friction due to the expansion of the universe.

We can write the energy momentum tensor of the scalar field

$$T_{\mu\nu} = \partial_\mu\phi\partial_\nu\phi - g_{\mu\nu}\,\mathcal{L}\,.$$

The corresponding energy density ρ_ϕ and pressure density P_ϕ are

$$T_{00} = \rho_\phi = \frac{\dot{\phi}^2}{2} + V(\phi) + \frac{(\nabla\phi)^2}{2a^2}, \tag{35}$$

$$T_{ii} = P_\phi = \frac{\dot{\phi}^2}{2} - V(\phi) - \frac{(\nabla\phi)^2}{6a^2}\,. \tag{36}$$

Note that, if the gradient term were dominant, we would obtain $P_\phi = -\frac{\rho_\phi}{3}$, not enough to drive inflation. We can now split the inflaton field in

$$\phi(t) = \phi_0(t) + \delta\phi(\mathbf{x}, t)\,,$$

where ϕ_0 is the 'classical' (infinite wavelength) field, that is the expectation value of the inflaton field on the initial isotropic and homogeneous state, while $\delta\phi(\mathbf{x}, t)$ represents the quantum fluctuations around ϕ_0. In this section, we shall be concerned only with the evolution of the classical field ϕ_0. The next section will be devoted to the crucial issue of the evolution of quantum perturbations during inflation. This separation is justified by the fact that quantum fluctuations are much smaller than the classical value and therefore negligible when looking at the classical evolution. Not to be overwhelmed by the notation,

we shall indicate the classical value of the inflaton field by ϕ from now on. The energy momentum tensor becomes

$$T_{00} = \rho_\phi = \frac{\dot{\phi}^2}{2} + V(\phi) \tag{37}$$

$$T_{ii} = P_\phi = \frac{\dot{\phi}^2}{2} - V(\phi). \tag{38}$$

If

$$V(\phi) \gg \dot{\phi}^2$$

we obtain the following condition

$$P_\phi \simeq -\rho_\phi \,.$$

From this simple calculation, we realize that a scalar field whose energy is dominant in the universe and whose potential energy dominates over the kinetic term gives inflation. Inflation is driven by the vacuum energy of the inflaton field.

4.4 Slow-roll conditions

Let us now quantify better under which circumstances a scalar field may give rise to a period of inflation. The equation of motion of the field is

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0\,. \tag{39}$$

If we require that $\dot{\phi}^2 \ll V(\phi)$, the scalar field is slowly rolling down its potential. This is the reason why such a period is called *slow-roll*. We may also expect that since the potential is flat, $\ddot{\phi}$ is negligible as well. We shall assume that this is true and we shall quantify this condition soon. The FRW equation (3) becomes

$$H^2 \simeq \frac{8\pi G}{3}\, V(\phi), \tag{40}$$

where we have assumed that the inflaton field dominates the energy density of the universe. The new equation of motion becomes

$$3H\dot{\phi} = -V'(\phi) \tag{41}$$

which gives $\dot{\phi}$ as a function of $V'(\phi)$. Using Eq. (41) slow-roll conditions then require

$$\dot{\phi}^2 \ll V(\phi) \Longrightarrow \frac{(V')^2}{V} \ll H^2$$

and

$$\ddot{\phi} \ll 3H\dot{\phi} \Longrightarrow V'' \ll H^2.$$

It is now useful to define the slow-roll parameters ϵ and η in the following way

$$\begin{aligned}
\epsilon &= -\frac{\dot{H}}{H^2} = 4\pi G \frac{\dot{\phi}^2}{H^2} = \frac{1}{16\pi G}\left(\frac{V'}{V}\right)^2, \\
\eta &= \frac{1}{8\pi G}\left(\frac{V''}{V}\right) = \frac{1}{3}\frac{V''}{H^2}, \\
\delta &= \eta - \epsilon = -\frac{\ddot{\phi}}{H\dot{\phi}}\,.
\end{aligned}$$

It might be useful to have the same parameters expressed in terms of conformal time

$$\begin{aligned}\epsilon &= 1-\frac{\mathcal{H}'}{\mathcal{H}^2}=4\pi G\frac{\phi'^2}{\mathcal{H}^2}\\ \delta &= \eta-\epsilon=1-\frac{\phi''}{\mathcal{H}\phi'}\,.\end{aligned}$$

The parameter ϵ quantifies how much the Hubble rate H changes with time during inflation. Notice that, since

$$\frac{\ddot{a}}{a}=\dot{H}+H^2=(1-\epsilon)\,H^2,$$

inflation can be attained only if $\epsilon<1$:

$$\text{INFLATION} \iff \epsilon<1.$$

As soon as this condition fails, inflation ends. In general, slow-roll inflation is attained if $\epsilon \ll 1$ and $|\eta| \ll 1$. During inflation the slow-roll parameters ϵ and η can be considered to be approximately constant since the potential $V(\phi)$ is very flat.

Comment: In the following, we shall work at *first-order* perturbation in the slow-roll parameters, that is we shall take only the first power of them. Since, using their definition, it is easy to see that $\dot{\epsilon},\dot{\eta}=\mathcal{O}\left(\epsilon^2,\eta^2\right)$, this amounts to saying that we shall treat the slow-roll parameters as constant in time.

Within these approximations, it is easy to compute the number of e-foldings between the beginning and the end of inflation. If we indicate by ϕ_i and ϕ_f the values of the inflaton field at the beginning and at the end of inflation, respectively, we find that the *total* number of e-foldings is

$$\begin{aligned}N &\equiv \int_{t_i}^{t_f} H\,dt\\ &\simeq H\int_{\phi_i}^{\phi_f}\frac{d\phi}{\dot{\phi}}\\ &\simeq -3H^2\int_{\phi_i}^{\phi_f}\frac{d\phi}{V'}\\ &\simeq -8\pi G\int_{\phi_i}^{\phi_f}\frac{V}{V'}\,d\phi\,.\end{aligned} \tag{42}$$

We may also compute the number of e-foldings ΔN which are left to go to the end of inflation

$$\Delta N\simeq 8\pi G\int_{\phi_f}^{\phi_{\Delta N}}\frac{V}{V'}\,d\phi, \tag{43}$$

where $\phi_{\Delta N}$ is the value of the inflaton field when there are ΔN e-foldings to the end of inflation.

1. Comment: According to the criterion given in Subsection 2.4, a given scale length $\lambda=a/k$ leaves the horizon when $k=aH_k$ where H_k is the the value of the Hubble rate at that time. One can easily compute the rate of change of H_k^2 as a function of k

$$\frac{d\ln H_k^2}{d\ln k}=\left(\frac{d\ln H_k^2}{dt}\right)\left(\frac{dt}{d\ln a}\right)\left(\frac{d\ln a}{d\ln k}\right)=2\frac{\dot{H}}{H}\times\frac{1}{H}\times 1=2\frac{\dot{H}}{H^2}=-2\epsilon. \tag{44}$$

2. Comment: Take a given physical scale λ today which crossed the horizon scale during inflation. This happened when

$$\lambda \left(\frac{a_f}{a_0}\right) e^{-\Delta N_\lambda} = \lambda \left(\frac{T_0}{T_f}\right) e^{-\Delta N_\lambda} = H_I^{-1}$$

where ΔN_λ indicates the number of e-foldings from the time the scale crossed the horizon during inflation and the end of inflation. This relation gives a way to determine the number of e-foldings to the end of inflation corresponding to a given scale

$$\Delta N_\lambda \simeq 65 + \ln\left(\frac{\lambda}{3000\ \mathrm{Mpc}}\right) + 2\ln\left(\frac{V^{1/4}}{10^{14}\ \mathrm{GeV}}\right) - \ln\left(\frac{T_f}{10^{10}\ \mathrm{GeV}}\right).$$

Scales relevant for the CMB anisotropies correspond to $\Delta N \sim$60.

Inflation ended when the potential energy associated with the inflaton field became smaller than the kinetic energy of the field. By that time, any pre-inflation entropy in the universe had been inflated away, and the energy of the universe was entirely in the form of coherent oscillations of the inflaton condensate around the minimum of its potential. The universe may be said to be frozen after the end of inflation. We know that somehow the low-entropy cold universe dominated by the energy of coherent motion of the ϕ field must be transformed into a high-entropy hot universe dominated by radiation. The process by which the energy of the inflaton field is transferred from the inflaton field to radiation has been dubbed *reheating*. In the theory of reheating, the simplest way to envisage this process is if the comoving energy density in the zero mode of the inflaton decays into normal particles, which then scatter and thermalize to form a thermal background. It is usually assumed that the decay width of this process is the same as the decay width of a free inflaton field.

Of particular interest is a quantity usually known as the reheat temperature, denoted as T_{RH}[5]. The reheat temperature is calculated by assuming an instantaneous conversion of the energy density in the inflaton field into radiation when the decay width of the inflaton energy, Γ_ϕ, is equal to H, the expansion rate of the universe.

The reheat temperature is calculated quite easily. After inflation the inflaton field executes coherent oscillations about the minimum of the potential. Averaged over several oscillations, the coherent oscillation energy density redshifts as matter: $\rho_\phi \propto a^{-3}$, where a is the Robertson–Walker scale factor. If we denote as ρ_I and a_I the total inflaton energy density and the scale factor at the initiation of coherent oscillations, then the Hubble expansion rate as a function of a is

$$H^2(a) = \frac{8\pi}{3}\frac{\rho_I}{{m_{\mathrm{Pl}}}^2}\left(\frac{a_I}{a}\right)^3. \tag{45}$$

Equating $H(a)$ and Γ_ϕ leads to an expression for a_I/a. Now if we assume that all available coherent energy density is instantaneously converted into radiation at this value of a_I/a, we can find the reheat temperature by setting the coherent energy density, $\rho_\phi = \rho_I(a_I/a)^3$, equal to the radiation energy density, $\rho_R = (\pi^2/30)g_* T_{RH}^4$, where g_* is the effective number of relativistic degrees of freedom at temperature T_{RH}. The result is

$$T_{RH} = \left(\frac{90}{8\pi^3 g_*}\right)^{1/4}\sqrt{\Gamma_\phi m_{\mathrm{Pl}}} = 0.2\left(\frac{200}{g_*}\right)^{1/4}\sqrt{\Gamma_\phi m_{\mathrm{Pl}}}\,. \tag{46}$$

5 Inflation and the cosmological perturbations

As we have seen in the previous section, the early universe was made very nearly uniform by a primordial inflationary stage. However, the important caveat in that statement is the word 'nearly'. Our current understanding of the origin of structure in the universe is that it originated from small 'seed' perturbations,

[5] So far, we have indicated it by T_f.

which over time grew to become all of the structure we observe. Once the universe becomes matter dominated (around 1000 yrs after the bang) primeval density inhomogeneities ($\delta\rho/\rho \sim 10^{-5}$) are amplified by gravity and grow into the structure we see today [4]. The fact that a fluid of self-gravitating particles is unstable to the growth of small inhomogeneities was first pointed out by Jeans and is known as the Jeans instability. Furthermore, the existence of these inhomogeneities was confirmed by the COBE discovery of CMB anisotropies; the temperature anisotropies detected almost certainly owe their existence to primeval density inhomogeneities, since, as we have seen, causality precludes microphysical processes from producing anisotropies on angular scales larger than about 1°, the angular size of the horizon at last-scattering.

The growth of small matter inhomogeneities of wavelength smaller than the Hubble scale ($\lambda \lesssim H^{-1}$) is governed by a Newtonian equation:

$$\ddot{\delta}_{\mathbf{k}} + 2H\dot{\delta}_{\mathbf{k}} + v_s^2 \frac{k^2}{a^2}\delta_{\mathbf{k}} = 4\pi G\rho_M \delta_{\mathbf{k}}, \tag{47}$$

where $v_s^2 = \partial P/\partial\rho_M$ is the square of the speed of sound and we have expanded the perturbation to the matter density in plane waves

$$\frac{\delta\rho_m(\mathbf{x},t)}{\rho_m} = \frac{1}{(2\pi)^3}\int d^3k\, \delta_{\mathbf{k}}(t) e^{-i\mathbf{k}\cdot\mathbf{x}}. \tag{48}$$

Competition between the pressure term and the gravity term on the rhs of Eq. (47) determines whether or not pressure can counteract gravity: perturbations with wavenumber larger than the Jeans wavenumber, $k_J^2 = 4\pi G a^2 \rho_m / v_s^2$, are Jeans stable and just oscillate; perturbations with smaller wavenumber are Jeans unstable and can grow.

Let us discuss solutions to this equation under different circumstances. First, consider the Jeans problem, evolution of perturbations in a static fluid, i.e., $H = 0$. In this case Jeans unstable perturbations grow exponentially, $\delta_{\mathbf{k}} \propto \exp(t/\tau)$ where $\tau = 1/\sqrt{4G\pi\rho_M}$. Next, consider the growth of Jeans unstable perturbations in a matter-dominated universe, i.e., $H^2 = 8\pi G\rho_M/3$ and $a \propto t^{2/3}$. Because the expansion tends to pull particles away from one another, the growth is only power law, $\delta_{\mathbf{k}} \propto t^{2/3}$; i.e., at the same rate as the scale factor. Finally, consider a radiation-dominated universe. In this case, the expansion is so rapid that matter perturbations grow very slowly, as $\ln a$ in a radiation-dominated epoch. Therefore, perturbations may grow only in a matter-dominated period. Once a perturbation reaches an overdensity of order unity or larger it separates from the expansion, i.e., it becomes its own self-gravitating system and ceases to expand any further. In the process of virial relaxation, its size decreases by a factor of two—density increases by a factor of 8; thereafter, its density contrast grows as a^3 since the average matter density is decreasing as a^{-3}, though smaller scales could become Jeans unstable and collapse further to form smaller objects of higher density.

In order for structure formation to occur via gravitational instability, there must have been small pre-existing fluctuations on physical length scales when they crossed the Hubble radius in the radiation-dominated and matter-dominated eras. In the standard Big Bang model these small perturbations have to be put in by hand, because it is impossible to produce fluctuations on any length scale while it is larger than the horizon. Since the goal of cosmology is to understand the universe on the basis of physical laws, this appeal to initial conditions is unsatisfactory. The challenge is therefore to give an explanation to the small seed perturbations which allow the gravitational growth of the matter perturbations.

Our best guess for the origin of these perturbations is quantum fluctuations during an inflationary era in the early universe. Although originally introduced as a possible solution to the cosmological conundrums such as the horizon, flatness and entropy problems, by far the most useful property of inflation is that it generates spectra of both density perturbations and gravitational waves. These perturbations extend from extremely short scales to scales considerably in excess of the size of the observable universe.

During inflation the scale factor grows quasi-exponentially, while the Hubble radius remains almost constant. Consequently the wavelength of a quantum fluctuation— either in the scalar field whose

potential energy drives inflation or in the graviton field—soon exceeds the Hubble radius. The amplitude of the fluctuation therefore becomes 'frozen in'. This is quantum mechanics in action at macroscopic scales.

According to quantum field theory, empty space is not entirely empty. It is filled with quantum fluctuations of all types of physical fields. The fluctuations can be regarded as waves of physical fields with all possible wavelenghts, moving in all possible directions. If the values of these fields, averaged over some macroscopically large time, vanish then the space filled with these fields seems to us empty and can be called the vacuum.

In the exponentially expanding universe the vacuum structure is much more complicated. The wavelenghts of all vacuum fluctuations of the inflaton field ϕ grow exponentially in the expanding universe. When the wavelength of any particular fluctuation becomes greater than H^{-1}, this fluctuation stops propagating, and its amplitude freezes at some non-zero value $\delta\phi$ because of the large friction term $3H\dot{\phi}$ the equation of motion of the field ϕ. The amplitude of this fluctuation then remains almost unchanged for a very long time, whereas its wavelength grows exponentially. Therefore, the appearance of such frozen fluctuation is equivalent to the appearance of a classical field $\delta\phi$ that does not vanish after having averaged over some macroscopic interval of time. Because the vacuum contains fluctuations of all possible wavelengths, inflation leads to the creation of more and more new perturbations of the classical field with wavelength larger than the horizon scale.

Once inflation has ended, however, the Hubble radius increases faster than the scale factor, so the fluctuations eventually re-enter the Hubble radius during the radiation- or matter-dominated eras. The fluctuations that exit around 60 e-foldings or so before reheating re-enter with physical wavelengths in the range accessible to cosmological observations. These spectra provide a distinctive signature of inflation. They can be measured in a variety of different ways including the analysis of microwave background anisotropies.

Quantum fluctuations of the inflaton field are generated during inflation. Since gravity talks to any component of the universe, small fluctuations of the inflaton field are intimately related to fluctuations of the space-time metric, giving rise to perturbations of the curvature $\mathcal{R}$ (which will be defined in the following; the reader may loosely think of it as a gravitational potential). The wavelengths λ of these perturbations grow exponentially and leave the horizon soon when $\lambda > R_H$. On superhorizon scales, curvature fluctuations are frozen in and may be considered as classical. Finally, when the wavelength of these fluctuations re-enters the horizon, at some radiation- or matter-dominated epoch, the curvature (gravitational potential) perturbations of the space-time give rise to matter (and temperature) perturbations $\delta\rho$ via the Poisson equation. These fluctuations will then start growing, giving rise to the structures we observe today.

In summary, these are the key ingredients for understanding the observed structures in the universe within the inflationary scenario:

- Quantum fluctuations of the inflaton field are excited during inflation and stretched to cosmological scales. At the same time, being the inflaton fluctuations connected to the metric perturbations through Einstein's equations, ripples on the metric are also excited and stretched to cosmological scales.
- Gravity acts a messenger since it communicates the small seed perturbations to photons and baryons once a given wavelength becomes smaller than the horizon scale after inflation.

Let us now see how quantum fluctuations are generated during inflation. we shall proceed by steps. First, we shall consider the simplest problem of studying the quantum fluctuations of a generic scalar field during inflation: we shall learn how perturbations evolve as a function of time and compute their spectrum. Then—since a satisfactory description of the generation of quantum fluctuations has to take both the inflaton and the metric perturbations into account— we shall study the system composed

by quantum fluctuations of the inflaton field and quantum fluctuations of the metric.

6 Quantum fluctuations of a generic massless scalar field during inflation

Let us first see how the fluctuations of a generic scalar field χ, which is *not* the inflaton field, behave during inflation. To warm up we first consider a de Sitter epoch during which the Hubble rate is constant.

6.1 Quantum fluctuations of a generic massless scalar field during a de Sitter stage

We assume this field to be massless. The massive case will be analysed in the next subsection.

Expanding the scalar field χ in Fourier modes

$$\delta\chi(\mathbf{x},t) = \int \frac{d^3\mathbf{k}}{(2\pi)^{3/2}}\, e^{i\mathbf{k}\cdot\mathbf{x}}\, \delta\chi_{\mathbf{k}}(t),$$

we can write the equation for the fluctuations as

$$\delta\ddot{\chi}_{\mathbf{k}} + 3H\,\delta\dot{\chi}_{\mathbf{k}} + \frac{k^2}{a^2}\,\delta\chi_{\mathbf{k}} = 0\,. \tag{49}$$

Let us study the qualitative behaviour of the solution to Eq. (49).

- For wavelengths within the horizon, $\lambda \ll H^{-1}$, the corresponding wave-number satisfies the relation $k \gg a\,H$. In this regime, we can neglect the friction term $3H\,\delta\dot{\chi}_{\mathbf{k}}$ and Eq. (49) reduces to

$$\delta\ddot{\chi}_{\mathbf{k}} + \frac{k^2}{a^2}\,\delta\chi_{\mathbf{k}} = 0, \tag{50}$$

  which is basically the equation of motion of an harmonic oscillator. Of course, the frequency term k^2/a^2 depends upon time because the scale factor a grows exponentially. On the qualitative level, however, one expects that when the wavelength of the fluctuation is within the horizon, the fluctuation oscillates.
- For wavelengths above the horizon, $\lambda \gg H^{-1}$, the corresponding wave-number satisfies the relation $k \ll aH$ and the term k^2/a^2 can be safely neglected. Equation (49) reduces to

$$\delta\ddot{\chi}_{\mathbf{k}} + 3H\,\delta\dot{\chi}_{\mathbf{k}} = 0, \tag{51}$$

  which tells us that on superhorizon scales $\delta\chi_{\mathbf{k}}$ remains constant.

We have therefore the following picture: take a given fluctuation whose initial wavelength $\lambda \sim a/k$ is within the horizon. The fluctuations oscillate till the wavelength becomes of the order of the horizon scale. When the wavelength crosses the horizon, the fluctuation ceases to oscillate and gets frozen in.

Let us now study the evolution of the fluctuation in a more quantitative way. To do so, we perform the following redefinition

$$\delta\chi_{\mathbf{k}} = \frac{\delta\sigma_{\mathbf{k}}}{a}$$

and we work in conformal time $d\tau = dt/a$. For the time being, we solve the problem for a pure de Sitter expansion and we take the scale factor exponentially growing as $a \sim e^{Ht}$; the corresponding conformal factor reads (after choosing properly the integration constants)

$$a(\tau) = -\frac{1}{H\tau} \quad (\tau < 0).$$

In the following we shall also solve the problem in the case of quasi de Sitter expansion. The beginning of inflation coincides with some initial time $\tau_i \ll 0$. We find that Eq. (49) becomes

$$\delta\sigma_{\mathbf{k}}'' + \left(k^2 - \frac{a''}{a}\right)\delta\sigma_{\mathbf{k}} = 0. \tag{52}$$

We obtain an equation which is very 'close' to the equation for a Klein–Gordon scalar field in flat space-time, the only difference being a negative time-dependent mass term $-a''/a = -2/\tau^2$. Equation (52) can be obtained from an action of the type

$$\delta S_{\mathbf{k}} = \int d\tau \left[\frac{1}{2}\delta\sigma_{\mathbf{k}}'^2 - \frac{1}{2}\left(k^2 - \frac{a''}{a}\right)\delta\sigma_{\mathbf{k}}^2\right], \tag{53}$$

which is the canonical action for a simple harmonic oscillator with canonical commutation relations $\delta\sigma_{\mathbf{k}}^*\delta\sigma_{\mathbf{k}}' - \delta\sigma_{\mathbf{k}}\delta\sigma_{\mathbf{k}}^{*\prime} = -i$.

Let us study the behaviour of this equation on subhorizon and superhorizon scales. Since

$$\frac{k}{aH} = -k\,\tau\,,$$

on subhorizon scales $k^2 \gg a''/a$ Equation (52) reduces to

$$\delta\sigma_{\mathbf{k}}'' + k^2\,\delta\sigma_{\mathbf{k}} = 0\,,$$

whose solution is a plane wave

$$\delta\sigma_{\mathbf{k}} = \frac{e^{-ik\tau}}{\sqrt{2k}} \quad (k \gg aH)\,. \tag{54}$$

We find again that fluctuations with wavelength within the horizon oscillate exactly like in flat space-time. This does not come as a surprise. In the ultraviolet regime, that is for wavelengths much smaller than the horizon scale, one expects that approximating the space-time as flat is a good approximation.

On superhorizon scales, $k^2 \ll a''/a$ Equation (52) reduces to

$$\delta\sigma_{\mathbf{k}}'' - \frac{a''}{a}\delta\sigma_{\mathbf{k}} = 0,$$

which is satisfied by

$$\delta\sigma_{\mathbf{k}} = B(k)\,a \quad (k \ll aH) \tag{55}$$

where $B(k)$ is a constant of integration. Roughly matching the (absolute values of the) solutions (54) and (55) at $k = aH$ $(-k\tau = 1)$, we can determine the (absolute value of the) constant $B(k)$

$$|B(k)|\,a = \frac{1}{\sqrt{2k}} \Longrightarrow |B(k)| = \frac{1}{a\sqrt{2k}} = \frac{H}{\sqrt{2k^3}}.$$

Going back to the original variable $\delta\chi_{\mathbf{k}}$, we obtain that the quantum fluctuation of the χ field on superhorizon scales is constant and approximately equal to

$$|\delta\chi_{\mathbf{k}}| \simeq \frac{H}{\sqrt{2k^3}} \quad \text{(ON SUPERHORIZON SCALES)}$$

In fact we can do much better, since Eq. (52) has an *exact* solution:

$$\delta\sigma_{\mathbf{k}} = \frac{e^{-ik\tau}}{\sqrt{2k}}\left(1 + \frac{i}{k\tau}\right). \tag{56}$$

This solution reproduces all that we have found by qualitative arguments in the two extreme regimes $k \ll aH$ and $k \gg aH$. We have performed the matching procedure to show that the latter can be very useful to determine the behaviour of the solution on superhorizon scales when the exact solution is not known.

6.2 The power spectrum

Let us define now the power spectrum, a useful quantity to characterize the properties of the perturbations. For a generic quantity $g(\mathbf{x}, t)$, which can expanded in Fourier space as

$$g(\mathbf{x}, t) = \int \frac{d^3\mathbf{k}}{(2\pi)^{3/2}} \, e^{i\mathbf{k}\cdot\mathbf{x}} \, g_{\mathbf{k}}(t),$$

the power spectrum can be defined as

$$\langle 0 | g^*_{\mathbf{k}_1} g_{\mathbf{k}_2} | 0 \rangle \equiv \delta^{(3)} \left(\mathbf{k}_1 - \mathbf{k}_2 \right) \, \frac{2\pi^2}{k^3} \, \mathcal{P}_g(k), \tag{57}$$

where $|0\rangle$ is the vacuum quantum state of the system. This definition leads to the usual relation

$$\langle 0 | g^2(\mathbf{x}, t) | 0 \rangle = \int \frac{dk}{k} \, \mathcal{P}_g(k). \tag{58}$$

6.3 Quantum fluctuations of a generic scalar field in a quasi de Sitter stage

So far, we have computed the time evolution and the spectrum of the quantum fluctuations of a generic scalar field χ supposing that the scale factor evolves like in a pure de Sitter expansion, $a(\tau) = -1/(H\tau)$. However, during inflation the Hubble rate is not exactly constant, but changes with time as $\dot{H} = -\epsilon \, H^2$ (quasi de Sitter expansion). In this subsection, we shall solve for the perturbations in a quasi de Sitter expansion. Using the definition of the conformal time, one can show that the scale factor for small values of ϵ becomes

$$a(\tau) = -\frac{1}{H} \frac{1}{\tau(1-\epsilon)}.$$

The fluctuation mass-squared mass term is

$$M^2(\tau) = m^2_\chi a^2 - \frac{a''}{a},$$

where

$$\begin{aligned} \frac{a''}{a} &= a^2 \left(\frac{\ddot{a}}{a} + H^2 \right) = a^2 \left(\dot{H} + 2\, H^2 \right) \\ &= a^2 \, (2 - \epsilon) \, H^2 = \frac{(2-\epsilon)}{\tau^2 \, (1-\epsilon)^2} \\ &\simeq \frac{1}{\tau^2} \, (2 + 3\epsilon) \, . \end{aligned} \tag{59}$$

Armed with these results, we may compute the variance of the perturbations of the generic χ field

$$\begin{aligned} \langle 0 | \left(\delta\chi(\mathbf{x}, t) \right)^2 | 0 \rangle &= \int \frac{d^3k}{(2\pi)^3} \, |\delta\chi_{\mathbf{k}}|^2 \\ &= \int \frac{dk}{k} \, \frac{k^3}{2\pi^2} \, |\delta\chi_{\mathbf{k}}|^2 \\ &= \int \frac{dk}{k} \, \mathcal{P}_{\delta\chi}(k), \end{aligned} \tag{60}$$

which defines the power spectrum of the fluctuations of the scalar field χ

$$\mathcal{P}_{\delta\chi}(k) \equiv \frac{k^3}{2\pi^2} \, |\delta\chi_{\mathbf{k}}|^2 \, . \tag{61}$$

Since we have seen that fluctuations are (nearly) frozen in on superhorizon scales, a way of characterizing the perturbations is to compute the spectrum on scales larger than the horizon. For a massive scalar field, we obtain

$$\mathcal{P}_{\delta\chi}(k) = \left(\frac{H}{2\pi}\right)^2 \left(\frac{k}{aH}\right)^{3-2\nu_\chi}, \tag{62}$$

where, taking $m_\chi^2/H^2 = 3\eta_\chi$ and expanding for small values of ϵ and η,

$$\nu_\chi \simeq \frac{3}{2} + \epsilon - \eta_\chi. \tag{63}$$

We may also define the *spectral index* $n_{\delta\chi}$ of the fluctuations as

$$n_{\delta\chi} - 1 = \frac{d\ln \mathcal{P}_{\delta\phi}}{d\ln k} = 3 - 2\nu_\chi = 2\eta_\chi - 2\epsilon.$$

The power spectrum of fluctuations of the scalar field χ is therefore *nearly flat*, that is is nearly independent of the wavelength $\lambda = \pi/k$: the amplitude of the fluctuation on superhorizon scales does almost not depend upon the time at which the fluctuation crosses the horizon and becomes frozen in. The small tilt of the power spectrum arises from the fact that the scalar field χ is massive and because during inflation the Hubble rate is not exactly constant, but nearly constant, where 'nearly' is quantified by the slow-roll parameters ϵ. Adopting the traditional terminology, we may say that the spectrum of perturbations is blue if $n_{\delta\chi} > 1$ (more power in the ultraviolet) and red if $n_{\delta\chi} < 1$ (more power in the infrared). The power spectrum of the perturbations of a generic scalar field χ generated during a period of slow-roll inflation may be either blue or red. This depends upon the relative magnitude between η_χ and ϵ.

Comment: We might have computed the spectral index of the spectrum $\mathcal{P}_{\delta\chi}(k)$ by first solving the equation for the perturbations of the field χ in a di Sitter stage, with $H =$ constant and therefore $\epsilon = 0$, and then taking into account the time evolution of the Hubble rate introducing the subscript in H_k whose time variation is determined by Eq. (44). Correspondingly, H_k is the value of the Hubble rate when a given wavelength $\sim k^{-1}$ crosses the horizon (from that point on the fluctuation remains frozen in). The power spectrum in such an approach would read

$$\mathcal{P}_{\delta\chi}(k) = \left(\frac{H_k}{2\pi}\right)^2 \left(\frac{k}{aH}\right)^{3-2\nu_\chi} \tag{64}$$

with $3 - 2\nu_\chi \simeq \eta_\chi$. Using Eq. (44), one finds

$$n_{\delta\chi} - 1 = \frac{d\ln \mathcal{P}_{\delta\phi}}{d\ln k} = \frac{d\ln H_k^2}{d\ln k} + 3 - 2\nu_\chi = 2\eta_\chi - 2\epsilon$$

which reproduces our previous findings.

Comment: Since on superhorizon scales

$$\delta\chi_{\mathbf{k}} \simeq \frac{H}{\sqrt{2k^3}} \left(\frac{k}{aH}\right)^{\eta_\chi - \epsilon} \simeq \frac{H}{\sqrt{2k^3}} \left[1 + (\eta_\chi - \epsilon) \ln\left(\frac{k}{aH}\right)\right],$$

we discover that

$$|\delta\dot{\chi}_{\mathbf{k}}| \simeq |H\,(\eta_\chi - \epsilon)\;\delta\chi_{\mathbf{k}}| \ll |H\,\delta\chi_{\mathbf{k}}|\,, \tag{65}$$

that is, on superhorizon scales the time variation of the perturbations can be safely neglected.

7 Quantum fluctuations during inflation

As we have mentioned in the previous section, the linear theory of the cosmological perturbations represents a cornerstone of modern cosmology and is used to describe the formation and evolution of structures in the universe as well as the anisotropies of the CMB. The seeds for these inhomogeneities were generated during inflation and stretched over astronomical scales because of the rapid superluminal expansion of the universe during the (quasi) de Sitter epoch.

In the previous section we have already seen that pertubations of a generic scalar field χ are generated during a (quasi) de Sitter expansion. The inflaton field is a scalar field and, as such, we conclude that inflaton fluctuations will be generated as well. However, the inflaton is special from the point of view of perturbations. The reason is very simple. By assumption, the inflaton field dominates the energy density of the universe during inflation. Any perturbation in the inflaton field means a perturbation of the stress energy momentum tensor

$$\delta\phi \Longrightarrow \delta T_{\mu\nu}.$$

A perturbation in the stress energy momentum tensor implies, through Einstein's equations of motion, a perturbation of the metric

$$\delta T_{\mu\nu} \Longrightarrow \left[\delta R_{\mu\nu} - \frac{1}{2}\delta\left(g_{\mu\nu}R\right)\right] = 8\pi G\delta T_{\mu\nu} \Longrightarrow \delta g_{\mu\nu}.$$

On the other hand, a pertubation of the metric induces a back-reaction on the evolution of the inflaton perturbation through the perturbed Klein–Gordon equation of the inflaton field

$$\delta g_{\mu\nu} \Longrightarrow \delta\left(\partial_\mu\partial^\mu\phi + \frac{\partial V}{\partial\phi}\right) = 0 \Longrightarrow \delta\phi.$$

This logic chain makes us conclude that the perturbations of the inflaton field and of the metric are tightly coupled to each other and have to be studied together

$$\delta\phi \Longleftrightarrow \delta g_{\mu\nu}\,.$$

As we shall see shortly, this relation is stronger than one might think because of the issue of gauge invariance.

Before launching ourselves into the problem of finding the evolution of the quantum perturbations of the inflaton field when they are coupled to gravity, let us give a heuristic explanation of why we expect that during inflation such fluctuations are indeed present.

If we take Eq. (34) and split the inflaton field as its classical value ϕ_0 plus the quantum flucutation $\delta\phi$, $\phi(\mathbf{x},t) = \phi_0(t) + \delta\phi(\mathbf{x},t)$, the quantum perturbation $\delta\phi$ satisfies the equation of motion

$$\delta\ddot{\phi} + 3H\,\delta\dot{\phi} - \frac{\nabla^2\delta\phi}{a^2} + V''\,\delta\phi = 0. \tag{66}$$

Differentiating Eq. (39) wrt time and taking H constant (de Sitter expansion) we find

$$(\phi_0)^{\cdots} + 3H\ddot{\phi}_0 + V''\,\dot{\phi}_0 = 0. \tag{67}$$

Let us consider for simplicity the limit $\mathbf{k}^2/a^2 \ll 1$ and let us disregard the gradient term. Under this condition we see that $\dot{\phi}_0$ and $\delta\phi$ solve the same equation. The solutions have therefore to be related to each other by a constant of proportionality which depends upon time

$$\delta\phi = -\dot{\phi}_0\, \delta t(\mathbf{x}). \tag{68}$$

This tells us that $\phi(\mathbf{x}, t)$ will have the form

$$\phi(\mathbf{x}, t) = \phi_0\left(\mathbf{x}, t - \delta t(\mathbf{x})\right).$$

This equation indicates that the inflaton field does not acquire the same value at a given time t in all the space. On the contrary, when the inflaton field is rolling down its potential, it acquires different values from one spatial point $\mathbf{x}$ to the next. The inflaton field is not homogeneous and fluctuations are present. These fluctuations, in turn, will induce fluctuations in the metric.

7.1 The metric fluctuations

The mathematical tool to describe the linear evolution of the cosmological perturbations is obtained by perturbing at the first order the FRW metric $g^{(0)}_{\mu\nu}$, see Eq. (1)

$$g_{\mu\nu} \quad = \quad g^{(0)}_{\mu\nu}(t) \,+\, g_{\mu\nu}(\mathbf{x}, t)\,; \qquad g_{\mu\nu} \ll g^{(0)}_{\mu\nu}\,. \tag{69}$$

The metric perturbations can be decomposed according to their spin with respect to a local rotation of the spatial coordinates on hypersurfaces of constant time. This leads to

- *scalar perturbations*
- *vector perturbations*
- *tensor perturbations*

Tensor perturbations or gravitational waves have spin 2 and are the true degrees of freedom of the gravitational fields in the sense that they can exist even in the vacuum. Vector perturbations are spin 1 modes arising from rotational velocity fields and are also called vorticity modes. Finally, scalar perturbations have spin 0.

Let us do a simple exercise to count how many scalar degrees of freedom are present. Take a space-time of dimensions $D = n + 1$, of which n coordinates are spatial coordinates. The symmetric metric tensor $g_{\mu\nu}$ has $\frac{1}{2}(n+2)(n+1)$ degrees of freedom. We can perform $(n+1)$ coordinate transformations in order to eliminate $(n+1)$ degrees of freedom, this leaves us with $\frac{1}{2}n(n+1)$ degrees of freedom. These $\frac{1}{2}n(n+1)$ degrees of freedom contain scalar, vector and tensor modes. According to Helmholtz's theorem we can always decompose a vector u_i $(i = 1, \cdots, n)$ as $u_i = \partial_i v + v_i$, where v is a scalar (usually called potential flow) which is curl-free, $v_{[i,j]} = 0$, and v_i is a real vector (usually called vorticity) which is divergence-free, $\nabla \cdot v = 0$. This means that the real vector (vorticity) modes are $(n-1)$. Furthermore, a generic traceless tensor Π_{ij} can always be decomposed as $\Pi_{ij} = \Pi^S_{ij} + \Pi^V_{ij} + \Pi^T_{ij}$, where $\Pi^S_{ij} = \left(-\frac{k_i k_j}{k^2} + \frac{1}{3}\delta_{ij}\right)\Pi$, $\Pi^V_{ij} = (-i/2k)\left(k_i\Pi_j + k_j\Pi_i\right)$ $(k_i\Pi_i = 0)$ and $k_i\Pi^T_{ij} = 0$. This means that the true symmetric, traceless and transverse tensor degreees of freedom are $\frac{1}{2}(n-2)(n+1)$.

The number of scalar degrees of freedom is therefore

$$\frac{1}{2}n(n+1) - (n-1) - \frac{1}{2}(n-2)(n+1) = 2,$$

while the degrees of freedom of true vector modes are $(n-1)$ and the number of degrees of freedom of true tensor modes (gravitational waves) is $\frac{1}{2}(n-2)(n+1)$. In four dimensions $n = 3$, meaning that one expects 2 scalar degrees of freedom, 2 vector degrees of freedom and 2 tensor degrees of freedom. As we shall see, to the 2 scalar degrees of freedom from the metric, one has to add another one, the inflaton field perturbation $\delta\phi$. However, since Einstein's equations will tell us that the two scalar degrees of freedom

from the metric are equal during inflation, we expect a total number of scalar degrees of freedom equal to 2.

At the linear order, the scalar, vector, and tensor perturbations evolve independently (they decouple) and it is therefore possible to analyse them separately. Vector perturbations are not excited during inflation because there are no rotational velocity fields during the inflationary stage. we shall analyse the generation of tensor modes (gravitational waves) in the following. For the time being we want to focus on the scalar degrees of freedom of the metric.

Considering only the scalar degrees of freedom of the perturbed metric, the most generic perturbed metric reads

$$g_{\mu\nu} \,=\, a^2 \left(\begin{array}{cc} -1 \,-\, 2\,\Phi & \partial_i B \\ \partial_i B & (1 \,-\, 2\,\psi)\,\delta_{ij} \,+\, D_{ij} E \end{array} \right), \tag{70}$$

while the line-element can be written as

$$ds^2 \,=\, a^2\big((-1 - 2\,\Phi)d\tau^2 \,+\, 2\,\partial_i B\,d\tau\,dx^i \,+\, ((1 - 2\,\psi)\delta_{ij} \,+\, D_{ij}E)\;dx^i\,dx^j\big). \tag{71}$$

Here $D_{ij} \,=\, \big(\partial_i\partial_j \,-\, \frac{1}{3}\,\delta_{ij}\,\nabla^2\big)$.

7.2 The issue of gauge invariance

When studying the cosmological density perturbations, what we are interested in is following the evolution of a space-time which is neither homogeneous nor isotropic. This is done by following the evolution of the differences between the actual space-time and a well understood reference space-time. So we shall consider small perturbations away from the homogeneous, isotropic space-time.

The reference system in our case is the spatially flat Friedmann–Robertson–Walker (FRW) space-time, with line element $ds^2 \,=\, a^2(\tau)\left\{d\tau^2 - \delta_{ij}dx^i dx^j\right\}$. Now, the key issue is that general relativity is a gauge theory where the gauge transformations are the generic coordinate transformations from one local reference frame to another.

When we compute the perturbation of a given quantity, this is defined to be the difference between the value that this quantity assumes on the real physical space-time and the value it assumes on the unperturbed background. Nonetheless, to perform a comparison between these two values, it is necessary to compute them at the same space-time point. Since the two values live on two different geometries, it is necessary to specify a map which allows one to link univocally the same point on the two different space-times. This correspondence is called a gauge choice and changing the map means performing a gauge transformation.

Fixing a gauge in general relativity implies choosing a coordinate system. A choice of coordinates defines a *threading* of space-time into lines (corresponding to fixed spatial coordinates $\mathbf{x}$) and a *slicing* into hypersurfaces (corresponding to fixed time τ). A choice of coordinates is called a *gauge* and there is no unique preferred gauge

GAUGE CHOICE $\iff$ SLICING AND THREADING

From a more formal point of view, operating an infinitesimal gauge transformation on the coordinates

$$\widetilde{x^\mu} \,=\, x^\mu \,+\, \delta x^\mu \tag{72}$$

implies on a generic quantity Q a transformation on its perturbation

$$\widetilde{\delta Q} \,=\, \delta Q \,+\, \pounds_{\delta x}\, Q_0 \tag{73}$$

where Q_0 is the value assumed by the quantity Q on the background and $\pounds_{\delta x}$ is the Lie-derivative of Q along the vector δx^μ.

Decomposing in the usual manner the vector δx^μ

$$\begin{aligned} \delta x^0 &= \xi^0(x^\mu)\,; \\ \delta x^i &= \partial^i \beta(x^\mu) + v^i(x^\mu)\,; \qquad \partial_i v^i = 0\,, \end{aligned} \tag{74}$$

we can easily deduce the transformation law of a scalar quantity f (like the inflaton scalar field ϕ and energy density ρ). Instead of applying the formal definition (73), we find the transformation law in an alternative (and more pedagogical) way. We first write $\delta f(x) = f(x) - f_0(x)$, where $f_0(x)$ is the background value. Under a gauge transformation we have $\widetilde{\delta f}(\widetilde{x^\mu}) = \widetilde{f}(\widetilde{x^\mu}) - \widetilde{f_0}(\widetilde{x^\mu})$. Since f is a scalar we can write $f(\widetilde{x^\mu}) = f(x^\mu)$ (the value of the scalar function in a given physical point is the same in all the coordinate system). On the other side, on the unperturbed background hypersurface $\widetilde{f_0} = f_0$. We have therefore

$$\begin{aligned} \widetilde{\delta f}(\widetilde{x^\mu}) &= \widetilde{f}(\widetilde{x^\mu}) - \widetilde{f_0}(\widetilde{x^\mu}) \\ &= f(x^\mu) - f_0(\widetilde{x^\mu}) \\ &= f\left(\widetilde{x^\mu}\right) - f_0(\widetilde{x^\mu}) \\ &= f(\widetilde{x^\mu}) - \delta x^\mu \frac{\partial f}{\partial x^\mu}(\widetilde{x}) - f_0(\widetilde{x^\mu}), \end{aligned} \tag{75}$$

from which we finally deduce, being $f_0 = f_0(x^0)$,

$$\boxed{\widetilde{\delta f} = \delta f - f'\,\xi^0}$$

For the spin-zero perturbations of the metric, we can proceed analogously. We use the following trick. Upon a coordinate transformation $x^\mu \to \widetilde{x^\mu} = x^\mu + \delta x^\mu$, the line element is left invariant, $ds^2 = \widetilde{ds^2}$. This implies, for instance, that $a^2(\widetilde{x^0})\left(1+\widetilde{\Phi}\right)\left(\widetilde{dx^0}\right)^2 = a^2(x^0)\,(1+\Phi)\,(dx^0)^2$. Since $a^2(\widetilde{x^0}) \simeq a^2(x^0) + 2a\,a'\,\xi^0$ and $\widetilde{dx^0} = \left(1+\xi^{0\prime}\right)dx^0 + \frac{\partial x^0}{\partial x^i}\,dx^i$, we obtain $1 + 2\Phi = 1 + 2\widetilde{\Phi} + 2\,\mathcal{H}\xi^0 + 2\xi^{0\prime}$. We now may introduce in detail some gauge-invariant quantities which play a major role in the computation of the density perturbations. In the following we shall be interested only in the coordinate transformations on constant time hypersurfaces and therefore gauge invariance will be equivalent to independence of the slicing.

7.3 The co-moving curvature perturbation

The intrinsic spatial curvature on hypersurfaces on constant conformal time τ and for a flat universe is given by

$$^{(3)}R = \frac{4}{a^2}\nabla^2\,\psi.$$

The quantity ψ is usually referred to as the *curvature perturbation*. We have seen, however, that the curvature potential ψ is *not* gauge invariant, but is defined only on a given slicing. Under a transformation on constant time hypersurfaces $t \to t + \delta\tau$ (change of the slicing)

$$\psi \to \psi + \mathcal{H}\,\delta\tau.$$

We now consider the *co-moving slicing* which is defined to be the slicing orthogonal to the worldlines of co-moving observers. The latter are are free-falling and the expansion defined by them is isotropic. In practice, what this means is that there is no flux of energy measured by these observers, that is $T_{0i} = 0$. During inflation this means that these observers measure $\delta\phi_{\rm com} = 0$ since T_{0i} goes like $\partial_i\delta\phi(\mathbf{x},\tau)\phi'(\tau)$.

Since $\delta\phi \to \delta\phi - \phi'\delta\tau$ for a transformation on constant time hypersurfaces, this means that

$$\delta\phi \to \delta\phi_{\rm com} = \delta\phi - \phi'\,\delta\tau = 0 \Longrightarrow \delta\tau = \frac{\delta\phi}{\phi'},$$

that is $\delta\tau = \frac{\delta\phi}{\phi'}$ is the time-displacement needed to go from a generic slicing with generic $\delta\phi$ to the co-moving slicing where $\delta\phi_{\rm com} = 0$. At the same time the curvature perturbation ψ transforms into

$$\psi \to \psi_{\rm com} = \psi + \mathcal{H}\,\delta\tau = \psi + \mathcal{H}\frac{\delta\phi}{\phi'}.$$

The quantity

$$\boxed{\mathcal{R} = \psi + \mathcal{H}\frac{\delta\phi}{\phi'} = \psi + H\frac{\delta\phi}{\dot\phi}}$$

is the *co-moving curvature perturbation*. This quantity is gauge invariant by construction and is related to the gauge-dependent curvature perturbation ψ on a generic slicing to the inflaton perturbation $\delta\phi$ in that gauge. By construction, the meaning of $\mathcal{R}$ is that it represents the gravitational potential on co-moving hypersurfaces where $\delta\phi = 0$ or the inflaton fluctuation hypersurfaces where $\psi = 0$:

$$\mathcal{R} = \psi|_{\delta\phi=0} = H\frac{\delta\phi}{\dot\phi}\Big|_{\psi=0}.$$

The power spectrum of the curvature perturbation may then be easily computed

$$\mathcal{R}_{\mathbf{k}} = H\,\frac{\delta\phi_{\mathbf{k}}}{\dot\phi}. \tag{76}$$

We may now compute the power spectrum of the co-moving curvature perturbation on superhorizon scales

$$\boxed{\mathcal{P}_{\mathcal{R}}(k) = \frac{1}{2m_{\rm Pl}{}^2\epsilon}\left(\frac{H}{2\pi}\right)^2\left(\frac{k}{aH}\right)^{n_{\mathcal{R}}-1} \equiv A^2_{\mathcal{R}}\left(\frac{k}{aH}\right)^{n_{\mathcal{R}}-1}}$$

where we have defined the *spectral index* $n_{\mathcal{R}}$ of the co-moving curvature perturbation as

$$\boxed{n_{\mathcal{R}} - 1 = \frac{d\ln\mathcal{P}_{\mathcal{R}}}{d\ln k} = 3 - 2\nu = 2\eta - 6\epsilon.}$$

We conclude that inflation is responsible for the generation of adiabatic/curvature perturbations with an almost scale-independent spectrum. To compute the spectral index of the spectrum $\mathcal{P}_{\mathcal{R}}(k)$ we have proceeded as follows: first solve the equation for the perturbation $\delta\phi_{\mathbf{k}}$ in a de Sitter stage, with $H =$ constant ($\epsilon = \eta = 0$), whose solution is Eq. (56) and then taking into account the time-evolution of the Hubble rate and of ϕ introducing the subscript in H_k and $\dot{\phi}_k$. The time variation of the latter is determined by

$$\frac{d\ln\dot{\phi}_k}{d\ln k} = \left(\frac{d\ln\dot{\phi}_k}{dt}\right)\left(\frac{dt}{d\ln a}\right)\left(\frac{d\ln a}{d\ln k}\right) = \frac{\ddot{\phi}_k}{\dot{\phi}_k}\times\frac{1}{H}\times 1 = -\delta = \epsilon - \eta. \tag{77}$$

Correspondingly, $\dot{\phi}_k$ is the value of the time derivative of the inflaton field when a given wavelength $\sim k^{-1}$ crosses the horizon (from that point on the fluctuations remains frozen in). The curvature perturbation in such an approach would read

$$\mathcal{R}_{\mathbf{k}} \simeq \frac{H_k}{\dot{\phi}_k}\,\delta\phi_{\mathbf{k}} \simeq \frac{1}{2\pi}\left(\frac{H_k^2}{\dot{\phi}_k}\right).$$

Correspondingly

$$n_{\mathcal{R}} - 1 = \frac{d\ln\mathcal{P}_{\mathcal{R}}}{d\ln k} = \frac{d\ln H_k^4}{d\ln k} - \frac{d\ln\dot{\phi}_k^2}{d\ln k} = -4\epsilon + (2\eta - 2\epsilon) = 2\eta - 6\epsilon.$$

During inflation the curvature perturbation is generated on superhorizon scales with a spectrum which is nearly scale invariant [13], that is, is nearly independent of the wavelength $\lambda = \pi/k$: the amplitude of the fluctuation on superhorizon scales does not (almost) depend upon the time at which the fluctuation crosses the horizon and becomes frozen in. The small tilt of the power spectrum arises from the fact that the inflaton field is massive, giving rise to a non-vanishing η and because during inflation the Hubble rate is not exactly constant, but nearly constant, where 'nearly' is quantified by the slow-roll parameters ϵ.

Comment: From what we have found so far, we may conclude that on superhorizon scales the co-moving curvature perturbation $\mathcal{R}$ and the uniform-density gauge curvature ζ satisfy on superhorizon scales the relation

$$\dot{\mathcal{R}}_{\mathbf{k}} \simeq 0.$$

7.4 Gravitational waves

Quantum fluctuations in the gravitational fields are generated in a similar fashion to that of the scalar perturbations discussed so far. A gravitational wave may be viewed as a ripple of space-time in the FRW background metric (1) and in general the linear tensor perturbations may be written as

$$g_{\mu\nu} = a^2(\tau)\left[-d\tau^2 + (\delta_{ij} + h_{ij})\,dx^i dx^j\right],$$

where $|h_{ij}| \ll 1$. The tensor h_{ij} has six degrees of freedom, but, as we studied in Subsection 7.1, the tensor perturbations are traceless, $\delta^{ij}h_{ij} = 0$, and transverse $\partial^i h_{ij} = 0$ ($i = 1, 2, 3$). With these four constraints, there remain two physical degrees of freedom, or polarizations, which are usually indicated $\lambda = +, \times$. More precisely, we can write

$$h_{ij} = h_+\, e^+_{ij} + h_\times\, e^\times_{ij},$$

where e^+ and $e^\times$ are the polarization tensors which have the following properties

$$e_{ij} = e_{ji}, \quad k^i e_{ij} = 0, \quad , e_{ii} = 0,$$
$$e_{ij}(-\mathbf{k}, \lambda) = e^*_{ij}(\mathbf{k}, \lambda), \quad \sum_\lambda e^*_{ij}(\mathbf{k}, \lambda) e^{ij}(\mathbf{k}, \lambda) = 4.$$

Notice also that the tensors h_{ij} are gauge-invariant and therefore represent physical degrees of freedom.

If the stress-energy momentum tensor is diagonal, as the one provided by the inflaton potential $T_{\mu\nu} = \partial_\mu\phi\partial_\nu\phi - g_{\mu\nu}\mathcal{L}$, the tensor modes do not have any source in their equation and their action can be written as

$$\frac{{m_{\rm Pl}}^2}{2} \int d^4x \sqrt{-g} \frac{1}{2} \partial_\sigma h_{ij} \partial^\sigma h_{ij},$$

that is the action of four independent massless scalar fields. The gauge-invariant tensor amplitude

$$v_{\mathbf{k}} = a m_{\rm Pl} \frac{1}{\sqrt{2}} h_{\mathbf{k}},$$

satisfies therefore the equation

$$v''_{\mathbf{k}} + \left(k^2 - \frac{a''}{a}\right) v_{\mathbf{k}} = 0,$$

which is the equation of motion of a massless scalar field in a quasi-de Sitter epoch. We can therefore make use of the results present in Subsection 6.5 and Eq. (63) to conclude that on superhorizon scales the tensor modes scale like

$$|v_{\mathbf{k}}| = \left(\frac{H}{2\pi}\right) \left(\frac{k}{aH}\right)^{\frac{3}{2} - \nu_T},$$

where

$$\nu_T \simeq \frac{3}{2} - \epsilon.$$

Since fluctuations are (nearly) frozen in on superhorizon scales, a way of characterizing the tensor perturbations is to compute the spectrum on scales larger than the horizon

$$\mathcal{P}_T(k) = \frac{k^3}{2\pi^2} \sum_\lambda |h_{\mathbf{k}}|^2 = 4 \times 2 \frac{k^3}{2\pi^2} |v_{\mathbf{k}}|^2. \tag{78}$$

This gives the power spectrum on superhorizon scales

$$\boxed{\mathcal{P}_T(k) = \frac{8}{{m_{\rm Pl}}^2} \left(\frac{H}{2\pi}\right)^2 \left(\frac{k}{aH}\right)^{n_T} \equiv A_T^2 \left(\frac{k}{aH}\right)^{n_T}}$$

where we have defined the *spectral index* n_T of the tensor perturbations as

$$\boxed{n_T = \frac{d\ln \mathcal{P}_T}{d\ln k} = 3 - 2\nu_T = -2\epsilon.}$$

The tensor perturbation is almost scale-invariant. Notice that the amplitude of the tensor modes depends only on the value of the Hubble rate during inflation. This amounts to saying that it depends only on the energy scale $V^{1/4}$ associated to the inflaton potential. A detection of gravitational waves from inflation will therefore be a direct measurement of the energy scale associated to inflation.

7.5 The consistency relation

The results obtained so far for the scalar and tensor perturbations allow one to predict a *consistency relation* which holds for the models of inflation addressed in these lectures, i.e., the models of inflation driven by one-single field ϕ. We define the tensor-to-scalar amplitude ratio to be

$$r = \frac{\frac{1}{100}A_T^2}{\frac{4}{25}A_{\mathcal{R}}^2} = \frac{\frac{1}{100}8\left(\frac{H}{2\,\pi\,m_{\rm Pl}}\right)^2}{\frac{4}{25}(2\epsilon)^{-1}\left(\frac{H}{2\,\pi\,m_{\rm Pl}}\right)^2} = \epsilon.$$

This means that

$$\boxed{r = -\frac{n_T}{2}}$$

One-single models of inflation predict that during inflation driven by a single scalar field, the ratio between the amplitude of the tensor modes and that of the curvature perturbations is equal to minus one-half of the tilt of the spectrum of tensor modes. If this relation turns out to be falsified by the future measurements of the CMB anisotropies, this does not mean that inflation is wrong, but only that inflation has not been driven by only one field.

7.6 From the inflationary seeds to the matter power spectrum

As the curvature perturbations enter the causal horizon during radiation- or matter-domination, they create density fluctuations $\delta\rho_{\mathbf{k}}$ via gravitational attractions of the potential wells. The density contrast $\delta_{\mathbf{k}} = \frac{\delta\rho_{\mathbf{k}}}{\overline{\rho}}$ can be deduced from the Poisson equation

$$\frac{k^2\Phi_{\mathbf{k}}}{a^2} = -4\pi G\,\delta\rho_{\mathbf{k}} = -4\pi G\,\frac{\delta\rho_{\mathbf{k}}}{\overline{\rho}}\,\overline{\rho} = \frac{3}{2}\,H^2\,\frac{\delta\rho_{\mathbf{k}}}{\overline{\rho}}$$

where $\overline{\rho}$ is the background average energy density. This means that

$$\delta_{\mathbf{k}} = \frac{2}{3}\left(\frac{k}{aH}\right)^2\,\Phi_{\mathbf{k}}.$$

From this expression we can compute the power spectrum of matter density perturbations induced by inflation when they re-enter the horizon during matter-domination:

$$\mathcal{P}_{\delta\rho} = \langle|\delta_{\mathbf{k}}|^2\rangle = A\left(\frac{k}{aH}\right)^n = \frac{2\pi^2}{k^3}\left(\frac{2}{5}\right)^2 A_{\mathcal{R}}^2\left(\frac{k}{aH}\right)^4\left(\frac{k}{aH}\right)^{n_{\mathcal{R}}-1}$$

from which we deduce that matter perturbations scale linearly with the wave-number and have a scalar tilt

$$n = n_{\mathcal{R}} = 1 + 2\eta - 6\epsilon.$$

The primordial spectrum $\mathcal{P}_{\delta\rho}$ is of course reprocessed by gravitational instabilities after the universe becomes matter-dominated. Indeed, as we have seen in Section 6, perturbations evolve after entering the horizon and the power spectrum will not remain constant. To see how the density contrast is reprocessed we have first to analyse how it evolves on superhorizon scales before horizon-crossing. We use the following trick. Consider a flat universe with average energy density $\overline{\rho}$. The corresponding Hubble rate is

$$H^2 = \frac{8\pi G}{3}\,\overline{\rho}.$$

A small positive fluctuation $\delta\rho$ will cause the universe to be closed:

$$H^2 = \frac{8\pi G}{3}\,(\overline{\rho} + \delta\rho) - \frac{k}{a^2}.$$

Substracting the two equations we find

$$\frac{\delta\rho}{\rho} = \frac{3}{8\pi G}\frac{k}{a^2\rho} \sim \left\{ \begin{array}{cc} a^2 & \mathrm{RD} \\ a & \mathrm{MD} \end{array} \right.$$

Notice that $\Phi_{\mathbf{k}} \sim \delta\rho a^2/k^2 \sim (\delta\rho/\rho)\rho a^2/k^2 =$ constant for both RD and MD which confirms our previous findings.

When the matter densities enter the horizon, they do not increase appreciably before matter-domination because before matter-domination pressure is too large and does not allow the matter inhomogeneities to grow. On the other hand, the suppression of growth due to radiation is restricted to scales smaller than the horizon, while large-scale perturbations remain unaffected. This is why the horizon size at equality sets an important scale for structure growth:

$$k_{\mathrm{EQ}} = H^{-1}\,(a_{\mathrm{EQ}}) \simeq 0.08\,h\,\mathrm{Mpc}^{-1}.$$

Therefore, perturbations with $k \gg k_{\mathrm{EQ}}$ are perturbations which have entered the horizon before matter-domination and have remained nearly constant till equality. This means that they are suppressed with respect to those perturbations having $k \ll k_{\mathrm{EQ}}$ by a factor $(a_{\mathrm{ENT}}/a_{\mathrm{EQ}})^2 = (k_{\mathrm{EQ}}/k)^2$. If we define the transfer function $T(k)$ by the relation $\mathcal{R}_{\mathrm{final}} = T(k)\,\mathcal{R}_{\mathrm{initial}}$ we find therefore that roughly speaking

$$T(k) = \left\{ \begin{array}{cc} 1 & k \ll k_{\mathrm{EQ}}, \\ (k_{\mathrm{EQ}}/k)^2 & k \gg k_{\mathrm{EQ}}. \end{array} \right.$$

The corresponding power spectrum will be

$$\mathcal{P}_{\delta\rho}(k) \sim \left\{ \begin{array}{cc} \left(\frac{k}{aH}\right) & k \ll k_{\mathrm{EQ}}, \\ \left(\frac{k}{aH}\right)^{-3} & k \gg k_{\mathrm{EQ}}. \end{array} \right.$$

Of course, a more careful computation needs to include many other effects such as neutrino free-streaming, photon diffusion and the diffusion of baryons along with photons. It is encouraging, however, that this rough estimate turns out to be confirmed by present data on large-scale structures [4].

The next step would be to investigate how the primordial perturbations generated by inflation flow into the CMB to produce their anisotropies.

8 From inflation to large-angle CMB anisotropy

As we have already mentioned, the high temperature of the early universe maintained a low equilibrium fraction of neutral atoms, and a correspondingly high number density of free electrons. Coulomb scattering between the ions and electrons kept them in local kinetic equilibrium, and Thomson scattering of

photons tended to maintain the isotropy of the CMB in the baryon rest frame. As the universe expanded and cooled, the dominant element hydrogen started to recombine when the temperature fell below $\sim$ 4000 K. This is a factor of 40 lower than might be anticipated from the 13.6 eV ionization potential of hydrogen, and is due to the large ratio of the number of photons to baryons. Through recombination, the mean-free path for Thomson scattering grew to the horizon size and CMB photons "decoupled" from matter. More precisely, the probability density that photons last scattered at some time defines the visibility function. This is now known to peak 380 kyr after the Big Bang with a width ~ 120 kyr. Since then, CMB photons have propagated relatively unimpeded for $13.7\,\mathrm{Gyr}$, covering a co-moving distance $\sim 14.1\,\mathrm{Gpc}$. The distribution of their energies carries the imprint of fluctuations in the radiation temperature, the gravitational potentials, and the peculiar velocity of the radiation where they last scattered, as the temperature anisotropies that we observe today.

Temperature fluctuations in the CMB arise due to various distinct physical effects: our peculiar velocity with respect to the cosmic rest frame; fluctuations in the gravitational potential on the last scattering surface; fluctuations intrinsic to the radiation field itself on the last scatteringsurface; the peculiar velocity of the last scatteringsurface and damping of anisotropies if the universe should be re-ionized after decoupling. The first effect gives rise to the dipole anisotropy. Finally, there is the contribution from the evolution of the anisotropies from the last scattering surface till today (which we shall neglect from now on).

The second effect, known as the Sachs–Wolfe effect is the dominat contribution to the anisotropy on large-angular scales, $\theta \gg \theta_{\mathrm{HOR}} \sim 1^\circ$. The last three effects provide the dominant contributions to the anisotropy on small-angular scales, $\theta \ll 1^\circ$.

8.1 Sachs–Wolfe plateau

We consider first the temperature fluctuations on large-angular scales that arise due to the Sachs–Wolfe effect. These anisotropies probe length scales that were superhorizon-sized at photon decoupling and therefore insensitive to microphysical processes. On the contrary, they provide a probe of the original spectrum of primeval fluctuations produced during inflation.

To proceed, we consider the CMB anisotropy measured at positions other than our own and at earlier times. This is called the brightness function $\Theta(t, \mathbf{x}, \mathbf{n}) \equiv \delta T(t, \mathbf{x}, \mathbf{n})/T(t)$. The photons with momentum $\mathbf{p}$ in a given range d^3p have intensity I proportional to $T^4(t, \mathbf{x}, \mathbf{n})$ and therefore $\delta I/I = 4\Theta$. The brightness function depends upon the direction $\mathbf{n}$ of the photon momentum or, equivalently, on the direction of observation $\mathbf{e} = -\mathbf{n}$. Because the CMB travels freely from the last-scattering, we can write

$$\frac{\delta T}{T} = \Theta\left(t_{\mathrm{LS}}, \mathbf{x}_{\mathrm{LS}}, \mathbf{n}\right) + \left(\frac{\delta T}{T}\right)_*,$$

where $\mathbf{x}_{\mathrm{LS}} = -x_{\mathrm{LS}}\mathbf{n}$ is the point of the origin of the photon coming from the direction $\mathbf{e}$. The co-moving distance of the last scatteringdistance is $x_{\mathrm{LS}} = 2/H_0$. The first term corresponds to the anisotropy already present at last scattering and the second term is the additional anisotropy acquired during the travel towards us, equal to minus the fractional pertubation in the redshift of the radiation. Notice that the separation between each term depends on the slicing, but the sum does not.

Consider the redshift perturbation on co-moving slicing. We imagine the universe populated by co-moving observers along the line of sight. The relative velocity of adjacent co-moving observers is equal to their distance times the velocity gradient measured along $\mathbf{n}$ of the photon. In the unperturbed universe, we have $\mathbf{u} = H\mathbf{r}$, leading to the velocity gradient $u_{ij} = \partial u_i/\partial r_j = u_{ij} = H(t)\delta_{ij}$ with zero vorticity and shear. Including a peculiar velocity field as perturbation, $\mathbf{u} = H\mathbf{r} + \mathbf{v}$ and $u_{ij} = H(t)\delta_{ij} + \frac{1}{a}\frac{\partial v_i}{\partial v_j}$. The corresponding Doppler shift is

$$\frac{d\lambda}{\lambda} = \frac{da}{a} + n_i n_j \frac{\partial v_i}{\partial x_j} dx.$$

The perturbed FRW equation is

$$\delta H = \frac{1}{3}\nabla \cdot \mathbf{v},$$

while

$$(\delta\rho)^{\cdot} = -3\rho\delta H - 3H\delta\rho.$$

Instead of $\delta\rho$, let us work with the density contrast $\delta = \delta\rho/\rho$. Remembering that $\rho \sim a^{-3}$, we find that $\dot{\delta} = -3\delta H$, which gives

$$\nabla \cdot \mathbf{v} = -\dot{\delta}_{\mathbf{k}}.$$

From the Euler equation $\dot{\mathbf{u}} = -\rho^{-1}\nabla p - \nabla\Phi$, we deduce $\dot{\mathbf{v}} + H\mathbf{v} = -\nabla\Phi - \rho^{-1}\nabla p$. Therefore, for $a \sim t^{2/3}$ and negligible pressure gradient, since the gravitational potential is constant, we find

$$\mathbf{v} = -t\nabla\Phi$$

leading to

$$\left(\frac{\delta T}{T}\right)_* = \int_0^{x_{\rm LS}} \frac{t}{a}\frac{d^2\Phi}{dx^2}\, dx. \tag{79}$$

The photon trajectory is $ad\mathbf{x}/dt = \mathbf{n}$. Using $a \sim t^{2/3}$ gives

$$x(t) = \int_t^{t_0} \frac{dt'}{a} = 3\left(\frac{a_0}{t_0} - \frac{t}{a}\right).$$

Integrating by parts Eq. (79), we finally find

$$\left(\frac{\delta T}{T}\right)_* = \frac{1}{3}\left[\Phi(\mathbf{x}_{\rm LS}) - \Phi(0)\right] + \mathbf{e}\cdot\left[\mathbf{v}(0,t_0) - \mathbf{v}(\mathbf{x}_{\rm LS}, t_{\rm LS})\right].$$

The potential at our position contributes only to the unobservable monopole and can be dropped. On scales outside the horizon, $\mathbf{v} = -t\nabla\Phi \sim 0$. The remaining term is the Sachs–Wolfe effect

$$\frac{\delta T(\mathbf{e})}{T} = \frac{1}{3}\Phi(\mathbf{x}_{\rm LS}) = \frac{1}{5}\mathcal{R}(\mathbf{x}_{\rm LS}).$$

This relation has been obtained as follows. The co-moving curvature perturbation is given during the radiation phase by $\mathcal{R} = \psi + H\delta\rho/\dot{\rho} = \psi - 1/3\delta\rho_\gamma/\rho_\gamma$. Einstein equations set $\psi = \Phi$ and $\delta\rho_\gamma/\rho_\gamma = -2\Phi$ on super-horizon scales. Therefore $\mathcal{R} = 5/3\Phi$ beyond the horizon.

At large angular scales, the theory of cosmological perturbations predicts a remarkably simple formula relating the CMB anisotropy to the curvature perturbation generated during inflation.

We have seen previously that the temperature anisotropy is commonly expanded in spherical harmonics $\frac{\Delta T}{T}(x_0, \tau_0, \mathbf{n}) = \sum_{\ell m} a_{\ell,m}(x_0) Y_{\ell m}(\mathbf{n})$, where x_0 and τ_0 are our position and the preset time, respectively, $\mathbf{n}$ is the direction of observation, ℓ's are the different multipoles, and $\langle a_{\ell m} a^*_{\ell' m'}\rangle = \delta_{\ell,\ell'}\delta_{m,m'}C_\ell$, where the deltas are due to the fact that the process that created the anisotropy is statistically isotropic. The C_ℓ's are the so-called CMB power spectrum. For homogeneity and isotropy, the C_ℓ's are neither a function of x_0, nor of m. The two-point-correlation function is related to the C_l's according to Eq. (23).

For adiabatic perturbations we have seen that on large scales, larger than the horizon on the last scatteringsurface (corresponding to angles larger than $\theta_{\rm HOR} \sim 1^\circ$) $\delta T/T = \frac{1}{3}\Phi(\mathbf{x}_{\rm LS})$. In Fourier transform

$$\frac{\delta T(\mathbf{k},\tau_0,\mathbf{n})}{T} = \frac{1}{3}\Phi_{\mathbf{k}}\, e^{i\,\mathbf{k}\cdot\mathbf{n}(\tau_0-\tau_{\rm LS})}. \tag{80}$$

Using the decomposition

$$\exp(i\,\mathbf{k}\cdot\mathbf{n}(\tau_0-\tau_{\rm LS})) = \sum_{\ell=0}^{\infty}(2\ell+1)i^\ell j_\ell(k(\tau_0-\tau_{\rm LS}))P_\ell(\mathbf{k}\cdot\mathbf{n}) \tag{81}$$

where j_ℓ is the spherical Bessel function of order ℓ and substituting, we get

$$\Big\langle \frac{\delta T(x_0,\tau_0,\mathbf{n})}{T}\frac{\delta T(x_0,\tau_0,\mathbf{n}')}{T}\Big\rangle = \tag{82}$$

$$= \frac{1}{V}\int d^3x \Big\langle \frac{\delta T(x_0,\tau_0,\mathbf{n})}{T}\frac{\delta T(x_0,\tau_0,\mathbf{n}')}{T}\Big\rangle =$$

$$= \frac{1}{(2\pi)^3}\int d^3k \Big\langle \frac{\delta T(\mathbf{k},\tau_0,\mathbf{n})}{T}\left(\frac{\delta T(\mathbf{k},\tau_0,\mathbf{n}')}{T}\right)^*\Big\rangle =$$

$$= \frac{1}{(2\pi)^3}\int d^3k \Big(\Big\langle \frac{1}{3}|\Phi|^2\Big\rangle \sum_{\ell,\ell'=0}^{\infty}(2\ell+1)(2\ell'+1)j_\ell(k(\tau_0-\tau_{\rm LS}))$$

$$j_{\ell'}(k(\tau_0-\tau_{\rm LS}))P_\ell(\mathbf{k}\cdot\mathbf{n})P_{\ell'}(\mathbf{k}'\cdot\mathbf{n}')\Big) \tag{83}$$

Inserting $P_\ell(\mathbf{k}\cdot\mathbf{n}) = \frac{4\pi}{2\ell+1}\sum_m Y^*_{lm}(\mathbf{k})Y_{\ell m}(\mathbf{n})$ and analogously for $P_\ell(\mathbf{k}'\cdot\mathbf{n}')$, integrating over the directions $d\Omega_k$ generates $\delta_{\ell\ell'}\delta_{mm'}\sum_m Y^*_{\ell m}(\mathbf{n})Y_{\ell m}(\mathbf{n}')$. Using as well $\sum_m Y^*_{\ell m}(\mathbf{n})Y_{\ell m}(\mathbf{n}') = \frac{2\ell+1}{4\pi}P_\ell(\mathbf{n}\cdot\mathbf{n}')$, we get

$$\Big\langle \frac{\delta T(x_0,\tau_0,\mathbf{n})}{T}\frac{\delta T(x_0,\tau_0,\mathbf{n}')}{T}\Big\rangle \tag{84}$$

$$= \Sigma_\ell \frac{2\ell+1}{4\pi}P_\ell(\mathbf{n}\cdot\mathbf{n}')\frac{2}{\pi}\int\frac{dk}{k}\Big\langle\frac{1}{9}|\Phi|^2\Big\rangle k^3 j_\ell^2(k(\tau_0-\tau_{\rm LS})).$$

Comparing this expression with that for the C_ℓ, we get the expression for the $C_\ell^{\rm AD}$, where the suffix "AD" stands for adiabatic:

$$C_\ell^{\rm AD} = \frac{2}{\pi}\int\frac{dk}{k}\Big\langle\frac{1}{9}\,|\Phi|^2\,\Big\rangle k^3 j_\ell^2(k(\tau_0-\tau_{\rm LS})) \tag{85}$$

which is valid for $2 \le \ell \ll (\tau_0-\tau_{\rm LS})/\tau_{\rm LS} \sim 100$.

If we generically indicate by $\langle|\Phi_{\mathbf{k}}|^2\rangle k^3 = A^2\,(k\tau_0)^{n-1}$, we can perform the integration and get

$$\frac{\ell(\ell+1)C_\ell^{\rm AD}}{2\pi} = \left[\frac{\sqrt{\pi}}{2}\ell(\ell+1)\frac{\Gamma(\frac{3-n}{2})\Gamma(\ell+\frac{n-)}{2})}{\Gamma\left(\frac{4-n}{2}\right)\Gamma\left(\ell+\frac{5-n}{2}\right)}\right]\frac{A^2}{9}\left(\frac{H_0}{2}\right)^{n-1}. \tag{86}$$

For $n \simeq 1$ and $\ell \gg 1$, we can approximate this expression to

$$\frac{\ell(\ell+1)C_l^{\rm AD}}{2\pi} = \frac{A^2}{9}. \tag{87}$$

This result shows that inflation predicts a very flat spectrum for low ℓ. Furthermore, since inflation predicts $\Phi_{\mathbf{k}} = \frac{3}{5}\mathcal{R}_{\mathbf{k}}$, we find that

$$\pi\,\ell(\ell+1)C_l^{\rm AD} = \frac{A_{\mathcal{R}}^2}{25} = \frac{1}{25}\frac{1}{2\,{m_{\rm Pl}}^2\,\epsilon}\left(\frac{H}{2\pi}\right)^2. \tag{88}$$

WMAP5 data imply that $\frac{\ell(\ell+1)C_l^{\rm AD}}{2\pi} \simeq 10^{-10}$ or

$$\left(\frac{V}{\epsilon}\right)^{1/4} \simeq 6.7 \times 10^{16}\,\mathrm{GeV}$$

8.2 Acoustic peaks

To be able to calculate the power spectrum of the anisotropies even on angular scales larger than 1°, we need to consider the evolution of the photon anistropies. As we already mentioned, before recombination Thomson scattering was very efficient. As a result it is a good approximation to treat photons and baryons as a single fluid. This treatment is called the tight-coupling approximation and will allow us to evolve the perturbations until recombination.

The equation for the photon density perturbations for one Fourier mode of wave-number k is that of a forced and damped harmonic oscillator

$$\begin{aligned}
&\ddot{\delta}_\gamma + \frac{\dot{R}}{(1+R)}\dot{\delta}_\gamma + k^2 c_s^2 \delta_\gamma = F, \\
&F = 4[\ddot{\psi} + \frac{\dot{R}}{(1+R)}\dot{\psi} - \frac{1}{3}k^2\Phi], \\
&\dot{\delta}_\gamma = -\frac{4}{3}kv_\gamma + 4\dot{\psi}.
\end{aligned} \tag{89}$$

The photon–baryon fluid can sustain acoustic oscillations. The inertia is provided by the baryons, while the pressure is provided by the photons. The sound speed is $c_s^2 = 1/3(1+R)$, with $R = 3\rho_b/4\rho_\gamma = 31.5\ (\Omega_b h^2)(T/2.7)^{-4}[(1+z)/10^3]^{-1}$. As the baryon fraction goes down, the sound speed approaches $c_s^2 \to 1/3$. The third equation above is the continuity equation.

As a toy problem, we shall solve Eq. (89) under some simplifying assumptions. If we consider a matter-dominated universe, the driving force becomes a constant, $F = -4/3k^2\Phi$, because the gravitational potential remains constant in time. We neglect anisotropic stresses so that $\psi = \Phi$, and, furthermore, we neglect the time dependence of R. Equation (89) becomes that of a harmonic oscillator that can be trivially solved. This is a very simplified picture, but it captures most of the relevant physics we want to discuss.

To obtain the final solution we need again to specify the initial conditions. we shall restrict ourselves to adiabatic initial conditions, the most natural outcome of inflation. In our context this means that initially $\Phi = \psi = \Phi_0$, $\delta_\gamma = -8/3\Phi_0$, and $v_\gamma = 0$. We have denoted Φ_0 the initial amplitude of the potential fluctuations. We shall take Φ_0 to be a Gaussian random variable with power spectrum P_{Φ_0}.

We have made enough approximations that the evaluation of the sources in the integral solution has become trivial. The solution for the density and velocity of the photon fluid at recombination is

$$\begin{aligned}
\left(\frac{\delta_\gamma}{4} + \Phi\right)|_{\rm LS} &= \frac{\Phi_0}{3}(1+3R)\cos(kc_s\tau_{\rm LS}) - \Phi_0 R, \\
v_\gamma|_{\tau_{\rm LS}} &= -\Phi_0(1+3R)c_s\sin(kc_s\tau_{\rm LS}).
\end{aligned} \tag{90}$$

Equation (90) is the solution for a single Fourier mode. All quantities have an additional spatial dependence ($e^{i\mathbf{k}\cdot\mathbf{x}}$), which we have not included in order to make the notation more compact. With that additional term the solution we have is

$$\begin{aligned}
\frac{\delta T}{T}(\mathbf{n}) &= e^{ikD_{\rm LS}\cos\theta}S \\
S &= \Phi_0\frac{(1+3R)}{3}[\cos(kc_s\tau_{\rm LS}) - \frac{3R}{(1+3R)},
\end{aligned}$$

$$-i\sqrt{\frac{3}{1+R}}\cos\theta\sin(kc_s\tau_{\rm LS})], \tag{91}$$

where we have neglected the Φ on the left-hand side because it is a constant. We have introduced $\cos\theta$, the cosine of the angle between the direction of observation and the wavevector $\mathbf{k}$; for example, $\mathbf{k}\cdot\mathbf{x} = kD_{\rm LS}\cos\theta$. The term proportional to $\cos\theta$ is the Doppler contribution.

Once the temperature perturbation produced by one Fourier mode has been calculated, we need to expand it into spherical harmonics. The power spectrum of temperature anisotropies is expressed in terms of the a_{lm} coefficients as $C_{T\ell} = \sum_m |a_{\ell m}|^2$. The contribution to C_{Tl} from each Fourier mode is weighted by the amplitude of primordial fluctuations in this mode, characterized by the power spectrum of $\Phi_0 = 3/5\mathcal{R}$, $P_{\Phi_0} = Ak^{-3}$ as dictated by inflation. In practice, fluctuations on angular scale ℓ receive most of their contributions from wavevectors around $k_* = \ell/D_{\rm LS}$, so roughly the amplitude of the power spectrum at multipole ℓ is given by the value of the sources in Eq. (90) at k_*.

After summing the contributions from all modes, the power spectrum is roughly given by

$$\begin{aligned}\ell(\ell+1)C_{Tl} &\approx A\{[\frac{(1+3R)}{3}\cos(k_*c_s\tau_{\rm LS}) - R]^2 + \\ &\quad \frac{(1+3R)^2}{3}c_s^2\sin^2(k_*c_s\tau_{\rm LS})\}.\end{aligned} \tag{92}$$

Equation (92) can be used to understand the basic features in the CMB power spectra. The baryon drag on the photon–baryon fluid reduces its sound speed below $1/3$ and makes the monopole contribution dominant (the one proportional to $\cos(k_*c_s\tau_{\rm LS})$. Thus, the C_{Tl} spectrum peaks where the monopole term peaks, $k_*c_s\tau_{\rm LS} = \pi, 2\pi, 3\pi, \cdots$, which correspond to $\ell_{\rm peak} = n\pi D_{\rm LS}/c_S\tau_{\rm LS}$.

It is very important to understand the origin of the acoustic peaks. In this model the universe is filled with standing waves; all modes of wave-number k are in phase, which leads to the oscillatory terms. The sine and cosine in Eq. (92) originate in the time dependence of the modes. Each mode ℓ receives contributions preferentially from Fourier modes of a particular wavelength k_* (but pointing in all directions), so to obtain peaks in C_ℓ, it is crucial that all modes of a given k be in phase. If this is not the case, the features in the $C_{T\ell}$ spectra will be blurred and can even disappear. This is what happens when one considers the spectra produced by topological defects. The phase coherence of all modes of a given wave-number can be traced to the fact that perturbations were produced very early on and had wavelengths larger than the horizon during many expansion times.

There are additional physical effects we have neglected. The universe was radiation dominated early on, and modes of wavelength smaller and bigger than the horizon at matter-radiation equality behave differently. During the radiation era the perturbations in the photon–baryon fluid are the main source for the gravitational potentials which decay once a mode enters into the horizon. The gravitational potential decay acts as a driving force for the oscillator in Eq. (89), so a feedback loop is established. As a result, the acoustic oscillations for modes that entered the horizon before matter-radiation equality have a higher amplitude. In the $C_{T\ell}$ spectrum the separation between modes that experience this feedback and those that do not occurs at $\ell \sim D_{\rm LS}/\tau_{\rm LS}$. Larger ℓ values receive their contributions from modes that entered the horizon before matter-radiation equality. Finally, when a mode is inside the horizon during the radiation era the gravitational potentials decay.

There is a competing effect, Silk damping, that reduces the amplitude of the large-l modes. The photon–baryon fluid is not a perfect fluid. Photons have a finite mean free path and thus can random-walk away from the peaks and valleys of the standing waves. Thus perturbations of wavelength comparable to or smaller than the distance the photons can random-walk get damped. This effect can be modelled by multiplying Eq. 91 by $\exp(-k^2/k_s^2)$, with $k_s^{-1} \propto \tau_{\rm LS}^{1/2}(\Omega_b h^2)^{-1/2}$. Silk damping is important for multipoles of order $\ell_{\rm Silk} \sim k_s D_{\rm LS}$. Finally, the last scatteringsurface has a finite width. Perturbations

with wavelength comparable to this width get smeared out due to cancellations along the line of sight. This effect introduces an additional damping with a characteristic scale $k_w^{-1} \propto \delta\tau_{\rm LS}$.

The location of the first peak is by itself a measurement of the geometry of the universe. In fact, photons propagating on geodesics from the last scattering surface to us feel the spatial geometry, whose properties we learned are dictated by Ω_0. In fact, the location of the first peak is given by $\ell_1 \simeq 220/\sqrt{\Omega_0}$. WMAP5 gives $\Omega_0 = 1.00^{+0.07}_{-0.03}$. This tells us that the spatial (local) geometry of the universe is flat. This is precisely what inflation predicts.

8.3 The polarization of the CMB anisotropies

The anisotropy field is characterized by a 2×2 intensity tensor I_{ij}. For convenience, we normalize this tensor so that it represents the fluctuations in units of the mean intensity ($I_{ij} = \delta I/I_0$). The intensity tensor is a function of direction on the sky, $\mathbf{n}$, and two directions perpendicular to $\mathbf{n}$ that are used to define its components ($\mathbf{e}_1$,$\mathbf{e}_2$). The Stokes parameters Q and U are defined as $Q = (I_{11} - I_{22})/4$ and $U = I_{12}/2$, while the temperature anisotropy is given by $T = (I_{11} + I_{22})/4$ (the factor of 4 relates fluctuations in the intensity with those in the temperature, $I \propto T^4$). When representing polarization using "rods" in a map, the magnitude is given by $P = \sqrt{Q^2 + U^2}$, and the orientation makes an angle $\alpha = \frac{1}{2}\arctan(U/Q)$ with $\mathbf{e}_1$. In principle the fourth Stokes parameter V that describes circular polarization is needed, but we ignore it because it cannot be generated through Thomson scattering, so the CMB is not expected to be circularly polarized. While the temperature is invariant under a right-handed rotation in the plane perpendicular to direction $\mathbf{n}$, Q and U transform under rotation by an angle ψ as

$$(Q \pm iU)'(\mathbf{n}) = e^{\mp 2i\psi}(Q \pm iU)(\mathbf{n}), \tag{93}$$

where $\mathbf{e}_1' = \cos\psi\ \mathbf{e}_1 + \sin\psi\ \mathbf{e}_2$ and $\mathbf{e}_2' = -\sin\psi\ \mathbf{e}_1 + \cos\psi\ \mathbf{e}_2$. The quantities $Q \pm iU$ are said to be spin 2.

We already mentioned that the statistical properties of the radiation field are usually described in terms of the spherical harmonic decomposition of the maps. This basis, basically the Fourier basis, is very natural because the statistical properties of anisotropies are rotationally invariant. The standard spherical harmonics are not the appropriate basis for $Q \pm iU$ because they are spin-2 variables, but generalizations (called ${}_{\pm 2}Y_{lm}$) exist. We can expand

$$(Q \pm iU)(\mathbf{n}) \ = \ \sum_{\ell m} a_{\pm 2,\ell m}\ {}_{\pm 2}Y_{\ell m}(\mathbf{n}). \tag{94}$$

Here Q and U are defined at each direction $\hat{\mathbf{n}}$ with respect to the spherical coordinate system $(\mathbf{e}_\theta, \mathbf{e}_\phi)$. To ensure that Q and U are real, the expansion coefficients must satisfy $a^*_{-2,\ell m} = a_{2,\ell -m}$. The equivalent relation for the temperature coefficients is $a^*_{T,\ell m} = a_{T,\ell -m}$. Instead of $a_{\pm 2,\ell m}$, it is convenient to introduce their linear combinations $a_{E,\ell m} = -(a_{2,\ell m} + a_{-2,\ell m})/2$ and $a_{B,\ell m} = i(a_{2,\ell m} - a_{-2,\ell m})/2$. We define two quantities in real space, $E(\mathbf{n}) = \sum_{\ell,m} a_{E,\ell m} Y_{\ell m}(\mathbf{n})$ and $B(\mathbf{n}) = \sum_{\ell,m} a_{B,\ell m} Y_{\ell m}(\mathbf{n})$. Here E and B completely specify the linear polarization field.

The temperature is a scalar quantity under a rotation of the coordinate system, $T'(\mathbf{n}' = \mathcal{R}\mathbf{n}) = T(\mathbf{n})$, where $\mathcal{R}$ is the rotation matrix. We denote with a prime the quantities in the transformed coordinate system. While $Q \pm iU$ are spin 2, $E(\mathbf{n})$ and $B(\mathbf{n})$ are invariant under rotations. Under parity, however, E and B behave differently, E remains unchanged, while B changes sign.

To characterize the statistics of the CMB perturbations, only four power spectra are needed, those for T, E, B and the cross correlation between T and E. The cross correlation between B and E or B and T vanishes if there are no parity-violating interactions because B has the opposite parity to T or E. The power spectra are defined as the rotationally invariant quantities $C_{T\ell} = \frac{1}{2\ell+1}\sum_m \langle a^*_{T,\ell m} a_{T,\ell m}\rangle$, $C_{E\ell} = \frac{1}{2\ell+1}\sum_m \langle a^*_{E,\ell m} a_{E,\ell m}\rangle$, $C_{B\ell} = \frac{1}{2\ell+1}\sum_m \langle a^*_{B,\ell m} a_{B,\ell m}\rangle$, and $C_{C\ell} = \frac{1}{2\ell+1}\sum_m \langle a^*_{T,\ell m} a_{E,\ell m}\rangle$. The brackets $\langle \cdots \rangle$ denote ensemble averages.

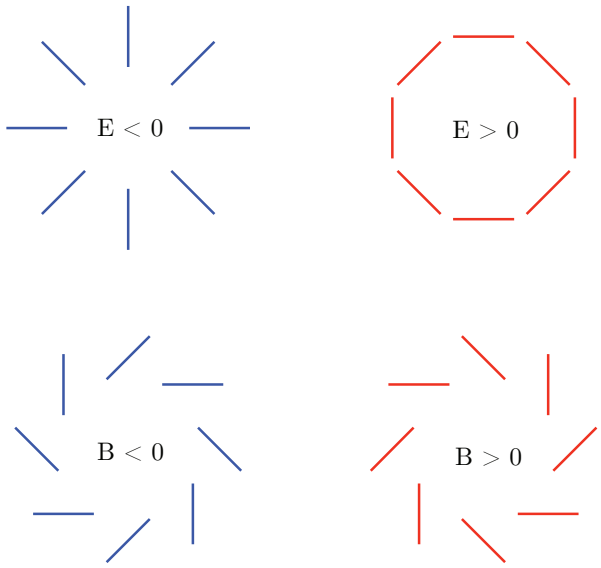


Fig. 4: Examples of E- and B-mode patterns of polarization

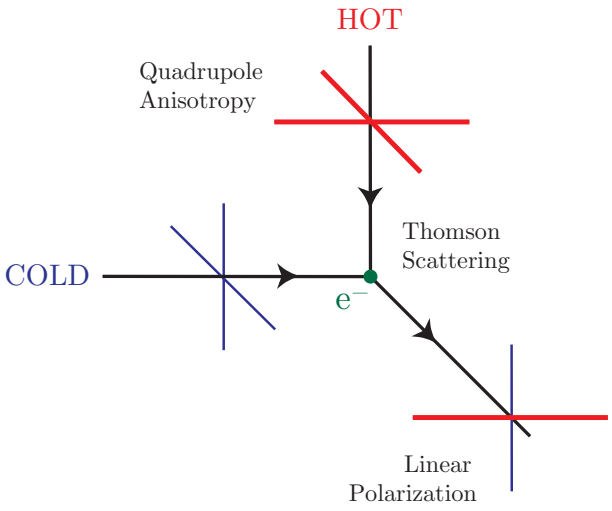


Fig. 5: Thomson scattering of radiation where quadrupole anisotropy generates linear polarization

Polarization is generated by Thomson scattering between photons and electrons, which means that polarization cannot be generated after recombination (except for re-ionization, which we shall discuss later). But Thomson scattering is not enough. The radiation incident on the electrons must also be anisotropic. In fact, its intensity needs to have a quadrupole moment. This requirement of having both Thomson scattering and anisotropies is what makes polarization relatively small. After recombination, anisotropies grow by free streaming, but there is no scattering to generate polarization. Before recombination there were so many scatterings that they erased any anisotropy present in the photon–baryon fluid.

In the context of anisotropies induced by density perturbations, velocity gradients in the photon–baryon fluid are responsible for the quadrupole that generates polarization. Let us consider a scattering occurring at position $\boldsymbol{x}_0$: the scattered photons came from a distance of order the mean free path (λ_T) away from this point. If we are considering photons traveling in direction $\hat{\boldsymbol{n}}$, they roughly come from $\boldsymbol{x} = \boldsymbol{x}_0 + \lambda_T \hat{\boldsymbol{n}}$. The photon–baryon fluid at that point was moving at velocity $\boldsymbol{v}(\boldsymbol{x}) \approx \boldsymbol{v}(\boldsymbol{x}_0) + \lambda_T \hat{\boldsymbol{n}}_i \partial_i \boldsymbol{v}(\boldsymbol{x}_0)$.

Due to the Doppler effect the temperature seen by the scatterer at $\boldsymbol{x}_0$ is $\delta T(\boldsymbol{x}_0, \hat{\boldsymbol{n}}) = \hat{\boldsymbol{n}} \cdot [\boldsymbol{v}(\boldsymbol{x}) - \boldsymbol{v}(\boldsymbol{x}_0)] \approx \lambda_T \hat{\boldsymbol{n}}_i \hat{\boldsymbol{n}}_j \partial_i \boldsymbol{v}_j(\boldsymbol{x}_0)$, which is quadratic in $\hat{\boldsymbol{n}}$ (i.e., it has a quadrupole). Velocity gradients in the photon–baryon fluid lead to a quadrupole component of the intensity distribution, which, through Thomson scattering, is converted into polarization.

The polarization of the scattered radiation field, expressed in terms of the Stokes parameters Q and U, is given by $(Q + iU) \propto \sigma_T \int d\Omega' (\mathbf{m} \cdot \hat{\mathbf{n}}')^2 T(\hat{\mathbf{n}}') \propto \lambda_p \mathbf{m}^i \mathbf{m}^j \partial_i v_j|_{\tau_{\rm LS}}$, where σ_T is the Thomson scattering cross-section and we have written the scattering matrix as $P(\mathbf{m}, \hat{\mathbf{n}}') = -3/4 \sigma_T (\mathbf{m} \cdot \hat{\mathbf{n}}')^2$, with $\mathbf{m} = \hat{\mathbf{e}}_1 + i \hat{\mathbf{e}}_2$. In the last step, we integrated over all directions of the incident photons $\hat{\mathbf{n}}'$. As photons decouple from the baryons, their mean free path grows very rapidly, so a more careful analysis is needed to obtain the final polarization:

$$(Q + iU)(\hat{\mathbf{n}}) \approx \epsilon \delta \tau_{\rm LS} \mathbf{m}^i \mathbf{m}^j \partial_i v_j|_{\tau_{\rm LS}}, \tag{95}$$

where $\delta \tau_{\rm LS}$ is the width of the last scattering surface and gives a measure of the distance that photons travel between their last two scatterings, and ϵ is a numerical constant that depends on the shape of the visibility function. The appearance of $\mathbf{m}^i \mathbf{m}^j$ in Eq. (95) ensures that $(Q + iU)$ transforms correctly under rotations of $(\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2)$.

If we evaluate Eq. (95) for each Fourier mode and combine them to obtain the total power, we get the equivalent of Eq. (92),

$$\ell(\ell + 1) C_{E\ell} \approx A \epsilon^2 (1 + 3R)^2 (k_* \delta \tau_{\rm LS})^2 \sin^2(k_* c_s \tau_{\rm LS}), \tag{96}$$

where we are assuming $n = 1$ and that ℓ is large enough that factors like $(\ell + 2)!/(\ell - 2)! \approx \ell^4$. The extra k_* in Eq. (96) originates in the gradient in Eq. (95). The large-angular scale polarization is greatly suppressed by the $k \delta \tau_{\rm LS}$ factor. Correlations over large angles can only be created by the long-wavelength perturbations, but these cannot produce a large polarization signal because of the tight coupling between photons and electrons prior to recombination. Multiple scatterings make the plasma very homogeneous; only wavelengths that are small enough to produce anisotropies over the mean free path of the photons will give rise to a significant quadrupole in the temperature distribution, and thus to polarization. Wavelengths much smaller than the mean free path decay due to photon diffusion (Silk damping) and so are unable to create a large quadrupole and polarization. As a result polarization peaks at the scale of the mean free path.

On sub-degree angular scales, temperature, polarization, and the cross-correlation power spectra show acoustic oscillations. In the polarization and cross-correlation spectra the peaks are much sharper. The polarization is produced by velocity gradients of the photon—baryon fluid at the last scatteringsurface. The temperature receives contributions from density and velocity perturbations, and the oscillations in each partially cancel one another, making the features in the temperature spectrum less sharp. The dominant contribution to the temperature comes from the oscillations in the density [Eq. (90)], which are out of phase with the velocity. This explains the difference in location between the temperature and polarization peaks. The extra gradient in the polarization signal, Eq. (95), explains why its overall amplitude peaks at a smaller angular scale.

Now, as photons travel in the metric perturbed by a GW [$ds^2 = a^2(\tau) \left[-d\tau^2 + (\delta_{ij} + h^T_{ij}) dx^i dx^j \right]$], they get redshifted or blueshifted depending on their direction of propagation relative to the direction of propagation of the GW and the polarization of the GW. For example, for a GW travelling along the z axis, the frequency shift is given by

$$\frac{1}{\nu} \frac{d\nu}{d\tau} = \frac{1}{2} \hat{n}^i \hat{n}^j \dot{h}^{T(\pm)}_{ij} = \frac{1}{2} (1 - \cos^2\theta) e^{\pm i 2\phi} \ \dot{h}_t \ \exp(i \mathbf{k} \cdot \mathbf{x}), \tag{97}$$

where (θ, ϕ) describe the direction of propagation of the photon, the $\pm$ correspond to the different polarizations of the GW, and h_t gives the time-dependent amplitude of the GW. During the matter-dominated

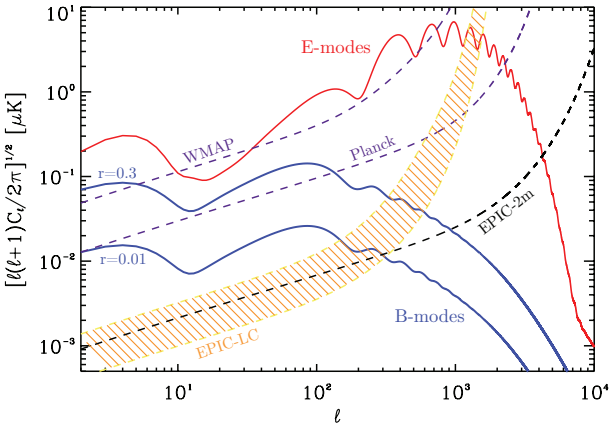


Fig. 6: E- and B-mode power spectra for a tensor-to-scalar ratio saturating the current bounds, $r = 0.3$ and for $r = 0.01$. Shown are the experimental sensitivities of WMAP, Planck and two different realizations of CMBPol (EPIC-LC and EPIC-2m)

era, for example, $h_t = 3j_1(k\tau)/k\tau$: time changes in the metric lead to frequency shifts (or equivalently shifts in the temperature of the black body spectrum). Notice that the angular dependence of this frequency shift is quadrupolar in nature. As a result, the temperature fluctuations induced by this effect as photons travel between successive scatterings before recombination produce a quadrupole intensity distribution, which, through Thomson scattering, lead to polarization. Both E and B power spectra are generated by GW. The current push to improve polarization measurements follows from the fact that density perturbations, to linear order in perturbation theory, cannot create any B-type polarization. As a rough rule of thumb, the amplitude of the peak in the B-mode power spectrum for GW is

$$[\ell(\ell+1)C_{Bl}/2\pi]^{1/2} = 0.024(V^{1/4}/10^{16}\mathrm{GeV})^2\mu\mathrm{K}$$

where

$$V^{1/4} \simeq 6.7\, r^{1/4} \times 10^{16}\,\mathrm{GeV} \tag{98}$$

is the energy scale of inflation. A future experiment like CMBPol [14] can probe values of r as small as 10^{-2}, corresponding to an inflation energy scale of about $2 \times \times 10^{16}$ GeV. Furthermore, using the consistency relation $r = \epsilon$ valid in one-single field models of inflation, one deduces that

$$\frac{\Delta\phi}{m_{\rm Pl}} \simeq \left(\frac{r}{10^{-2}}\right)^{1/2}, \tag{99}$$

meaning that a future measurement of the B-mode of CMB polarization will imply an inflaton excursus of Planckian values. Therefore, A future measurement of the B-mode polarization of the CMB will allow a determination of the value of the energy scale of inflation. This explains the utility of CMB polarization measurements as probes of the physics of inflation. A detection of primordial B-mode polarization would also demonstrate that inflation occurred at a very high energy scale, and that the inflaton traversed a super-Planckian distance in field space.

9 The dark puzzles

Having explored the physics of the primordial epochs of the evolution of the universe, such as inflation, and its impact on the present-day observables, we now devote the remaining space to a short discussion

of the dark puzzles of the present-day universe: the dark energy and the dark matter puzzles.

9.1 A present-day accelerating universe

In 1998 the accelerated expansion of the universe was pointed out by two groups from the observations of Type Ia Supernova (SN Ia) [15, 16]. Let us see how this came about.

An important concept related to observational tools in an expanding background is associated with the definition of a distance. A way of defining a distance is through the luminosity of a stellar object. The distance d_L known as the luminosity distance, plays a very important role in astronomy including in supernovae observations. It proves to be convenient to write the metric as

$$ds^2 = -dt^2 + a^2(t)\left[dr^2 + f_K^2(r)(d\theta^2 + \sin^2\theta d\phi^2)\right] , \tag{100}$$

where

$$f_K(r) = \begin{cases} \sin r\,, & K = +1\,, \\ r\,, & K = 0\,, \\ \sinh r\,, & K = -1\,. \end{cases} \tag{101}$$

In Minkowski space time the absolute luminosity L_s of the source and the energy flux $\mathcal{F}$ at a distance d is related through $\mathcal{F} = L_s/(4\pi d^2)$. By generalizing this to an expanding universe, the luminosity distance, d_L, is defined as

$$d_L^2 \equiv \frac{L_s}{4\pi\mathcal{F}}\,. \tag{102}$$

Let us consider an object with absolute luminosity L_s located at a co-moving coordinate distance r from an observer at $r = 0$. The energy of light emitted from the object with time interval Δt_e is denoted as ΔE_e, whereas the energy which reaches at the sphere with radius r is written as ΔE_r. We note that ΔE_e and ΔE_r are proportional to the frequencies of light at r and $r = 0$, respectively, i.e., $\Delta E_e \propto \nu_e$ and $\Delta E_r \propto \nu_r$. The luminosities L_r and L_e are given by

$$L_r = \frac{\Delta E_e}{\Delta t_e}\,, \quad L_e = \frac{\Delta E_e}{\Delta t_e}. \tag{103}$$

The speed of light is given by $c = \nu_e\lambda_e = \nu_r\lambda_r$, where λ_e and λ_r are the wavelengths at r and $r = 0$. Then, we find

$$\frac{\lambda_r}{\lambda_e} = \frac{\nu_e}{\nu_r} = \frac{\Delta t_r}{\Delta t_e} = \frac{\Delta E_e}{\Delta E_r} = 1 + z, \tag{104}$$

where we have also used $\nu_r\Delta t_r = \nu_e\Delta t_e$. Combining Eq. (103) with Eq. (104), we obtain

$$L_e = L_r(1 + z)^2. \tag{105}$$

The light travelling along the r direction satisfies the geodesic equation $ds^2 = -dt^2 + a^2(t)dr^2 = 0$. We then obtain

$$r = \int_0^r dr' = \int_{t_e}^{t_r} \frac{dt}{a(t)} \tag{106}$$

From the metric (100) we find that the area of the sphere at $t = t_r$ is given by $S = 4\pi(a_r f_K(r))^2$. Hence the observed energy flux is

$$\mathcal{F} = \frac{L_r}{4\pi(a_r f_K(r))^2}\,. \tag{107}$$

Substituting Eqs. (106) and (107) for Eq. (102), we obtain the luminosity distance in an expanding universe:

$$d_L = a_r f_K(r)(1 + z). \tag{108}$$

In the flat FRW background with $f_K(r) = r$ we find

$$d_L = \left(\frac{1+z}{H_0}\right)\int_0^z dz' \frac{H_0}{H(z')}. \tag{109}$$

Then the Hubble rate $H(z)$ can be expressed in terms of $d_L(z)$:

$$H(z) = \left\{\frac{d}{dz}\left(\frac{d_L(z)}{1+z}\right)\right\}^{-1}. \tag{110}$$

If we measure the luminosity distance observationally, we can determine the expansion rate of the universe.

The energy density ρ on the right-hand side of Einstein equations includes all components present in the universe, namely, non-relativistic particles, relativistic particles, cosmological constant and so on

$$\rho = \sum_i \rho_i^{(0)}(a/a_0)^{-3(1+w_i)} = \sum_i \rho_i^{(0)}(1+z)^{3(1+w_i)}. \tag{111}$$

Here w_i and $\rho_i^{(0)}$ correspond to the equation of state and the present energy density of each component, respectively. The Hubble parameter takes the convenient form

$$H^2 = H_0^2 \sum_i \Omega_i^{(0)}(1+z)^{3(1+w_i)}, \tag{112}$$

where $\Omega_i^{(0)} \equiv 8\pi G\rho_i^{(0)}/(3H_0^2) = \rho_i^{(0)}/\rho_c^{(0)}$ is the density parameter for an individual component at the present epoch. Hence the luminosity distance in a flat geometry is given by

$$d_L = \frac{(1+z)}{H_0}\int_0^z \frac{dz'}{\sqrt{\sum_i \Omega_i^{(0)}(1+z')^{3(1+w_i)}}}. \tag{113}$$

The direct evidence for the current acceleration of the universe is related to the observation of luminosity distances of high redshift supernovae [15, 16]. The apparent magnitude m of the source with an absolute magnitude M is related to the luminosity distance d_L via the relation [17]

$$m - M = 5\log_{10}\left(\frac{d_L}{\mathrm{Mpc}}\right) + 25\,. \tag{114}$$

This comes from taking the logarithm of Eq. (102) by noting that m and M are related to the logarithms of $\mathcal{F}$ and L_s, respectively. The numerical factors arise because of conventional definitions of m and M in astronomy. Type Ia supernovae (SN Ia) can be observed when white dwarf stars exceed the mass of the Chandrasekhar limit and explode. The belief is that SN Ia are formed in the same way irrespective of where they are in the universe, which means that they have a common absolute magnitude M independent of the redshift z. Thus they can be treated as an ideal standard candle. We can measure the apparent magnitude m and the redshift z observationally, which of course depends upon the objects we observe.

In order to get a feeling of the phenomenon let us consider two supernovae 1992P at low-redshift $z = 0.026$ with $m = 16.08$ and 1997ap at high-redshift $z = 0.83$ with $m = 24.32$ [15]. As we have already mentioned, the luminosity distance is approximately given by $d_L(z) \simeq z/H_0$ for $z \ll 1$. Using the apparent magnitude $m = 16.08$ of 1992P at $z = 0.026$, we find that the absolute magnitude is estimated by $M = -19.09$ from Eq. (114). Here we adopted the value $H_0^{-1} = 2998h^{-1}\,\mathrm{Mpc}$ with $h = 0.72$. Then the luminosity distance of 1997ap is obtained by substituting $m = 24.32$ and $M = -19.09$ for Eq. (114): $H_0 d_L \simeq 1.16$ for $z = 0.83$. From Eq. (113) the theoretical estimate for the luminosity distance in a two-component flat universe is $H_0 d_L \simeq 0.95$ for $\Omega_m^{(0)} \simeq 1$ and $H_0 d_L \simeq 1.23$ for $\Omega_m^{(0)} \simeq 0.3,\ \Omega_{\mathrm{DE}} \simeq 0.7$

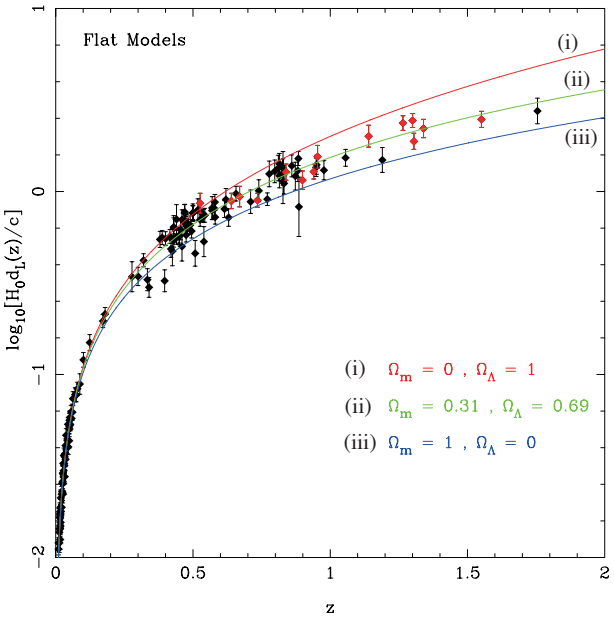


Fig. 7: The luminosity distance versus redshift for a flat cosmological model. Black points are from the "Gold" data sets [18]; red points are from recent data from HST

In 2004 Riess *et al.* [18] reported the measurement of 16 high-redshift SN Ia with redshift $z > 1.25$ with the Hubble Space Telescope (HST). By including 170 previously known SN Ia data points, they showed that the universe exhibited a transition from deceleration to acceleration at $> 99\%$ confidence level. A best-fit value of $\Omega_m^{(0)}$ was found to be $\Omega_m^{(0)} = 0.29^{+0.05}_{-0.03}$ (the error bar is 1σ). This shows that a matter-dominated universe without a cosmological constant does not fit the data.

We should emphasize that the accelerated expansion is by cosmological standards really a late-time phenomenon, starting at a redshift $z \sim 1$. From Eq. (112) the deceleration parameter, $q \equiv -a\ddot{a}/\dot{a}^2$, is given by

$$q(z) = \frac{3}{2}\frac{\sum_i \Omega_i^{(0)}(1+w_i)(1+z)^{3(1+w_i)}}{\sum_i \Omega_i^{(0)}(1+z)^{3(1+w_i)}} - 1.$$

For the two-component flat cosmology, the universe enters an accelerating phase ($q < 0$) for $z < z_c \equiv (2\Omega_{\rm DE}/\Omega_{\rm DM})^{1/3} - 1$. When $\Omega_{\rm DM} = 0.3$ and $\Omega_{\rm DE} = 0.7$, we have $z_c = 0.67$. The problem of why an accelerated expansion should occur now in the long history of the universe is called the "coincidence problem".

9.1.1 The origin of the acceleration

Once the idea of the accelerating universe is accepted, the next pressing question is: Why? There are various explanations available that we may mention briefly. The general trend is to accept that there is a form of Dark Energy (DE) fluid dominating the energy density of the present day. Its pressure is $P = w\rho$ and w needs to be smaller than $-1/3$ for this fluid to cause the acceleration. Having learned how to use scalar fields to accelerate the universe at primordial epochs, the most natural way to explain DE whould be to introduce a scalar field ϕ dubbed quintessence, with potential

$$\mathcal{V}(\phi) = V_0 + V(\phi), \tag{115}$$

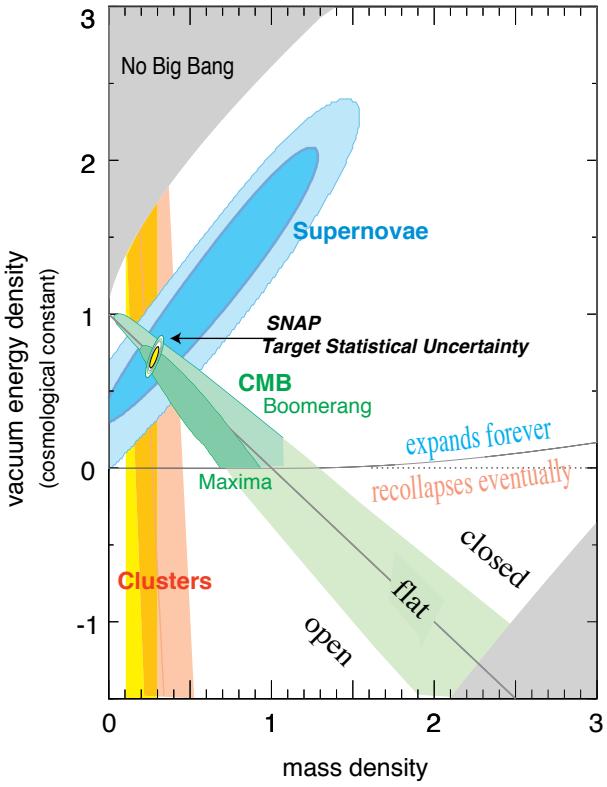


Fig. 8: The dark energy (vacuum energy) and dark matter (mass density) abundances from SN, CMB, and galaxy clustering observations

Now, if $V_0 \gg V(\phi)$ (at least at present epochs), the DE is in practice a Cosmological Constant (CC). Its value must be extremely small, $V_0^{1/4} \simeq (H_0 m_{\rm Pl})^{1/2} \simeq 10^{-3}$ eV. Why it is so small is a mystery that earned the name "the CC problem". On the other hand, if $V_0 \ll V(\phi)$, then the dynamics of the quintessence field dominates. However, another problem arises at this stage. Having learned from inflation that the field must be slow-rolling to cause the acceleration of the universe, we have to assume that $(m_{\rm Pl}^2 V''/V)$ is smaller than unity. This implies that ϕ is of order of the Planck scale and that its mass squared is such that $V'' \sim H_0^2 \sim (10^{-33}\,{\rm eV})^2$. The quintessence field has a Compton wavelegth as large as the entire observed universe.

If the reader does not like all this fine-tuning, there are at least two other explanations for the acceleration of the universe. The first one goes under the name of modified gravity and is in fact rather intuitive. If gravity gets weaker at large distances, objects far from us may recede at a velocity larger than what they would do in the traditional Newtonian gravity case. For this to work, we have to suppose that the gravitational force has a transient at some critical (and cosmological) scale r_c, from the usual $1/r^2$ to, say $1/r^3$. How to get this transition is unfortunately beyond the scope of these lectures. Another alternative goes under the name of the "anthropic principle" and is based on the following point. As we have seen, in a static universe, overdense regions will increase their density at an exponential rate. In an expanding universe, however, there is a competition between the expansion and the gravitational collapse. More rapid expansion, as induced by DE, retards the growth of structure. General relativity provides the following useful relation in linear perturbation theory between the growth factor $g(z)$ and the expansion history of the universe

$$\ddot{g} + 2H\dot{g} = 4\pi G\rho_m = \frac{3\Omega_{\rm DM} H_0^2}{2a^3} g. \tag{116}$$

If the universe is always matter-dominated, then $g \sim a$; however, in a DE dominated universe g scales slower than the scale factor. Now, if the CC is too large, structure does not have time to develop: the initial condition is $\delta\rho_m/\rho_m \sim 10^{-5}$ at the last scattering surface ($z \sim 10^3$) and needs to becomes order unity by now. Now, if we impose that structures might have been able to develop by now even in the presence of a CC, one obtains a reassuring bound, the CC $V_0^{1/4}$ must be smaller than about 10^{-1} eV. In other words, the CC may not be far from the value we observe (if it is non-zero) because otherwise we would not be here to discuss about it. A great deal of observational effort of the next decades will be devoted to understand the cause of the acceleration of the universe [19]. Four observational techniques are currently receiving much attention: 1) Baryoniuc Acoustic Oscillations (BAO) are observed on large-scale surveys of the spatial distribution of matter. They are caused by the same oscillations that left an imprint in the CMB under the form of acoustic peaks. The BAO technique is sensitive to the DE through its effect on the angular-diameter distance vs. redshift relation and through its effect on the time evolution of the expansion rate; 2) Galaxy Cluster (CL) surveys measure the spatial density and distribution of galaxy clusters. The CL technique is sensitive to DE through its effect in the angular-diameter distance vs. redshift relation and through its effect on the time evolution of the expansion rate and the growth rate of perturbations; 3) supernovae as standard candles to determine the luminosity distance vs. redshift relation; 4) Weak Lensing (WL) surveys measure the distortion of background images due to the bending of light as it passes by galaxies or clusters of galaxies. The WL technique is sensitive to DE through its effect on the angular-diameter distance vs. redshift relation and the growth rate of perturbations. All these techniques will not only shed light on the nature of DE, but will also help us to discriminate the various possibilities to explain the present-day acceleration. For instance, the modified gravity scenario predicts a growth function which is different from the one predicted in a CC dominated universe. Future applications of the techniques briefly summarized above should be able to determine which scenario is more likely.

9.2 Dark matter

The evidence that 95% of the mass of galaxies and clusters is made of some unknown component of Dark Matter (DM) comes from (i) rotation curves (out to tens of kpc), (ii) gravitational lensing (out to 200 kpc), and (iii) hot gas in clusters. They lead us to believe that DM makes up about 30% of the entire energy of the universe. A nice review about DM can be found in Ref. [20].

In the 1970s, Ford and Rubin discovered that rotation curves of galaxies are flat. The velocities of objects (stars or gas) orbiting the centres of galaxies, rather than decreasing as a function of the distance from the galactic centres as had been expected, remain constant out to very large radii. Similar observations of flat rotation curves have now been found for all galaxies studied, including our Milky Way. The simplest explanation is that galaxies contain far more mass than can be explained by the bright stellar objects residing in galactic disks. This mass provides the force to speed up the orbits. To explain the data, galaxies must have enormous dark haloes made of unknown matter. Indeed, more than 95% of the mass of galaxies consists of dark matter. The baryonic matter which accounts for the gas and disk cannot alone explain the galactic rotation curve. However, adding a DM halo allows a good fit to data.

The limitations of rotation curves are that one can only look out as far as there is light or neutral hydrogen (21 cm), namely to distances of tens of kpc. Thus one can see the beginnings of DM haloes, but cannot trace where most of the DM is. The lensing experiments discussed in the next section go beyond these limitations.

Einstein's theory of General Relativity predicts that mass bends, or lenses, light. This effect can be used to gravitationally ascertain the existence of mass even when it emits no light. Lensing measurements confirm the existence of enormous quantities of DM both in galaxies and in clusters of galaxies. Observations are made of distant bright objects such as galaxies or quasars. As the result of intervening matter, the light from these distant objects is bent towards the regions of large mass. Hence there may be multiple images of the distant objects, or, if these images cannot be individually resolved, the back-

ground object may appear brighter. Some of these images may be distorted or sheared. The Sloan Digital Sky Survey used weak lensing (statistical studies of lensed galaxies) to conclude that galaxies, including the Milky Way, are even larger and more massive than previously thought, and require even more DM out to great distances. Again, the predominance of DM in galaxies is observed. The key success of the lensing of DM to date is the evidence that DM is seen out to much larger distances than could be probed by rotation curves: the DM is seen in galaxies out to 200 kpc from the centres of galaxies, in agreement with N-body simulations. On even larger Mpc scales, there is evidence for DM in filaments (the cosmic web). Another piece of gravitational evidence for DM is the hot gas in clusters. The X-ray data indicates the presence of hot gas. The existence of this gas in the cluster can only be explained by a large DM component that provides the potential well to hold on to the gas. In summary, the evidence is overwhelming for the existence of an unknown component of DM that comprises 95% of the mass in galaxies and clusters.

There is another basic reason why DM is necessary: to form structures as we observe them. Let us assume that the matter content of the universe is dominated by a pressureless and self-gravitating fluid. This approximation holds if we are dealing with the evolution of the perturbations in the DM component or in case we are dealing with structures whose size is much larger than the typical Jeans scale length of baryons. Let us also define $\mathbf{x}$ to be the co-moving coordinate and $\mathbf{r} = a(t)\mathbf{x}$ the proper coordinate, $a(t)$ being the cosmic expansion factor. Furthermore, if $\mathbf{v} = \dot{\mathbf{r}}$ is the physical velocity, then $\mathbf{v} = \dot{a}\mathbf{x} + \mathbf{u}$, where the first term describes the Hubble flow, while the second term, $\mathbf{u} = a(t)\dot{\mathbf{x}}$, gives the peculiar velocity of a fluid element which moves in an expanding background.

In this case the equations that regulate the Newtonian description of the evolution of density perturbations are the continuity equation:

$$\frac{\partial \delta}{\partial t} + \nabla \cdot [(1+\delta)\mathbf{u}] = 0\,, \tag{117}$$

which gives the mass conservation, the Euler equation

$$\frac{\partial \mathbf{u}}{\partial t} + 2H(t)\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\frac{\nabla\phi}{a^2}\,, \tag{118}$$

which gives the relation between the acceleration of the fluid element and the gravitational force, and the Poisson equation

$$\nabla^2\phi = 4\pi G\bar{\rho}a^2\delta \tag{119}$$

which specifies the Newtonian nature of the gravitational force. In the above equations, ∇ is the gradient computed with respect to the co-moving coordinate $\mathbf{x}$, $\phi(\mathbf{x})$ describes the fluctuations of the gravitational potential, and $H(t) = \dot{a}/a$ is the Hubble parameter at the time t. Its time-dependence is given by $H(t) = E(t)H_0$, where

$$E(z) = [(1+z)^3\Omega_m + (1+z)^2(1-\Omega_m-\Omega_{DE}) + (1+z)^{3(1+w)}\Omega_{DE}]^{1/2}. \tag{120}$$

In the case of small perturbations, these equations can be linearized by neglecting all the terms which are of second order in the fields δ and $\mathbf{u}$. In this case, using the Euler equation to eliminate the term $\partial\mathbf{u}/\partial$t, and using the Poisson equation to eliminate $\nabla^2\phi$, one ends up with

$$\frac{\partial^2\delta}{\partial t^2} + 2H(t)\frac{\partial\delta}{\partial t} - 4\pi G\bar{\rho}\delta = 0\,. \tag{121}$$

This equation describes the Jeans instability of a pressureless fluid, with the additional "Hubble drag" term $2H(t)\partial\delta/\partial t$, which describes the counter-action of the expanding background on the perturbation growth. Its effect is to prevent the exponential growth of the gravitational instability taking place in a non-expanding background. The solution of the above equation can be cast in the form:

$$\delta(\mathbf{x},t) = \delta_+(\mathbf{x},t_i)D_+(t) + \delta_-(\mathbf{x},t_i)D_-(t)\,, \tag{122}$$

where D_+ and D_- describe the growing and decaying modes of the density perturbation, respectively. In the case of an Einstein–de-Sitter (EdS) universe ($\Omega_m = 1$, $\Omega_{DE} = 0$), it is $H(t) = 2/(3t)$, so that $D_+(t) = (t/t_i)^{2/3}$ and $D_-(t) = (t/t_i)^{-1}$. The fact that $D_+(t) \propto a(t)$ for an EdS universe should not be surprising. Indeed, the dynamical time-scale for the collapse of a perturbation of uniform density ρ is $t_{\rm dyn} \propto (G\rho)^{-1//2}$, while the expansion time-scale for the EdS model is $t_{\rm exp} \propto (G\bar{\rho})^{-1//2}$, where $\bar{\rho}$ is the mean cosmic density. Since for a linear (small) perturbation it is $\rho \simeq \bar{\rho}$, then $t_{\rm dyn} \sim t_{\rm exp}$, thus showing that the cosmic expansion and the perturbation evolution take place at the same pace. This argument also leads to understanding the behaviour for a $\Omega_m < 1$ model. In this case, the expansion time scale becomes shorter than the above one at the redshift at which the universe recognizes that $\Omega_m < 1$. This happens at $1 + z \simeq \Omega_m^{-1/3}$ or at $1 + z \simeq \Omega_m^{-1}$ in the presence or absence of a cosmological constant term, respectively. Therefore, after this redshift, cosmic expansion takes place at a quicker pace than gravitational instability, with the result that the perturbation growth is frozen.

The exact expression for the growing model of perturbations is given by

$$D_+(z) \, = \, \frac{5}{2} \, \Omega_m E(z) \int_z^\infty \frac{1+z'}{E(z')^3} \, dz'. \tag{123}$$

The EdS has the faster evolution, while the slowing down of the perturbation growth is more apparent for the open low-density model, the presence of a cosmological constant providing an intermediate degree of evolution. The key point is, however, that a pressureless fluid such as DM is needed for the perturbations to grow to give rise to collapsed objects. Baryon perturbations, being coupled to photons till the last-scattering epoch, feel a non-vanishing pressure and therefore they may not grow. After the last-scattering stage, the baryons fall into the gravitational potential generated by DM and the baryonic perturbations may promptly catch up with those of DM.

9.2.1 Dark matter candidates

There is a plethora of dark matter candidates. MACHOs, or Massive Compact Halo Objects, are made of ordinary matter in the form of faint stars or stellar remnants; they could also be primordial black holes or mirror matter. However, there are not enough of them to completely resolve the question. Of the non-baryonic candidates, the most popular are the WIMPS (Weakly Interacting Massive Particles) and the axions, as these particles have been proposed for other reasons in particle physics. Ordinary massive neutrinos are too light to be cosmologically significant, though sterile neutrinos remain a possibility. Other candidates include primordial black holes, non-thermal WIMPzillas, and Kaluza–Klein particles which arise in higher dimensional theories.

About axions, the good news is that cosmologists do not need to "invent" new particles. Two candidates already exist in particle physics for other reasons: axions and WIMPs. Axions with masses in the range $10^{-(3-6)}$ eV arise in the Peccei–Quinn solution to the strong-CP problem in the theory of strong interactions.

WIMPs are also natural dark matter candidates from particle physics. These particles, if present in thermal abundances in the early universe, annihilate with one another so that a predictable number of them remain today. The relic density of these particles comes out to be the right value:

$$\Omega_{\rm DM} h^2 = (3 \times 10^{-26} {\rm cm}^3/{\rm s})/\langle \sigma v \rangle_{\rm A} \tag{124}$$

where the annihilation cross-section $\langle \sigma v \rangle_{\rm A}$ of weak interaction strength automatically gives the right answer. The reason why the final abundance is inversely proportional to the annihilation cross-section is rather clear: the larger the annihilation cross-section, the more WIMPs annihilate and the fewer of them are left behind. Furthermore, annihilation is not eternal: owing to the expansion of the universe, annihilation stops when its rate becomes smaller than the expansion rate of the universe. When this happens, the abundance is said to freeze-out.

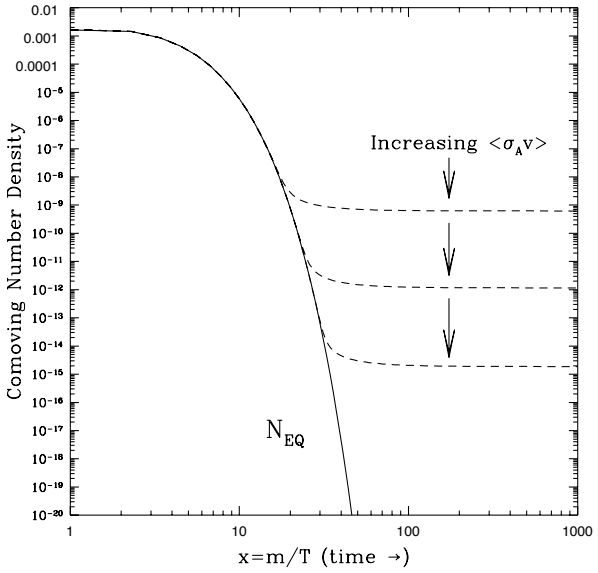


Fig. 9: The abundance of WIMPs of a given mass m as a function of temperature and for various annihilation cross-sections

This coincidence is known as 'the WIMP miracle' and is the reason why WIMPs are taken so seriously as DM candidates. The best WIMP candidate is motivated by Supersymmetry (SUSY): the lightest neutralino in the Minimal Supersymmetric Standard Model. Supersymmetry in particle theory is designed to keep particle masses at the right value. As a consequence, each particle we know has a partner: the photino is the partner of the photon, the squark is the quark's partner, and the selectron is the partner of the electron. The lightest superysmmetric partner is a good dark matter candidate.

There are several ways to search for dark WIMPs. SUSY particles may be discovered at the LHC as missing energy in an event. In that case one knows that the particles live long enough to escape the detector, but it will still be unclear whether they are long-lived enough to be the dark matter. Thus complementary astrophysical experiments are needed. In direct detection experiments, the WIMP scatters off a nucleus in the detector, and a number of experimental signatures of the interaction can be detected. In indirect detection experiments, neutrinos that arise as annihilation products of captured WIMPs exit from the Sun and can be detected on Earth. Another way to detect WIMPs is to look for anomalous cosmic rays from the Galactic Halo: WIMPs in the Halo can annihilate with one another to give rise to antiprotons, positrons, or neutrinos. In addition, neutrinos, gamma rays, and radio waves may be detected as WIMP annihilation products from the Galactic Centre. For lack of time these issues were not discussed extensively in the lectures. The interested reader may find more about these issues in Ref. [20].

10 Conclusions

The period when we say that cosmology is entering a golden age has already passed: cosmology *is* in the middle of its golden age. Present observational data pose various puzzles whose solutions might either be around the corner or decades far in the future. It will require some young and creative researcher sitting in this room to solve them. This is why the cosmological puzzles are dark, but the future is brighter.

Acknowledgements

It is a great pleasure to thank all the organizers, N. Ellis, E. Lillistol, D. Metral, and especially M. Losada and E. Nardi, for having created such a stimulating atmosphere. All students are also acknowledged for their never-ending enthusiasm.

References

[1] A. D. Linde, *Particle Physics and Inflationary Cosmology* (Harwood, Chur, Switzerland, 1990) and references therein.

[2] E. W. Kolb and M. S. Turner, *The Early Universe* (Addison-Wesley, Redwood City, CA, 1990).

[3] A. R. Liddle and D. H. Lyth, Phys. Rep. **231**, 1 (1993) and references therein.

[4] A. R. Liddle and D. H. Lyth, *Cosmological Inflation and Large-Scale Structure* (Cambridge University Press, 2000) and references therein.

[5] D. H. Lyth and A. Riotto, Phys. Rep. **314**, 1 (1999) and references therein.

[6] A. Riotto, arXiv:hep-ph/0210162 and references therein.

[7] For a review, see N. Bartolo, S. Matarrese, and A. Riotto, arXiv:astro-ph/0703496 and references therein.

[8] E. J. Copeland, M. Sami, and S. Tsujikawa, Int. J. Mod. Phys. D **15**, 1753 (2006) and references therein.

[9] E. Komatsu *et al.* [WMAP Collaboration], Astrophys. J. Suppl. **180**, 330 (2009) [arXiv:0803.0547 [astro-ph]].

[10] For a review, see, for instance, N. A. Bahcall, J. P. Ostriker, S. Perlmutter, and P. J. Steinhardt, Science **284**, 1481 (1999).

[11] R. K. Sachs and A. M. Wolfe, Astrophys. J. **147**, 73 (1967).

[12] S. Dodelson, *Modern Cosmology* (Academic Press, New York, 2003).

[13] V. F. Mukhanov, H. A. Feldman, and R. H. Brandenberger, Phys. Rep. **215**, 203 (1992) and references therein.

[14] S. Dodelson *et al.*, CMBPol Science White Paper submitted to the US Astro2010 Decadal Survey, arXiv:0902.3796v1.

[15] S. Perlmutter *et al.*, Astrophys. J. **517**, 565 (1999).

[16] A. G. Riess *et al.*, Astron. J. **116**, 1009 (1998); Astron. J. **117**, 707 (1999).

[17] T. Padmanabhan, Phys. Rep. **380**, 235 (2003); T. Padmanabhan, Current Science, **88**, 1057 (2005) [arXiv:astro-ph/0510492].

[18] A. G. Riess *et al.* [Supernova Search Team Collaboration], Astrophys. J. **607**, 665 (2004).

[19] A. J. Albrecht *et al.*, arXiv:astro-ph/0609591.

[20] G. Bertone, D. Hooper, and J. Silk, Phys. Rep. **405**, 279 (2005).

High-energy astroparticle physics

D. Semikoz
APC, Paris, France

Abstract

In these three lectures I discuss the present status of high-energy astroparticle physics including Ultra-High-Energy Cosmic Rays (UHECR), high-energy gamma rays, and neutrinos. The first lecture is devoted to ultra-high-energy cosmic rays. After a brief introduction to UHECR I discuss the acceleration of charged particles to highest energies in the astrophysical objects, their propagation in the intergalactic space, recent observational results by the Auger and HiRes experiments, anisotropies of UHECR arrival directions, and secondary gamma rays produced by UHECR. In the second lecture I review recent results on TeV gamma rays. After a short introduction to detection techniques, I discuss recent exciting results of the H.E.S.S., MAGIC, and Milagro experiments on the point-like and diffuse sources of TeV gamma rays. A special section is devoted to the detection of extragalactic magnetic fields with TeV gamma-ray measurements. Finally, in the third lecture I discuss Ultra-High-Energy (UHE) neutrinos. I review three different UHE neutrino detection techniques and show the present status of searches for diffuse neutrino flux and point sources of neutrinos.

1 Ultra-high-energy cosmic rays

1.1 Introduction

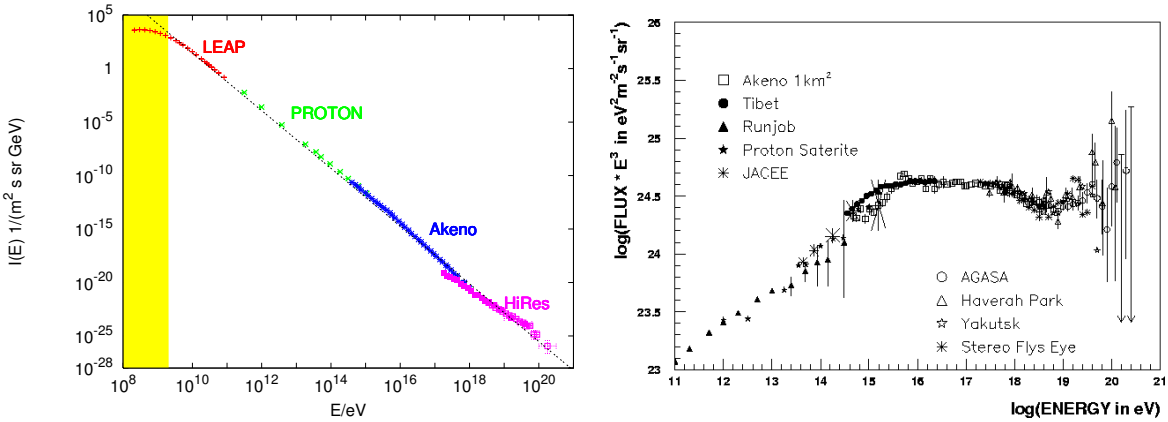


Fig. 1: Left: The cosmic ray spectrum $I(E)$ as function of kinetic energy E, compiled using results from the LEAP, proton, Akeno, and HiRes experiments [1,2]. The energy region influenced by the Sun is marked in yellow and an $1/E^{2.7}$ power-law is also shown. **Right:** The same spectrum at high energies $E > 10^{11}$ eV multiplied by E^3 [3]. Spectrum changes are called the 'knee' at 10^{15} eV and the 'ankle' at 10^{19} eV.

Particles coming from space to the atmosphere of the Earth historically were called cosmic rays. Most cosmic rays, however, are not 'rays' or photons, but charged particles, protons and nuclei. Real high-energy gamma rays coming from space to the Earth are only a small fraction of total flux, and they will be discussed in Section 2. The measured spectrum of cosmic rays from 100 GeV to highest energies $E > 10^{20}$ eV is presented in Fig. 1 (left). The yellow strip at low energies presents the contribution of the Sun. The remaining spectrum can be fitted with a single power law $1/E^{2.7}$ up to highest energies.

The main contribution to it above 100 GeV gives galactic sources. After multiplication of the spectrum on the energy cube, one can see changes of power law in Fig. 1 (right). At $E > 10^{15}$ eV the spectrum becomes steeper. This change in the spectrum called the 'knee' and associated energy $E = 10^{15}$ eV is the maximum energy up to which galactic sources accelerate cosmic rays. The next change of the spectrum is located at $E = 3 \cdot 10^{18}$ eV and has two possible interpretations. Either this is the place where extragalactic sources start to dominate or it is the result of pair- production energy loss by extragalactic protons (see Section 1.3). At the end of the spectrum there is a cutoff, which was not seen in the old experiments presented in Fig. 1 (right) due to small statistics, but it was observed recently by the HiRes [2] and Auger [4] experiments.

In this lecture I briefly discuss the theory and observations of Ultra-High Energy Cosmic Rays (UHECR), the highest-energy particles measured on Earth with energy $E > 10^{18}$ eV. Such particles, protons and nuclei, can be accelerated in astrophysical objects, propagate through intergalactic space, losing energy in the interactions with Cosmic Microwave Background (CMB). UHECR are charged particles. Therefore they are also deflected in the Galactic and intergalactic magnetic fields on the way from the source to the Earth. For a more detailed introduction to UHECR I recommend recent lectures by M. Kachelriess [5].

There are several important scales commonly used in astroparticle physics. Distance is usually measured in parsecs, $1\ \mathrm{pc} = 3 \cdot 10^{18}$ cm. Corresponding larger units are kiloparsec $1\ \mathrm{kpc} = 10^3\ \mathrm{pc}$ and megaparsec $1\ \mathrm{Mpc} = 10^6\ \mathrm{pc}$. Energy at highest energies is usually expressed in units of $\mathrm{EeV} = 10^{18}$ eV.

The plan of this lecture is as follows. In Section 1.2 I shall discuss possible acceleration mechanisms of cosmic rays and astrophysical objects which potentially can be their sources. In Section 1.3 I present the main energy loss processes for UHECR particles and briefly discuss their deflection in the magnetic fields. In Section 1.4 I sum up recent observational results from the Pierre Auger Observatory and other experiments. In Section 1.5 results on anisotropy at highest energy are discussed. In Section 1.6 I review expectations on secondary photons and neutrinos from UHECR protons. Results are summed up in Section 1.7.

1.2 Acceleration

There are several possible acceleration mechanisms that can work in astrophysical objects. These include first-order Fermi acceleration on the shocks in plasma or acceleration in the potential difference, which we call one-shot acceleration below. However, in any case, the Larmor radius of a particle does not exceed the accelerator size, otherwise the particle escapes from the accelerator and cannot gain energy further. This criterion is called the Hillas condition [6] and sets the limit

$$\mathcal{E} \leq \mathcal{E}_{\mathrm{H}} = qBR \tag{1}$$

for the energy $\mathcal{E}$ gained by a particle with charge q in the region of size R with the magnetic field B.

The maximum energy of the accelerated particle can be restricted even more than required by Eq. (1) if one takes into account energy losses during acceleration. Unavoidable losses come from particle emission in the external magnetic field, which can be either synchrotron-dominated if the velocity of the particle is not parallel to the magnetic field, or curvature-dominated in the opposite case.

In Fig. 2 in the plane magnetic field versus acceleration region size, the Hillas condition Eq. (1) is shown by a thick black line. The left figure is for protons and the right one for iron nuclei. Possible acceleration in different astrophysical objects is shown with thin solid figures. Notations are the following: NS are neutron stars, GRB are gamma-ray bursts, BH are black holes, AD are accretion disks, jets are jets in active galaxies, K and HS are knots and hot spots in the jets, L are lobes of radio galaxies, clusters are clusters of galaxies, starbursts are starburst galaxies, voids are voids in large-scale structure. Additional notations in brackets are subtypes of active galaxies: Sy for Seyfert galaxies, BL for BL Lac galaxies and RG for radio galaxies. Only objects above the Hillas line have the potential possibility to

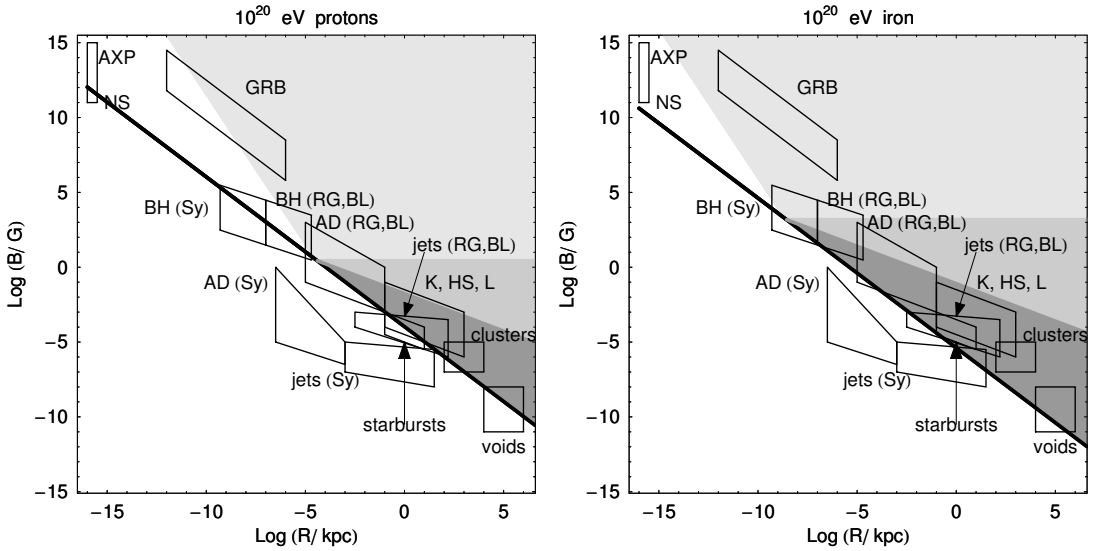


Fig. 2: The Hillas plot with constraints from geometry and radiation losses for 10^{20} eV protons (left) and iron (right). The thick line represents the lower boundary of the area allowed by the Hillas criterion, Eq. (1). Shaded areas are allowed by the radiation-loss constraints as well: light grey corresponds to one-shot acceleration in the curvature-dominated regime only; grey allows also for one-shot acceleration in the synchrotron-dominated regime; dark grey allows for both one-shot and diffusive (e.g., shock) acceleration.

accelerate particles to 10^{20} eV. This is a necessary condition, but not enough for a specific acceleration mechanism. As seen from Fig. 2, for example, neutron stars cannot accelerate particles to highest energies under any condition, while shock acceleration would work only for objects presented in the dark grey corner of this plot.

1.3 Propagation

Owing to expansion of the Universe, particles which come from sources at redshift z lose their energy as

$$E_P \to E'_P = E_P/(1+z)\,. \tag{2}$$

A typical energy loss distance, i.e., distance at which particles lose a significant part of their energy for this process, is of the order of $z \sim 1$ (50% of energy), i.e., $R \sim 3$ Gpc $= 10^{28}$ cm.

As well as during propagation in the intergalactic space, protons lose energy due to two other main processes of interactions with Cosmic Microwave Background (CMB) photons. Those are electron–positron pair production and pion production. In both processes massive particles have to be produced and they have threshold energy. Since the typical energy of CMB photons is very small, $\epsilon_{CMB} = 6 \times 10^{-4}$ eV, the threshold for those processes is very high. Only at energies above $E_{th} = m_e^2/\epsilon_{CMB} \sim 10^{15}$ eV does the electron–positron pair-production process become important:

$$P + \gamma_{CMB} \to P + e^+ + e^-. \tag{3}$$

The typical energy loss distance for this process is

$$R = \frac{M_P}{2m_e}\frac{1}{\sigma_{Pe^+e^-} n_{CMB}} = 600\ \text{Mpc} = 2 \times 10^{27}\ \text{cm}\,, \tag{4}$$

where $n_{CMB} = 400/\text{cm}^3$ is the density of CMB photons, $\sigma_{Pe^+e^-} \approx 10^{-27}/\text{cm}^2$ is the proton-pair production cross-section. The factor $M_P/2m_e$ comes here from the fact that in every interaction a proton loses only a tiny fraction of its energy proportional to the proton/electron mass ratio.

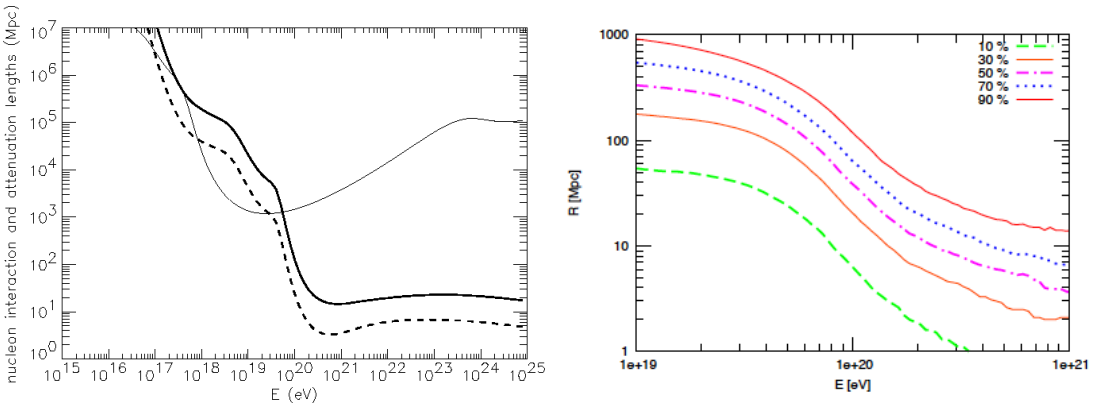


Fig. 3: Left: Nucleon interaction length as function of energy from Ref. [7]. Attenuation due to pion production Eq. (5) presented by the thick solid line, the same for pair production Eq. (3) presented by the thin solid line. **Right:** Horizon (maximal distance) from which protons with given or higher energy can arrive. Lines for 10%, 30%, 50%, 70% and 90% of events are shown [8].

At energies above threshold $E_{th} \approx m_\pi M_P/\epsilon_{CMB} = 10^{20}$ eV, the pion production process dominates energy losses. This process for cosmic rays was first considered by Greizen, Zatsepin, and Kuzmin in 1966 [9] and is now named the GZK process.

$$P + \gamma_{CMB} \to \left\{ \begin{array}{l} P + \pi^0 + \sum_i \pi_i \\ N + \pi^+ + \sum_i \pi_i \end{array} \right. \tag{5}$$

The typical energy loss distance for this process is

$$R = \frac{M_P}{m_\pi} \frac{1}{\sigma_{P\gamma} n_{CMB}} = 100 \text{ Mpc} = 3 \times 10^{26} \text{ cm} \,, \tag{6}$$

where $\sigma_{P\gamma} \approx 6 \times 10^{-28}/\text{cm}^2$ is the proton pion production cross-section. The factor M_P/m_π comes here from the fact that in every interaction the proton loses only 15–20% of its energy proportional to the proton/pion mass ratio. Note that at higher energies the dominating process in Eq. (5) is multi-pion production, in which the proton loses 50% of its energy in every interaction, however, the cross-section for this process $\sigma_{\sum \pi} = 10^{-28}/\text{cm}^2$ is a factor 6 lower than single pion production.

None of the above processes allows a proton with high energy to come from a very large distance. The distance from which protons can come as a function of energy is presented in Fig. 3 (left) [7]. The interaction length for pion production [Eq. (5)] is shown by the dashed line. Attenuation due to pion production [Eq. (6)] is presented by the thick solid line, the same for pair production [Eq. (3)] is presented by the thin solid line. Figure 3 (left) shows the average distance travelled by a single particle. However, for searches of UHECR sources the important question is the maximum distance or horizon from which UHECR can come to the detector. In Fig. 3 (right) we present the horizon as a function of minimal proton energy. The lines 10%, 30%, 50%, 70% and 90% show the fraction of events which come from a given distance. For example, 90% of events with $E > 10^{20}$ eV should come from distances $R < 100$ Mpc. This distance is sometimes called the GZK distance, because energy losses in this case are dominated by the GZK process of Eq. (5).

The dominant loss process for nuclei of energy $E \gtrsim 10^{19}$ eV is photodisintegration $A + \gamma \to (A-1) + N$ in the CMB and the infrared background due to the giant dipole resonance [10]. The threshold for this reaction follows from the binding energy per nucleon, ~ 10 MeV. Photo-disintegration leads to a suppression of the flux of nuclei above an energy that varies between 3×10^{19} eV for He and 8×10^{19} eV for Fe.

Fig. 4: Sky map of the UHECR proton deflections for energy $E = 40$ EeV in two different models of Galactic magnetic field from Ref. [11]. The colours show deflections from 0 to 10 degrees.

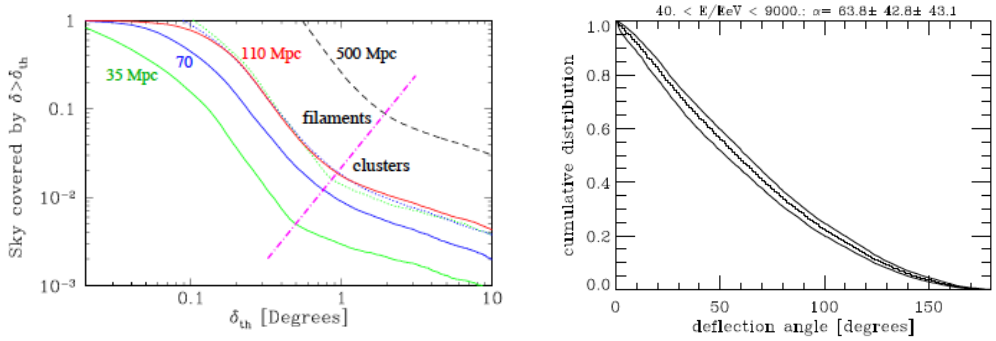


Fig. 5: Fraction of the sky in which the deflection in the extra-Galactic magnetic field is bigger than the given value. **Left:** Constraint simulation of K. Dolag et al. [12]. **Right:** Simulation of Sigl et al. [13].

Since UHECR are charged particles, they not only lose energy in the interactions with background photons, but also when deflected by Galactic and intergalactic magnetic fields.

The magnetic field of the Milky Way galaxy is conventionally modelled as a sum of the regular and turbulent components of the field in the disk and halo of the Galaxy. This means that the deflection in the Galactic field $\theta_{\rm Gal}$ is a superposition of at least four terms:

$$\theta_{\rm Gal} = \theta^{\rm regular}_{\rm Disk} + \theta^{\rm turbulent}_{\rm Disk} + \theta^{\rm regular}_{\rm Halo} + \theta^{\rm turbulent}_{\rm Halo} \ . \tag{7}$$

The deflection angle of UHECR in a regular magnetic field after propagation of distance D is given by:

$$\theta^{\rm regular} \simeq \frac{ZeB_\perp D}{E_{\rm UHECR}} \simeq 5^\circ Z \left[\frac{E_{\rm UHECR}}{4\cdot 10^{19}\ {\rm eV}}\right]^{-1} \left[\frac{B_\perp}{2\cdot 10^{-6}\ {\rm G}}\right] \left[\frac{D}{2\ {\rm kpc}}\right] \ , \tag{8}$$

where where $B_\perp$ is the magnetic field component orthogonal to the line of sight, $E_{\rm UHECR}$ is the particle energy, and Z is the atomic charge. In the case of deflection by the turbulent field on the distance D much larger than the correlation length of the field λ_B and where the deflection angle is small, the deflection is given by

$$\theta^{\rm turb} \simeq \frac{1}{\sqrt{2}} \frac{ZeB_\perp \sqrt{D\lambda_B}}{E_{\rm UHECR}} \simeq 1.2^\circ Z \left[\frac{E_{\rm UHECR}}{4\cdot 10^{19}\ {\rm eV}}\right]^{-1} \left[\frac{B_\perp}{4\cdot 10^{-6}\ {\rm G}}\right] \left[\frac{D}{2\ {\rm kpc}}\right]^{1/2} \left[\frac{\lambda_B}{50\ {\rm pc}}\right]^{1/2} \tag{9}$$

The deflection angles by regular and turbulent components of Galactic Disk and Halo, $\theta^{\rm Disk}_{\rm regular}$ and $\theta^{\rm Disk}_{\rm turbulent}$ are given by Eqs. (8), (9). Contributions of the Halo fields $\theta^{\rm Halo}_{\rm regular}$ and $\theta^{\rm Halo}_{\rm turbulent}$ are less certain, but the result in deflections is usually assumed to be less than the one for disk fields.

Deflections of UHECR by the regular field in the disk $\theta^{\rm regular}_{\rm Disk}$ have been studied in many theoretical models. A sky map of the deflections of UHECR with $E = 40$ EeV in two different models is presented in Fig. 4. Despite both models being consistent on average with the expectation of Eq. (8), predictions in any given direction are strongly model-dependent.

Extragalactic magnetic fields are unknown except in the centres of galaxy clusters. Therefore one has to use theoretical models for the evolution of the magnetic fields. In such models magnetic fields follow the formation of large-scale structures. Because the structure of extragalactic magnetic fields is very non-trivial with very large fields near large-scale structures and tiny fields in the voids, one cannot use Eq. (9) everywhere. Instead one can introduce a fraction of the sky with deflections lower than a given value. Unfortunately, modern models give a very broad range of predictions. In Fig. 5 we show calculations by two groups that show very different results. The group of K. Dolag et al. made constraint simulations of local large-scale structures within 100 Mpc around the Earth [12]. This means that all big structures such as clusters of galaxies are located in exactly the same places as in the real sky. Also the density of points in this simulation is adaptive with more points at clusters and fewer on filaments. The results of this simulation are shown in Fig. 5 (left). According to this simulation, at 100 Mpc from the Earth only in 2% of the sky are deflections bigger than $\theta^{EGMF} = 1^\circ$. Those places are centres of galaxy clusters. In contrast the simulation of G. Sigl et al. [13] uses a uniform grid with more points on filaments and fewer on clusters. Unfortunately this simulation is not constrained and thus cannot be directly compared to local large-scale structures. In this simulation $\theta^{EGMF} > 50^\circ$ in 60% of the sky.

Let us note here that propagation energy losses are extremely important for the understanding of experimental results on spectrum and composition discussed in the next section. Deflections in turn are a key issue in the anisotropy studies in Section 1.5.

1.4 Observations

In this section we shall discuss the present status of UHECR observations.

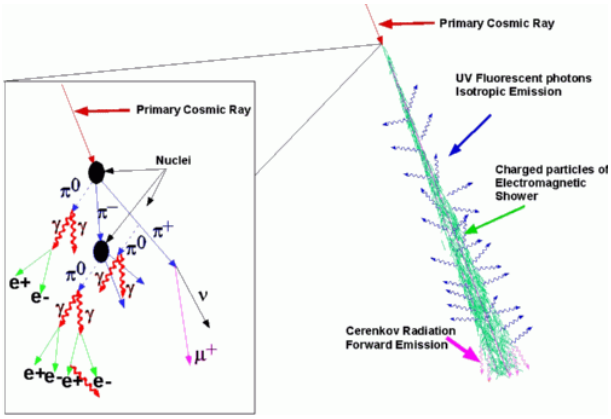


Fig. 6: Extensive air shower produced by a UHECR particle in the atmosphere

We start with the detection of UHECR in the atmosphere. The typical column density of the atmosphere is 1000 g/cm^2 and in 1g there are $N = 10^{24}$ protons, while a strong cross-section is $\sigma_{PP} \sim 10^{-25}$ cm^2. Thus UHECR protons or nuclei should interact within the atmosphere many times before they reach the Earth's surface. In these interactions it would produce extensive air showers. An example of such a shower is illustrated in Fig. 6. After first interaction the primary proton or nuclei would produce a large number of pions. Neutral pions would start an electromagnetic cascade, while charge pions would produce muons. At the maximum of shower development one expects $N = 10^{9-10}$ particles distributed in an area with a radius of a few kilometres. At this point the shower mostly consists of 10 MeV electrons and photons and only 5–10% of its energy is in muons. If the shower is not vertical, the

column density increases as $1/\cos(\theta_{zenith})$ and reaches 2000 g/cm^2 for $\theta_{zenith} = 60°$. At such a depth all electromagnetic components of the shower disappear and the shower would consist of muons only. Another way to detect the shower is to look for fluorescent UV light of nitrogen atoms in the atmosphere. This method is called 'calorimetric' and gives a 3-dimensional image of the shower. The main problem with this method is that detection is possible only on a moonless night, making the duty cycle of such detectors possible only 10–15% of the time. Finally one can detect direct Cherenkov light of the charged particles, but since this light is concentrated only within the central kilometre of the shower, one cannot use this technique at highest energies.

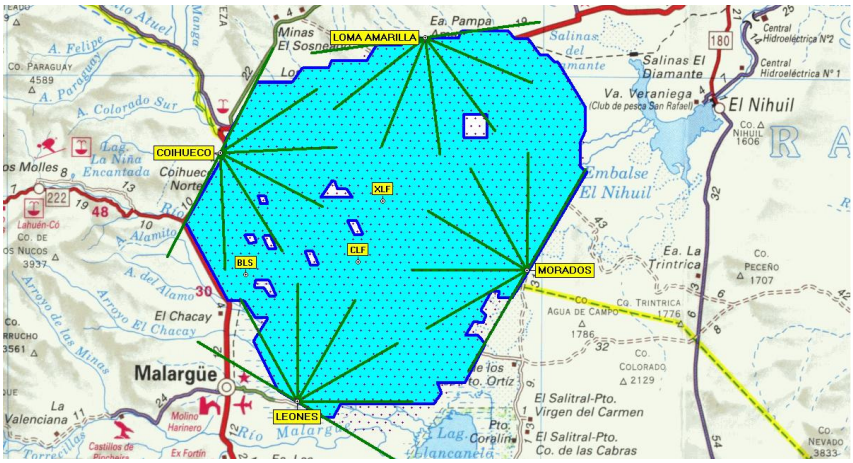


Fig. 7: Pierre Auger Observatory detector with more than 1600 water tanks and 4 fluorescence telescopes (see Ref. [14] for details)

The Pierre Auger Observatory (Auger) is the largest UHECR detector in the world at the moment with an area of 3000 km^2. Such a big area is required to collect enough statistics with UHECR at highest energies $E > 60$ EeV, because the flux of such UHECR is tiny, one particle per 100 km^2 per year. Auger is located on the plateau at an altitude of 1000 metres in the Mendoza province of Argentina. The ground detector consists of 1600 water tanks distributed 1.5 km from each other as presented in Fig. 7. Also there are four fluorescence telescopes pointed at the atmosphere above the ground detectors as shown in Fig. 7. Detection of 10% of showers both by fluorescence detectors (FD) and by ground detectors guarantees a good quality of events and at the same time allows one to calibrate the ground detectors by FD.

In Fig. 8 we show a recent energy spectrum which was measured by Auger before 31 March 2009 [4]. The steeply falling flux of UHECR is multiplied by E^3 in order to show details of the spectrum. The total systematic energy error is 22% and is shown in the top right corner of the figure. For energy bins with $E < 3 \cdot 10^{19}$ eV statistical errors are not important, while at highest energies $E > 60$ EeV the shape of the spectrum is still uncertain and more statistics are needed. On the other hand, the suppression of the spectrum is statistically significant and is clearly seen in Fig. 8.

This is an important experimental result, since it is independent confirmation of similar observations made by the HiRes experiment [2]. Thus cutoff in the energy spectrum exists. However, there are several questions to be answered before one can tell that this really is a GZK cutoff. First, is this cutoff due to the maximum energy of sources, or to energy losses? In Section 1.2 we have seen that indeed the maximum energy for many types of sources is close to 10^{20} eV. The ultimate answer to this question would be the detection of several sources at different distances with cutoff following expectations of energy losses.

Second, is the chemical composition of UHECR proton-dominated at those energies? Since in our Galaxy all elements up to iron are accelerated to energies around the knee $E = 10^{15}$ eV, the same

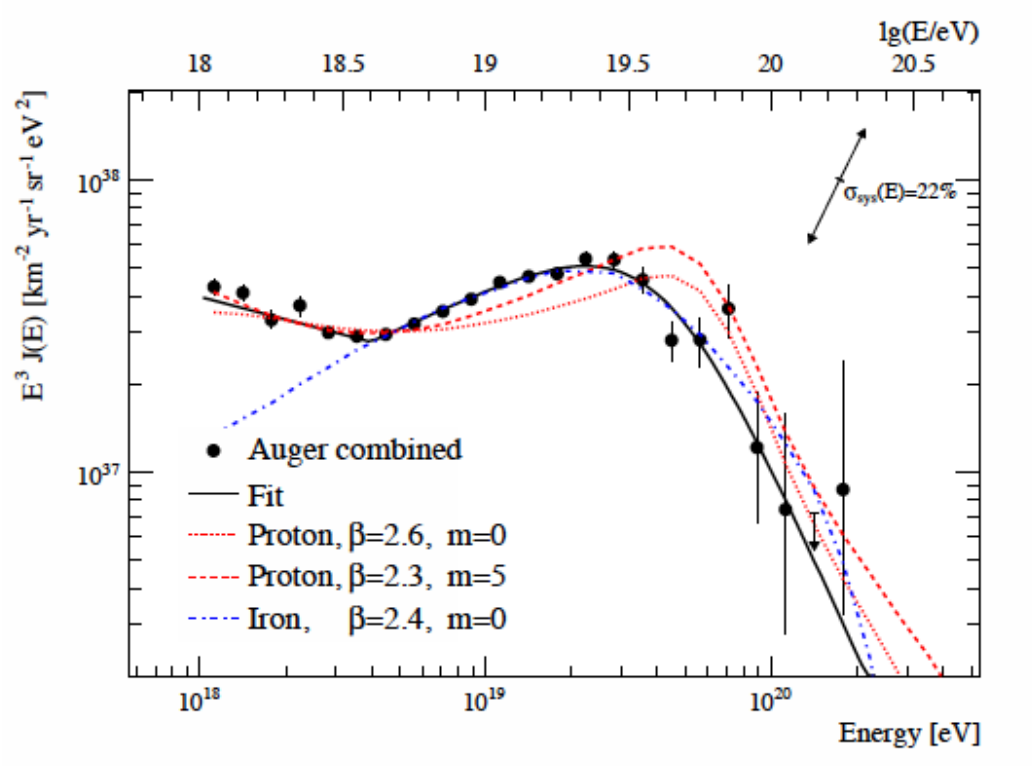


Fig. 8: Energy spectrum of UHECR as a function of energy measured by the Pierre Auger Observatory and model predictions for iron nuclei (blue) and protons (red) [4]

situation can exist in astrophysical objects which accelerate to highest energies. Experimental answers to this question can be found in the future even by Auger, but already current data show that the composition becomes heavy at high energies. Indeed, recent Auger results from Refs. [15, 16] are shown in Fig. 9. In Fig. 9 (left) we present the shape of the shower development in the atmosphere as seen by a fluorescence telescope. The signal is proportional to the number of electrons and positrons in the shower. Signals grow due to the development of electromagnetic cascades. The maximum of the signal corresponds to the maximum development of the cascade in the atmosphere. After that the shower loses its energy due to dissipation effects. The depth of the atmosphere corresponding to the maximum of the shower development is called X_{max}. For the example presented, this maximum is at X_{max} = 753 g/cm^2 and the energy of the event is $E = 1.6 \cdot 10^{19}$ eV [15]. At the same energy, protons on average interact much deeper in the atmosphere than heavy nuclei.

On the top panel of Fig. 9 (right) one can see the results of the most common hadronic models presented with red lines for protons and with blue lines for iron. The example event in Fig. 9 (left) is definitely proton-like. The averaged X_{max} values in each bin are presented in the same figure for both the Auger and HiRes experiments. Both results are consistent with each other showing a relatively light composition from 10^{18} eV to 10^{19} eV. However, Auger shows heavier composition at highest energies.

The main problem when measuring the composition with X_{max} is its strong model dependence, as seen on the top panel of Fig. 9 (right). There are two complementary ways out. One is to use a composition-sensitive parameter that weakly depends on the model choice. Such a parameter is $RMS(X_{max}) = \sqrt{\langle X_{max}^2 \rangle - \langle X_{max} \rangle^2}$, presented in the lower panel of Fig. 9 (right). One can see that according to this measurement the composition becomes heavier at high energy. Another important way is to test models and find the best one. For this purpose a dedicated experiment LHC forward (LHCf) was constructed at CERN. The idea of this experiment is to measure the neutral particles emitted

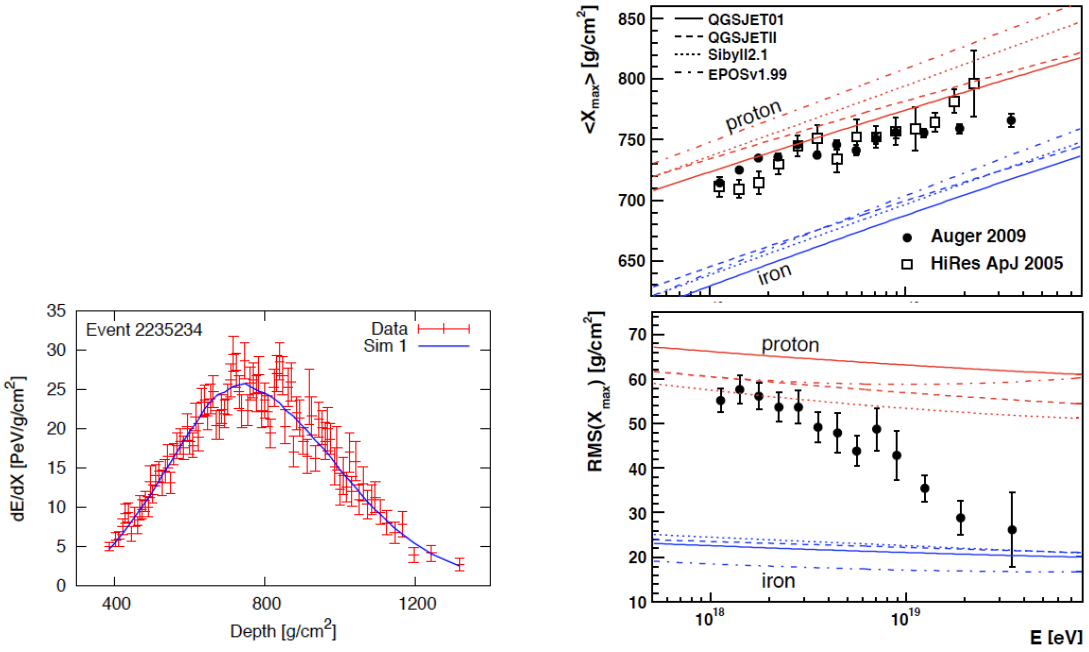


Fig. 9: Left: Measurement of shower development by signals in the fluorescence detectors as a function of depth in the atmosphere. The maximum of shower development in this example is $X_{max} = 753\ g/\mathrm{cm}^2$ [15]. **Right:** Average X_{max} of showers measured by HiRes and Auger and RMS of X_{max} measured by Auger in 2009 [16].

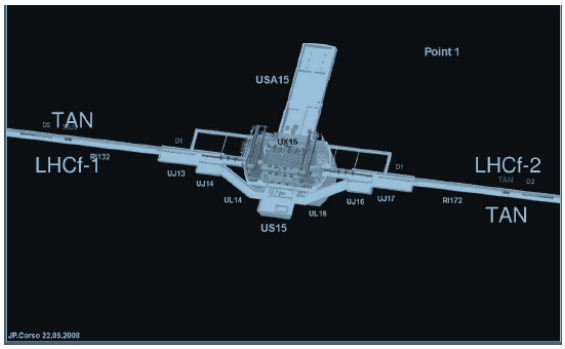


Fig. 10: The layout around the Interaction Point 1 (IP1) of the LHC. The structure at the centre indicates the ATLAS detector surrounding the collision point. The LHCf detectors are installed in the instrumentation slot of the TANs located ± 140 m from IP1. Two independent detectors, LHCf Arm1 and LHCf Arm2 are installed at either side of IP1 [17].

in the very forward region of LHC collisions at low luminosity. The configuration of this experiment is presented in Fig. 10. Data required for testing the hadronic models will be collected in the first scientific runs of the LHC [17]. Thus in the near future we shall have better knowledge of hadronic models and more understanding of the composition of UHECR at highest energy. At present the Auger results indicate heavy composition; this was not confirmed by independent measurements and the fraction of light nuclei in the data remains uncertain.

This is a very important question for searches of UHECR sources, which we shall discuss in the next section.

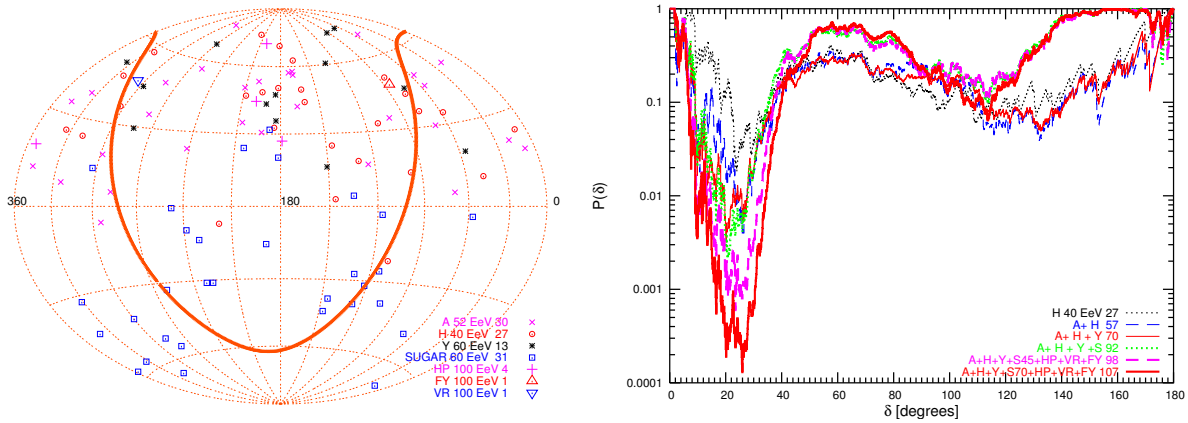


Fig. 11: Left: Sky map of arrival directions of UHECR with $E > 40$ EeV in old experiments. **Right:** Probability that this anisotropy is a function of angular distance between events [18].

1.5 Anisotropy

Since for every UHECR event the arrival direction is detected, for tens of years many attempts were made to find sources of UHECR in the experimental data. Unfortunately none of them has been confirmed so far. There are two ways to look for the sources. One is to look for the data itself and try to find anisotropy in autocorrelation factions. The second is to pick up a catalogue of possible sources and look for the cross-correlations with this catalogue. This second way always requires confirmation by an independent data set, since completeness of the catalogue is a very complicated issue and it is difficult to estimate the probability due to the parameter choice a posteriori.

Here we start with autocorrelations. In the left panel of Fig. 11 one can see the sky map with the arrival direction of events with $E > 40$ EeV in several old experiments, including SUGAR, AGASA, HiRes, Yakutsk, Havera Park, Volcano Ranch, and Fly's Eye. On the right panel of the same figure one can see the probability that autocorrelations between selected events are by chance within a given angle. One can see that the probability is minimal at angles 20–25 degrees. After penalization on the choice of angle, the probability that this happened by chance is $P = 0.3\%$ [18]. This clustering of events on rather moderate scales can be due to the location of the sources in the Large Scale Structure.

The same probability in the first Auger data is presented in Fig. 12 as a function of both energy and angle. One can see that for exactly the same energy $E = 40$ EeV and angle $\theta = 20°$–$25°$ the probability is $P \sim 10^{-2}$. However, recent results with larger statistics did not show more significant anisotropy at such energies [20]. This makes the situation with anisotropy in the data less clear.

Now let us discuss correlations with astrophysical objects. First Auger data have shown strong correlations with nearby active galaxies called Active Galactic Nuclei (AGN). Namely, 12 out of 14 events with E > 57 EeV were correlated within $\theta < 3.1°$ from 472 AGNs from the Veron catalogue with distances $R < 75$ Mpc. This correlation was considered by the Auger Collaboration as a formal way to study the deviation of cosmic rays from isotropic distribution. Data from Period I was tested with the prescription during Period II, where 13 new events were detected, out of which 9 obeyed prescription parameters. The prescription was fulfilled, i.e., the observed sky was considered anisotropic at the 99% confidence level [21]. Data used in this publication and shown in Fig. 13 (left) correspond to Periods I (not shown) and II shown in Fig. 13 (right) before the vertical line. Unfortunately this correlation was not confirmed in the later data [Period III in Fig. 13 (right)].

It does not mean that all anisotropy signals in Auger have completely disappeared. There is still a remaining excess of events around the Cen A galaxy on scales of 20 degrees, see Fig. 14. This anisotropy has to be tested by future data.

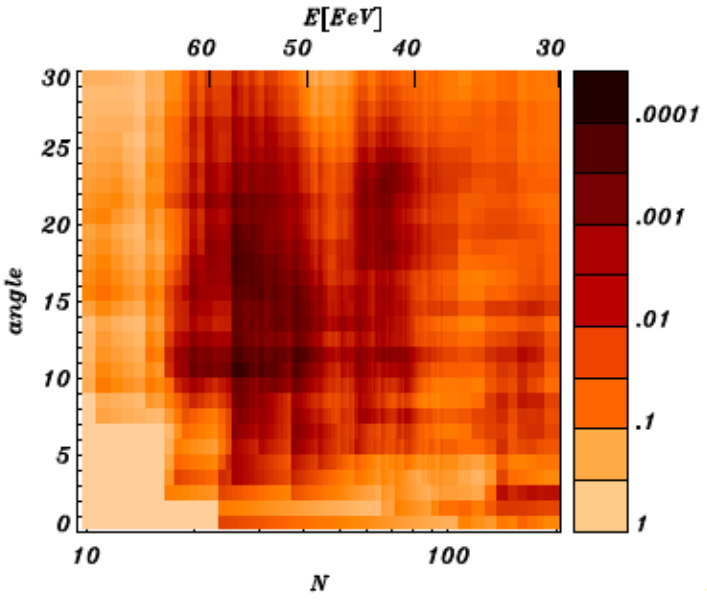


Fig. 12: Probability of autocorrelations as a function of energy and angular distance between events, see Ref. [19]

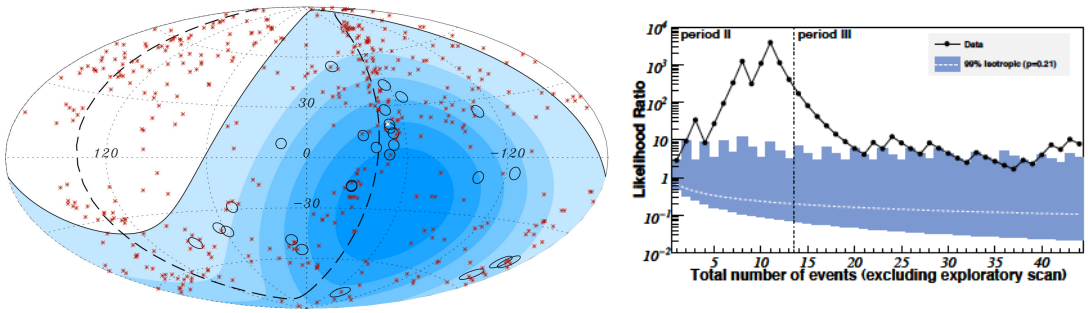


Fig. 13: Left: Sky map of arrival directions of 27 UHECR with $E > 57$ EeV measured by the Pierre Auger Observatory before August 2007 in galactic coordinates (circles) and 472 nearby AGNs (red stars) [21]. Blue contours show the Auger exposure. **Right:** Likelihood ratio for events after formulation of the prescription. Period II is for data on the left panel. Period III is for new data up to March 2009 [20].

1.6 Secondary photons and neutrinos from UHECR

As was discussed in Section 1.3, protons lose their energy in pair production and pair production reactions. Since secondary pions quickly decay, secondary photons and neutrinos are produced. Neutrinos propagate to the Earth without interactions on the way, but photons cannot. They start to interact with background photons and produce pairs. Electrons and positrons in turn up-scatter CMB photons or produce synchrotron radiation:

$$\begin{aligned} \gamma + \gamma_{background} &\rightarrow e^+ + e^- \\ e^\pm + \gamma_{background} &\rightarrow e^\pm + \gamma \\ e^\pm + B &\rightarrow e^\pm + \gamma_{synch} \end{aligned} \qquad (10)$$

The sequence of processes in Eq. (10) is called an electromagnetic cascade. At energies above 10^{15} eV the cascade proceeds on the CMB background (400/cm^3), but at lower energies pair production on CMB is impossible. At such energies the cascade continues on a much less abundant infrared back-

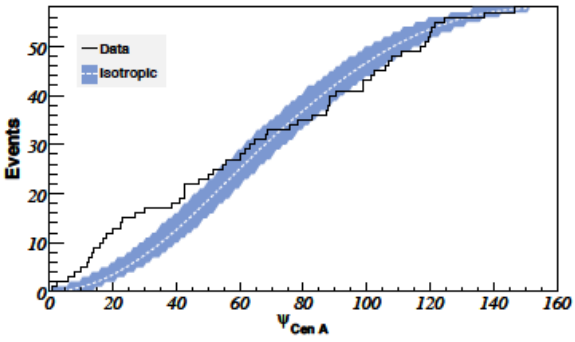


Fig. 14: Angular distribution of events around the Cen A galaxy in Auger data compared to isotropic ones

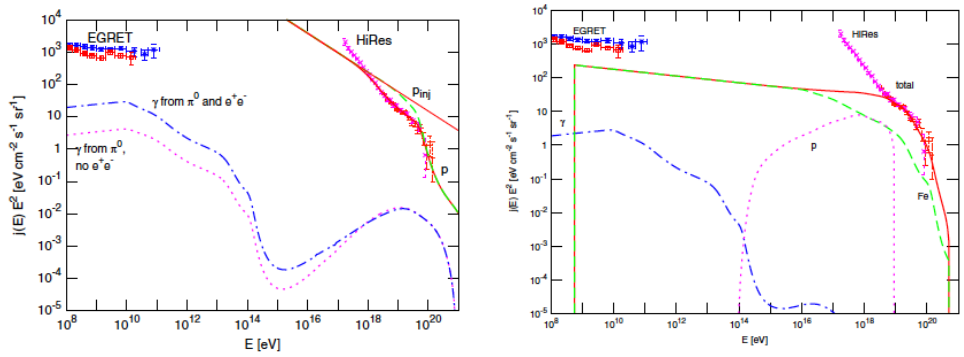


Fig. 15: Left: Fluxes of protons and secondary photons as a function of energy. Primary protons with spectrum $1/E^{2.6}$ and maximum energy $E_{max} = 10^{21}$ eV are shown by the thin red line. Secondary protons fit the UHECR spectrum from $E > 10^{18}$ eV (thick red line). Secondary photons from all reactions are shown by the blue dashed line and from pion production only, Eq. (5) by the magenta line. **Right:** Fluxes of UHECR and secondary photons in the case of iron nuclei primaries with spectrum $1/E^{2.1}$ and maximum energy $E_{max} = 10^{21}$ eV. The remaining iron nuclei are shown by the green line. Secondary protons by the magenta line. Secondary photons by the blue line.

ground (1/cm^3) and at lower energies on optical background (0.01/cm^3). Then it stops at the multi-GeV energies of gamma rays.

In Fig. 15 we plot primary cosmic-ray and secondary photon fluxes from primary protons (left) and iron (right) from Ref. [22]. Secondary protons after interaction fit the UHECR spectrum from $E > 10^{18}$ eV in Fig. 15 (left). Secondary photons cascade down to the GeV region. Only a small fraction of photons come from the pion production reaction (magenta dotted line). Most of the photons generated are from the e^+e^- production reaction with total flux shown by the dash-dotted blue line. The number of secondary protons is much lower in the case of iron primaries, as shown by the dotted magenta line in Fig. 15 (right). As a result, the secondary photon flux in the GeV region is much smaller in this case, on the level of 0.2% of the EGRET measurement. Also very high energy photons are absent in this case due to low maximum proton energy.

In Fig. 16 we compare the range of the electromagnetic cascade fluxes from UHECR with other possible astrophysical contributions in the EGRET band. Note that most of the uncertainty of the UHECR cascade flux comes from an unknown source evolution. The scatter for a given class of sources is thus much smaller, as seen from Fig. 16 for the case of AGNs.

In Fig. 17 (left) we plot the possible range of GZK gamma-ray fluxes for a given proton flux which fit the UHECR spectrum. The range of fluxes comes from the variation of possible values of the

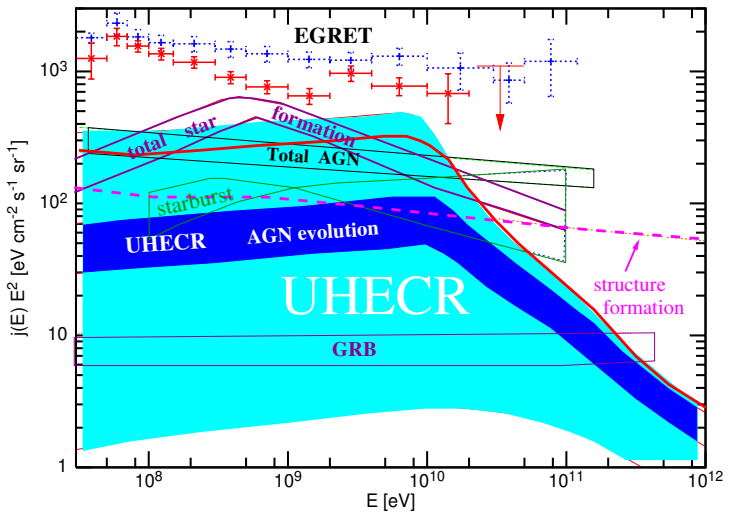


Fig. 16: Contribution of secondary photons from UHECR to the extragalactic gamma-ray background as a function of energy and other possible sources which contribute to the same background [22]

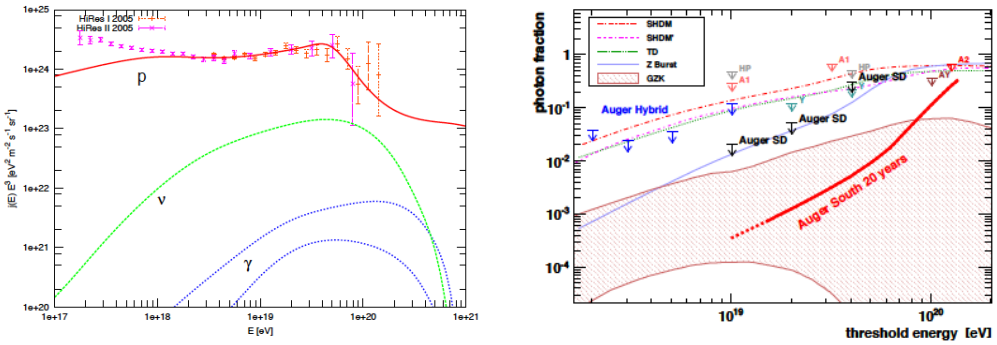


Fig. 17: Left: Example of GZK photon flux from Ref. [23]. UHECR protons fit the HiRes spectrum. Secondary neutrinos are shown by a green line. The remaining secondary photons are in the range between the blue lines. **Right:** Experimental upper limits on the photon fraction in the UHECR spectrum from Ref. [20].

extragalactic magnetic field and the range of the models for extragalactic radio background. Also on the same figure the corresponding neutrino flux is shown by a green line. In Fig. 17 (right) we show the experimental upper limits on the fraction of photons in the UHECR flux. The range of possible GZK photon fluxes corresponds to protons with a range of power law injection spectra and source evolution fitting the UHECR spectrum. One can note that the current best upper limits of Auger are still above the range of expected theoretical values. On the other hand, existing limits already exclude some exotic models.

1.7 Summary

In the first lecture we briefly discussed many aspects of UHECR physics.

Observed cosmic rays have energies up to 10^{20} eV. Acceleration in astrophysical objects to such energies is a very non-trivial task and there are no objects in our Galaxy which can do this job. There are very few classes of exceptionally powerful objects in the Universe, some of which can be real sources of UHECR. Accelerated particles lose their energy in interactions with the CMB background and are also deflected by electromagnetic fields during their propagation from sources to the Earth.

There are three important experimental challenges in UHECR physics: the spectrum of cosmic rays, the chemical composition of cosmic rays, and the search for anisotropies in the sky with the ultimate goal of finding UHECR sources.

The cutoff in the energy spectrum at highest energies $E > 6 \cdot 10^{19}$ eV has now been established by two independent experiments, HiRes and Auger.

The most striking result of 2009 was evidence of heavy composition, shown by the Auger experiment at highest energies, Fig. 9. This result still needs independent confirmation. Also the interpretation of composition measurements is affected by uncertainty in the hadronic models. This question can be clarified in the near future by the LHCf experiment.

Finally, most challenging is the search for UHECR sources. The last result in this direction was made by Auger in 2007. They found that the sky is anisotropic at the highest energies, at least at the 99% C.L., by looking at the correlations with nearby AGNs. Unfortunately those correlations were not confirmed in the new data, and the only anisotropy excess remaining in the Auger data at the highest energies is an excess around the Cen A galaxy, see Fig. 14.

During energy losses the UHECR protons produce secondary photons and neutrinos. Most of the secondary photons cascade down to the GeV energies, where this contributes to the diffuse extragalactic background. An experimental search for the remaining gamma rays at highest energies $E > 10^{18}$ eV is challenging and existing upper limits are just above theoretical predictions, see Fig. 17.

Thus there are many unsolved problems in UHECR physics. They require both theoretical and experimental efforts in the near and more distant future.

2 High-energy gamma rays

2.1 Introduction

In this lecture I shall discuss the theory of TeV gamma rays and recent observations made in this field. I shall give a brief introduction to the experimental detection techniques and present some selected results on the subject. For more detailed study I would like to recommend the recent review by F. Aharonian, J. Buckley, T. Kifune, and G. Sinnis [24].

Relativistic particles can travel with a speed larger than the speed of light in the medium $V > V_M = c/n$. Here $n > 1$ is the refractive index of the medium. This index in the air is $n_{\rm a} = 1.008$ and in water $n_{\rm w} = 1.33$.

The charged particles polarize the molecules of the medium, which then return rapidly to their ground state, emitting prompt radiation called Cherenkov radiation. This radiation is emitted under a constant Cherenkov angle with the particle trajectory, given by

$$\cos\delta = \frac{V_M}{V} = \frac{c}{nV} = \frac{1}{\beta n}\,. \tag{11}$$

The minimal energy of a charged particle is

$$\gamma_{min} = \frac{E_{min}}{M} = \frac{n}{\sqrt{n^2-1}}\,. \tag{12}$$

Particles with higher energy will produce a cone of Cherenkov light. This effect is used by Cherenkov telescopes for air (H.E.S.S., MAGIC, Veritas, CTA) and by ground experiments in water (Milagro, HAWK). Detection of the shower in air and in water is illustrated in Fig. 18.

We present examples of such experiments in Fig. 19. On the left panel we show a view of the H.E.S.S. experiment. This experiment made the most significant contribution to the development of TeV gamma-ray astrophysics in recent years. On the right panel we show the Milagro experiment, a pioneering experiment in water Cherenkov techniques.

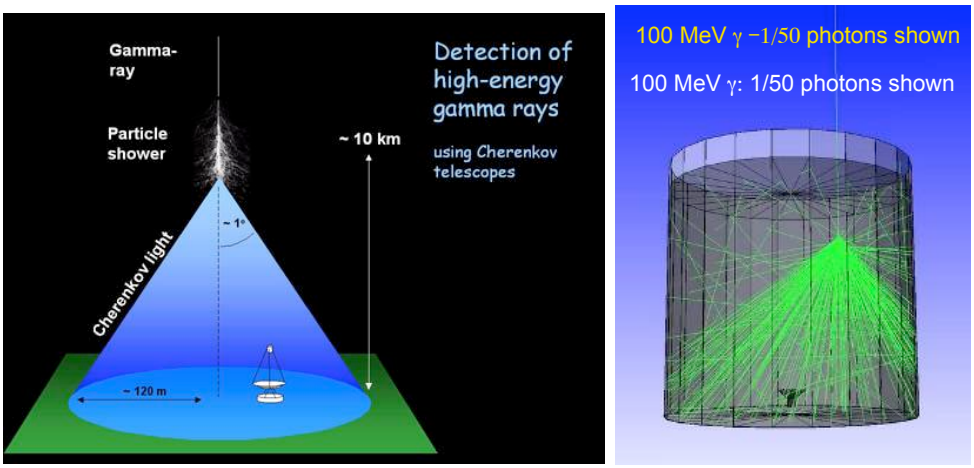


Fig. 18: Detection of high-energy gamma rays by Cherenkov telescopes in air (left) and in water (right)

Fig. 19: Examples of gamma-ray experiments: Cherenkov telescope H.E.S.S. (left) and water pool Milagro (right)

2.2 Point sources of TeV gamma rays

TeV gamma-ray astrophysics is developing very quickly. One can see the number of detected sources in the sky as a function of time in Fig. 20. From 3 sources in 1995 one has 32 sources in 2005 and 80 sources in 2008. In addition not only does the number of observed sources grow, but also the number of different populations of sources. This is a very important fact for future experiments with better sensitivity like the Cherenkov Telescope Array (CTA). They would have a very large potential for detecting many different classes of sources.

In particular, in Fig. 20 on the bottom panel, red circles show extragalactic sources which contain BL Lac objects, radio galaxies, and starburst galaxies. Also in the galactic plane there are many different classes of objects, which include supernova shells, pulsar wind nebulas, pulsars, binary systems and dark objects. Dark objects mean they were detected in gamma rays, but there is no corresponding source in other wavebands.

The sensitivity of gamma-ray detectors to point sources as a function of energy is shown in Fig. 21. The sensitivity of air telescopes is shown for 50 hours of observation for one source. The sensitivity for ground experiments is shown for 5 years, but they observe all the sky (2πsr). At low energies $E < 10$ GeV the sensitivity of the GLAST (Fermi) satellite is the best. One year of observations are shown. At large energies $E > 10$ TeV the ground air-shower experiments (Tibet) have the best sensitivity. Future CTA projects will be orders of magnitude better than present-day experiments from 10 GeV to 10 TeV energies.

Another important fact is that gamma rays cannot travel freely in the intergalactic space. They in-

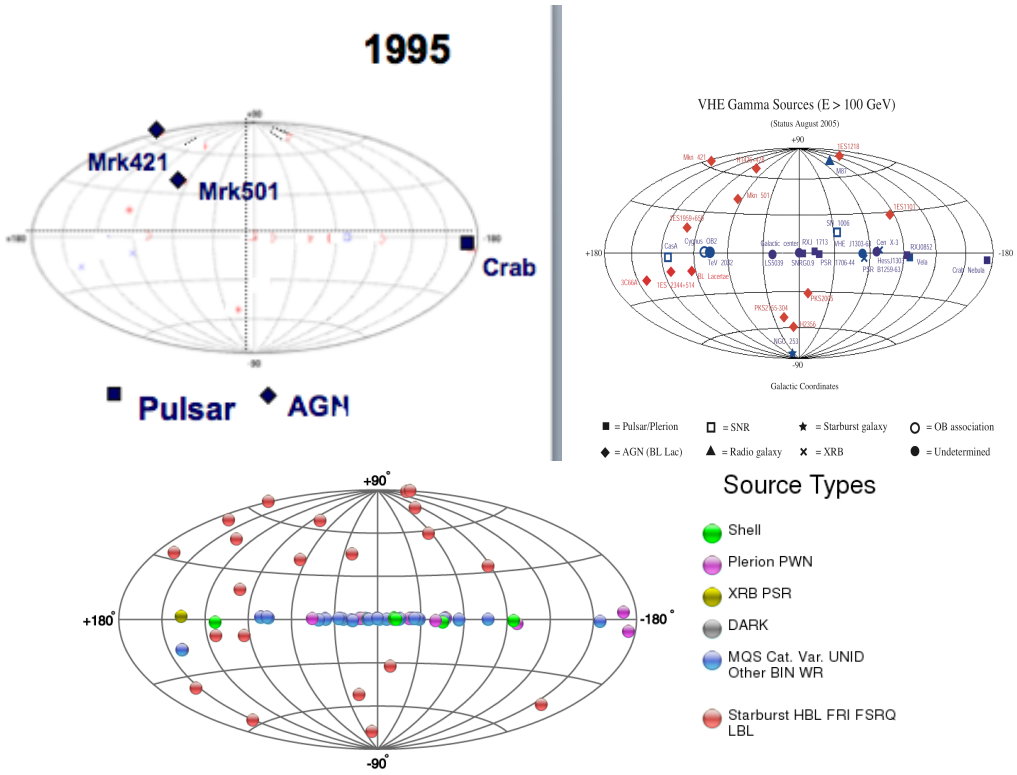


Fig. 20: Sky in the TeV gamma rays with 3 sources in 1995 (top left), 32 sources in 2005 (top right), and 80 sources in 2009 (bottom)

teract with optical/infrared background photons and disappear producing pairs of electrons and positrons. In Fig. 22 one can see the main backgrounds for gamma-ray propagation. They are shown in units of photon density per cm^3. The largest contribution comes from the CMB background with 400 photons per cm^3. However, owing to the small energy of CMB photons, this background is important only for $E > 1000$ TeV. For the experimentally interesting energy range $E < 100$ TeV the main backgrounds are infrared and optical. Since those backgrounds are created by galaxies and partly by dust they are strongly model dependent both as a function of energy and as a function of redshift.

Optical depth can be defined as

$$\tau(E) = R \cdot \sigma_{\gamma\gamma}(E) \cdot n_{back}(z, \epsilon) \, , \tag{13}$$

where R is the distance travelled by photons, $\sigma_{\gamma\gamma}(E)$ is the pair-production cross section, and $n_{back}(z, \epsilon)$ is the density of background photons. Distances on the cosmological scale are often expressed in terms of redshift. One can express it through the Hubble law $R = z \cdot c/H_0$, where $H_0 = 70$ km/s/Mpc is the Hubble constant. In Fig. 23 contours of constant optical depth $\tau(E)$ are shown on the plane redshift versus energy for $\tau(E) = 1, 3, 10$ in two different models of IR/O background.

There is one important difference between air Cherenkov telescopes and water Cherenkov detectors. In Fig. 24 we plot world-wide monitoring of the nearby BL Lac object Mkn 421 as a function of time. One can see that air Cherenkov telescopes can see a signal only on moonless nights, which restricts their operation to the corresponding intervals of time. On the contrary, water Cherenkov telescopes operate all the time they can see a source, which will allow source activity to be detected all the time. On the other hand, the problem of water Cherenkov experiments is poor sensitivity, which will prevent them from detection of relatively low fluxes and very fast variations in time. Thus both techniques are complementary to each other.

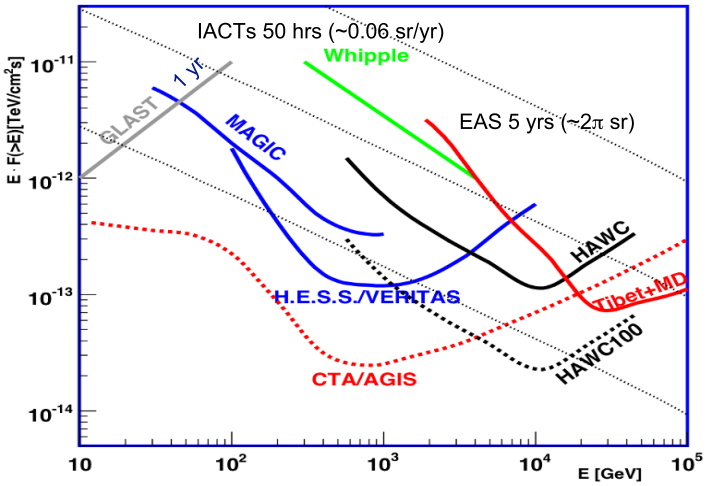


Fig. 21: Sensitivity of gamma-ray detectors to point sources, from Ref. [24]

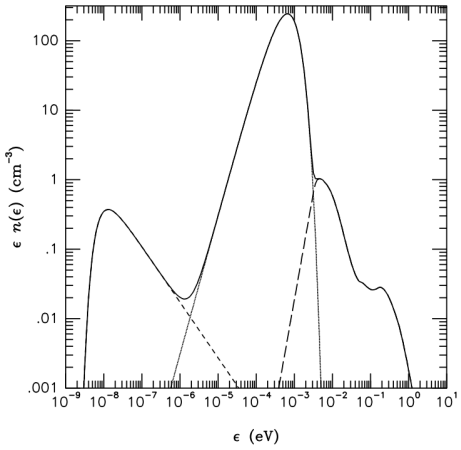


Fig. 22: Redshift for gamma rays as a function of energy. Lines show constant optical depth in two models of IR/O background.

In Fig. 25 one can see a view of the central part of the Milky Way galaxy in three energy bands: optical, infrared, and TeV gamma rays. At least three astronomical source populations: supernova remnants (SNRs), pulsar wind nebulae (PWNe), and binary systems (BSs) are represented in this figure. In addition, the H.E.S.S. observations of the central region of our Galaxy revealed a diffuse TeV γ-ray emission component which is apparently dominated by contributions from giant molecular clouds (GMCs). These massive complexes of gas and dust most likely serve as effective targets for interactions of relativistic particles from nearby active or recent accelerators. Thus one may claim that four galactic source populations are already firmly established as effective TeV γ-ray emitters. Meanwhile, many sources discovered by H.E.S.S. in the galactic plane remain unidentified. Although some of these sources might have direct or indirect links to SNRs, PWNe, and GMCs, one cannot exclude that a fraction of the H.E.S.S. unidentified sources are related to other source classes.

The Milagro telescope has made the first measurement of the diffuse TeV gamma-ray flux from the Galactic Disk. Figure 26 shows the Galaxy (as visible from the Northern Hemisphere) in TeV gamma rays. In addition to the individual sources discussed above, the image (compiled from Milagro data) shows the existence of a diffuse TeV gamma-ray flux between galactic longitudes of 30° and 90°.

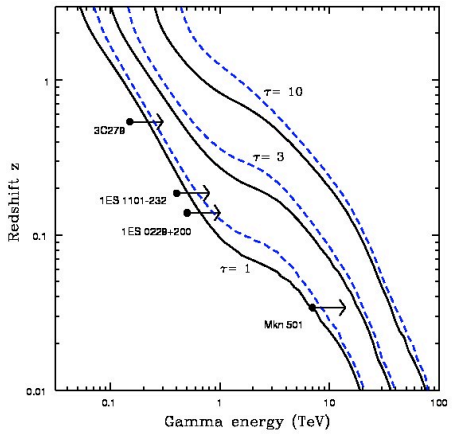


Fig. 23: Redshift for gamma rays as a function of energy. Lines show constant optical depth in two models of IR/O background.

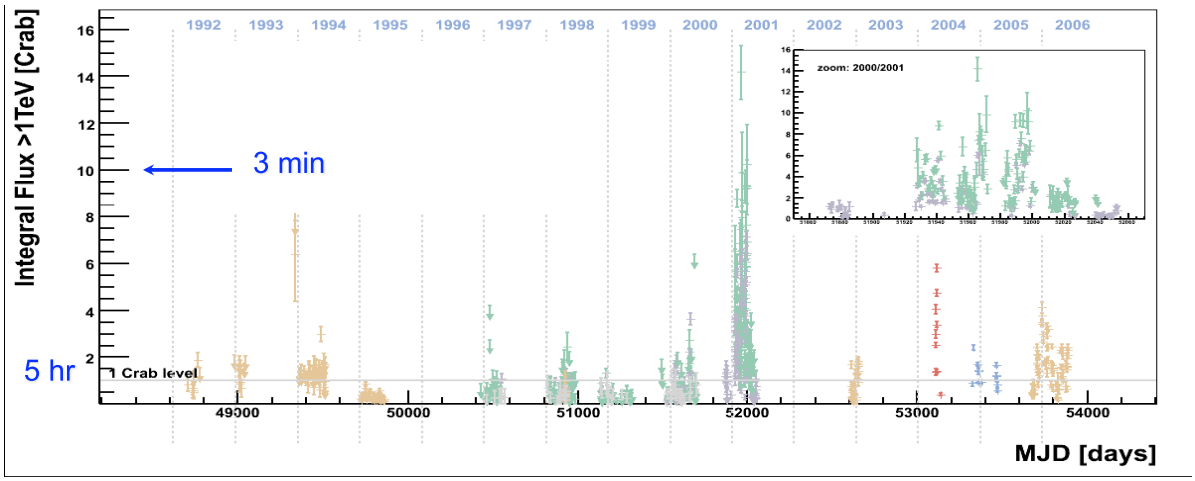


Fig. 24: Observation of Mkn 421 as a function of time

2.3 Extragalactic magnetic fields

Another very important field which will benefit in the near future from TeV gamma rays is Extragalactic Magnetic Fields.

Indeed, as discussed above, TeV gamma rays emitted by astrophysical sources can be measured by detectors on Earth. Practically all TeV gamma rays from galactic sources come directly to the detectors. However, this is not true for extragalactic sources. As one can see from Fig. 23, even for nearby sources like Mkn 501, gamma rays with $E > 10$ TeV cannot come freely to the detector. The pair production on Extragalactic Background Light (EBL) reduces the flux of γ-rays from the source by

$$F(E_{\gamma_0}) = F_0(E'_{\gamma_0}(z_E))e^{-\tau(E_{\gamma_0},z_E)}, \tag{14}$$

where $F(E_{\gamma_0})$ is the detected spectrum, $F_0(E'_{\gamma_0})$ is the initial spectrum of the source, and $\tau(E_{\gamma_0}, z_E)$ is the optical depth Eq. (13). The typical distance which a primary gamma ray travels is

$$D_{\gamma_0} = D_\gamma(E'_{\gamma_0}, z) = 40\frac{\kappa}{(1+z)^2}\left[\frac{E'_{\gamma_0}}{20\text{ TeV}}\right]^{-1}\text{ Mpc}\,, \tag{15}$$

where a numerical factor $\kappa = \kappa(E_{\gamma_0}, z) \sim 1$ accounts for the model uncertainties.

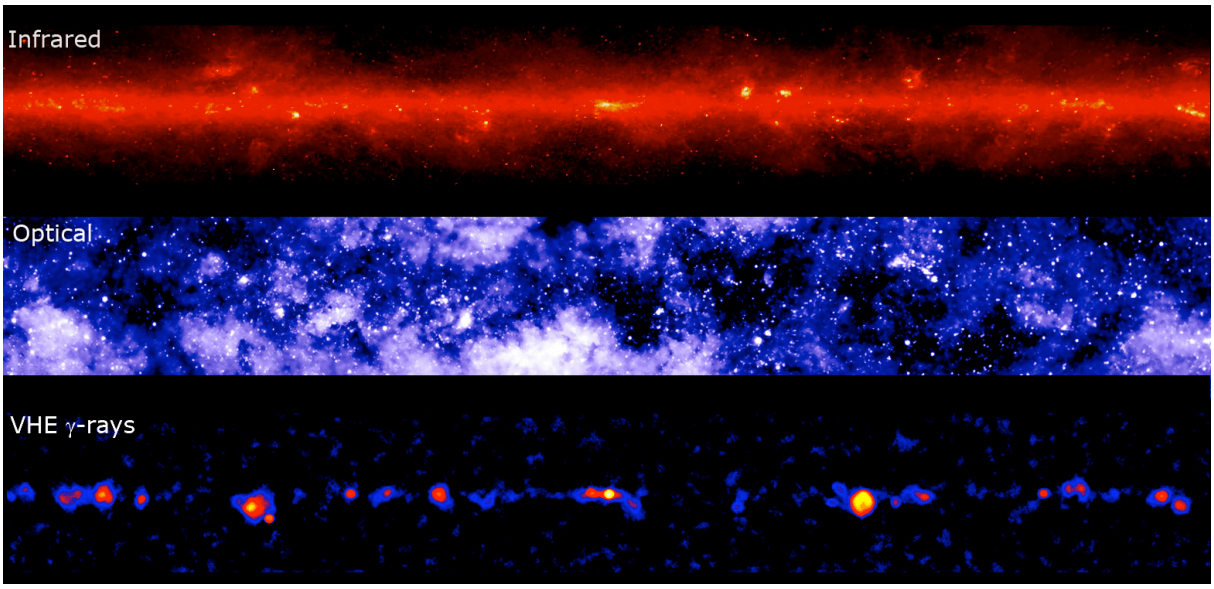


Fig. 25: Central part of the Milky Way galaxy in infrared, optical, and in TeV gamma rays. The TeV gamma-ray sky from H.E.S.S. observations with a large number of sources.

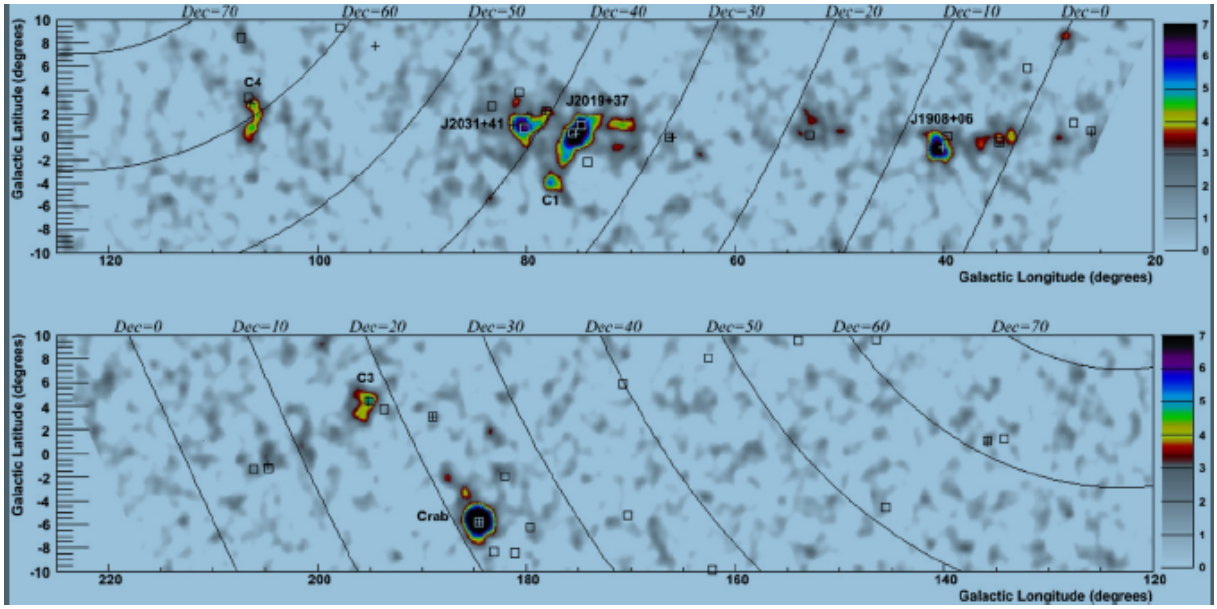


Fig. 26: The Milky Way galaxy in TeV gamma rays from galactic longitude 20° to 220° and galactic latitude from $-10°$ to $10°$. The image is the culmination of a seven-year exposure by the Milagro instrument.

The cascade electrons lose their energy via Inverse Compton (IC) scattering of the CMB photons within the distance

$$D_e = \frac{3m_e^2c^3}{4\sigma_T U'_{\rm CMB} E'_e} \simeq 10^{23}(1+z_{\gamma\gamma})^{-4}\left[\frac{E'_e}{10\ {\rm TeV}}\right]^{-1}\ {\rm cm} \tag{16}$$

The deflection angle of the e^+e^- pairs, accumulated over the cooling distance, depends on the correlation length of the magnetic field, λ_B. Note also that electrons and positrons travel much shorter distances than primary photons.

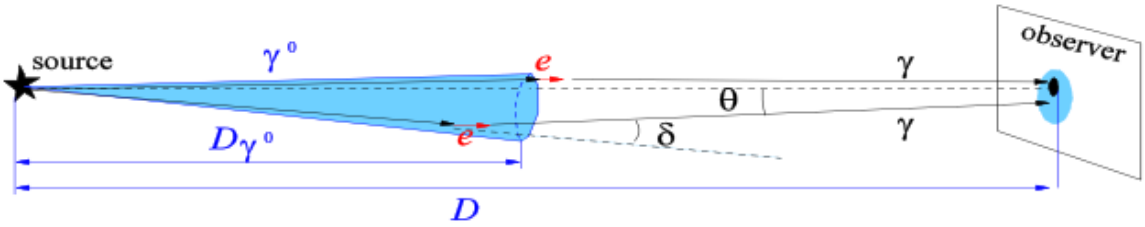


Fig. 27: Detection of EGMF through observation of secondary emissions around a point source [25]

The e^+e^- pairs produced in interactions of multi-TeV γ-rays with EBL photons produce secondary γ-rays via IC scattering of the Cosmic Microwave Background (CMB) photons. Typical energies of the IC photons reaching the Earth are

$$E_\gamma = \frac{4}{3}(1+z_{\gamma\gamma})^{-1}\epsilon'_{CMB}\frac{E'^2_e}{m_e^2} \simeq 0.32\left[\frac{E'_{\gamma_0}}{20\ \mathrm{TeV}}\right]^2\ \mathrm{TeV} \tag{17}$$

where $\epsilon'_{CMB} = 6\times10^{-4}(1+z_{\gamma\gamma})$ eV is the typical energy of CMB photons. In the above equation we have assumed that the energy of a primary γ-ray is $E'_{\gamma_0} \simeq 2E'_e$ with E'_{γ_0} being the energy of the primary γ-rays at the redshift of the pair production. Upscattering of the infrared/optical background photons gives a sub-dominant contribution to the IC scattering spectrum because the energy density of CMB is much higher than the density of the infrared/optical background.

Deflections of e^+e^- pairs produced by the γ-rays, which were initially emitted slightly away from the observer, could lead to 'redirection' of the secondary cascade photons toward the observer. This effect leads to the appearance of two potentially observable effects: extended emission around an initially point source of γ-rays [25–27] and delayed 'echo' of γ-ray flares of extragalactic sources [28, 29].

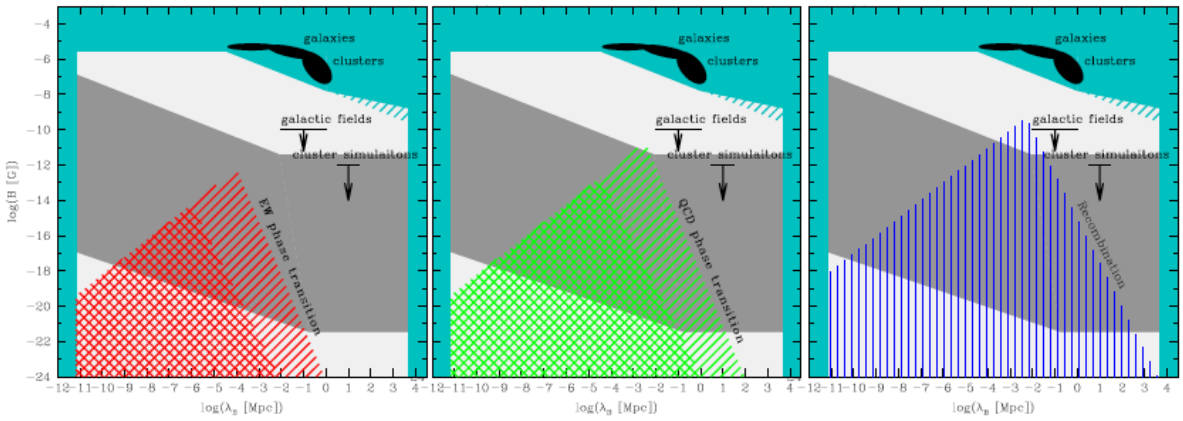


Fig. 28: Model predictions and estimates for the EGMF strength. Cyan shaded region excluded by present day measurements. Black ellipses show measurements of the field in the Galaxy and galaxy clusters. **Left panel:** left and right hatched regions show theoretically allowed range of values of (λ_B,B) for non-helical and helical fields generated at the epoch of electroweak phase transition during radiation-dominated era. **Middle panel:** left and right hatched region show ranges of possible (λ_B,B) for non-helical and helical magnetic fields produced during the QCD phase transition. **Right panel:** hatched region is the range of possible (λ_B,B) for EGMF generated during recombination epoch. Dark grey shaded region shows the range of (λ_B,B) parameter space accessible for the γ-ray measurements via γ-ray observations [30].

The above processes are illustrated in Fig. 27. Electron deflection δ depends on the magnetic field in the region of deflection. Note, that, in principle, EGMF depends on the redshift, $B' = B'(z)$. In the

simplest case, when the magnetic field strength changes only as a result of expansion of the Universe, $B'(z) \sim B_0(1+z)^2$, where B_0 is the present epoch EGMF strength. This gives

$$\begin{aligned} \delta &= \frac{D_e}{R_L} \simeq 3\times 10^{-6}(1+z_{\gamma\gamma})^{-4}\left[\frac{B'}{10^{-18}\,\mathrm{G}}\right]\left[\frac{E'_e}{10\,\mathrm{TeV}}\right]^{-2} \\ &\simeq 3\times 10^{-6}(1+z_{\gamma\gamma})^{-2}\left[\frac{B_0}{10^{-18}\,\mathrm{G}}\right]\left[\frac{E'_e}{10\,\mathrm{TeV}}\right]^{-2} \end{aligned} \tag{18}$$

Knowing the deflection angle of electrons, one can readily find the angular extension of the secondary IC emission from the e^+e^- pairs

$$\Theta_{\rm ext} \simeq \begin{cases} 0.5^\circ(1+z)^{-2}\left[\frac{\tau_\theta}{10}\right]^{-1} \\ \left[\frac{E_\gamma}{0.1\,\mathrm{TeV}}\right]^{-1}\left[\frac{B_0}{10^{-14}\,\mathrm{G}}\right], & \lambda'_B \gg D_e \\ 0.07^\circ(1+z)^{-1/2}\left[\frac{\tau_\theta}{10}\right]^{-1} \\ \left[\frac{E_\gamma}{0.1\,\mathrm{TeV}}\right]^{-3/4}\left[\frac{B_0}{10^{-14}\,\mathrm{G}}\right]\left[\frac{\lambda_{B0}}{1\,\mathrm{kpc}}\right]^{1/2}, & \lambda'_B \ll D_e \end{cases} \tag{19}$$

This is a key point for detection of the field, since extended emission depends on energy in a well-defined way and can be reconstructed using independent measurements at different energies.

The possible ranges of the (λ_B,B) parameter space are shown in Fig. 28 for the cases when magnetogenesis proceeds during electroweak or QCD phase transitions or at the moment of recombination.

It is interesting to note that predictions for the strength and correlation length of the primordial magnetic fields fall in a region of (λ_B,B) parameter space which is not accessible for the existing measurement techniques, such as Faraday rotation or Zeeman splitting methods. However, it turns out that this region of (λ_B,B) parameter space is accessible for the measurement techniques which exploit the potential of the newly opened field of very-high-energy (VHE) γ-ray astronomy [30].

2.4 Summary

Gamma-ray astronomy works, hundreds of sources have been detected in the GeV energy range and about one hundred in TeV energies.

There are several major questions to be answered in the near future:

- One needs to understand the hadronic component in a variety of astrophysical sources.
- Extragalactic IR/O backgrounds have already been constrained by observations of TeV sources to factor two uncertainty. The next step is precision determination of those backgrounds using measurements of many sources at different redshifts.
- For the first time one has a possibility to study primordial magnetic fields through TeV gamma-ray measurements. We can test models of primordial magnetic fields in the near future.

There are several other important issues which were not discussed in this Lecture due to lack of time. The corresponding questions are:

- Good measurements of blazar flairs can help to understand gravity near black holes.
- TeV gamma rays give one more constraint/signature on Dark Matter.
- Constraints on exotic physics (LIV, etc.) will be improved.

3 High-energy neutrinos

3.1 Introduction

In this lecture we discuss theoretical predictions and experimental efforts to detect Ultra-High Energy neutrinos. In Section 3.2 we discuss possible ways to detect UHE neutrinos and their corresponding experiments. In Section 3.3 we show theoretical predictions for UHE neutrino fluxes and present the status of experimental searches for such fluxes. In Section 3.4 we discuss another possibility to detect Galactic neutrino sources at multi-TeV energies. In Section 3.5 we summarize all the results of this lecture.

3.2 High-energy neutrino experiments

There are three types of ultra-high energy (UHE) neutrino experiments.

First, neutrinos can be detected by UHECR experiments. There are two possibilities for this. First, one can use the fact that the atmosphere horizontally has depth 36 times the vertical depth. Relatively young electromagnetic horizontal showers can be caused by neutrinos only. Hadronic showers at such a depth consist of muons only. Second, one can look for events penetrating the Earth in the tau-neutrino channel, i.e., look for upward-going events. This was used by the Auger experiment (see Fig. 7). The resulting limit on neutrino flux is shown in Fig. 31. Also less significant limits were presented by previous UHECR experiments including Fly's Eye, AGASA, and HiRes (the HiRes limit is also shown in Fig. 31).

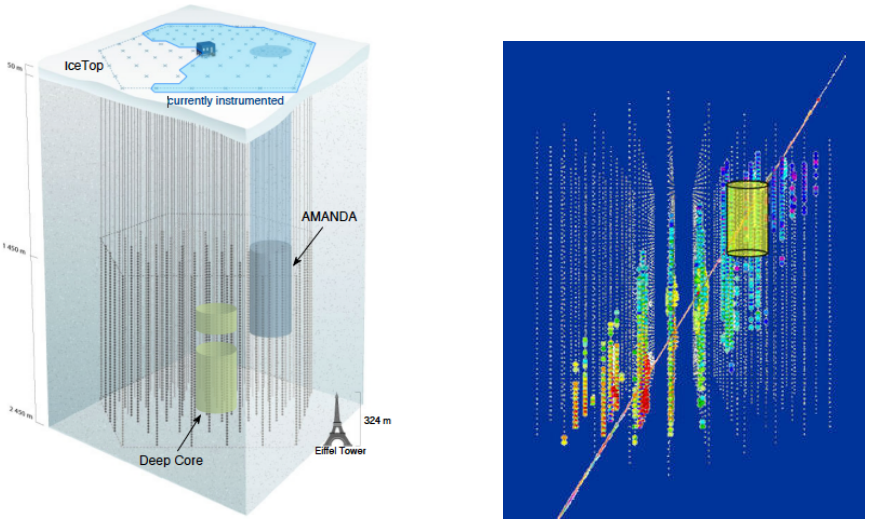


Fig. 29: IceCube detector. **Left:** Configuration of the IceCube detector. Eighty strings will be located at a depth of 1.5 km in the Antarctic ice filling a volume of one cubic kilometre. The present construction stage is also shown [31]. **Right:** Simulation of a high-energy neutrino event in the IceCube detector [32].

Second, one can detect neutrinos in the water or in the ice by detecting Cherenkov light created by corresponding leptons after neutrino interaction in the medium. There are two important backgrounds for such measurement. First, secondary leptons, mostly muons, should not be confused with secondary muons from extensive air showers in the atmosphere. In order to reduce the background of atmospheric muons one has to put the detector at a depth greater than one kilometre from the surface. Second, there are atmospheric neutrinos created by the same cosmic rays, which would produce isotropy in the space energy-dependent background. In order to fight this background, one either has to go to high energies $E > 10^{15-16}$ eV, where it is small, or look for point sources on top of this background.

Experiments that worked with these techniques in the past were Baikal and ANTARES in water and AMANDA in ice. All those experiments had a volume 0.1 km^3 or less. The new-generation ex-

periment IceCube with a volume of 1 km^3 is in the construction stage at the moment. In Fig. 29 in the left panel one can see the configuration of this experiment, which consists of 80 strings, filling a cubic kilometre volume in the Antarctic ice at a depth of 1.5 km from the surface. Strings already implemented are shaded blue on top of the picture (see Ref. [31] for more details). Also, as shown in the figure the top of the detector is covered by an array of ice tanks (ice top). In the right panel one can see a Monte Carlo simulation of a high-energy neutrino event, detected by the IceCube experiment. First results of this experiment will be discussed in Section 3.4.

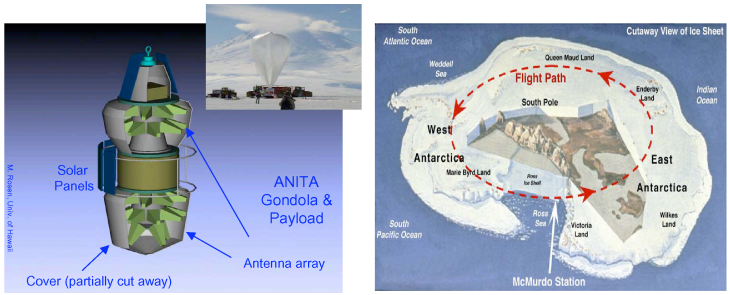


Fig. 30: ANtarctic Impulsive Transient Array (ANITA) radio balloon experiment. Array of radio antennas flying in the ballon, as shown on the left panel. It flies in circles over the Antarctic ice at a height of 37 km (see right panel) and looks for radio signals which UHE neutrinos create in the ice.

Finally, radio neutrino experiments exploit the Askaryan effect in which strong coherent radio emission arises from electromagnetic showers in any dielectric medium. High-energy neutrinos trigger a cascade of electromagnetic particles in the medium, which has net charge and can emit an analogue of Cherenkov light in the radio energy range. The main point of this effect is that the length of the radio wave is macroscopic (tens of centimetres) and is bigger than the size of the cascade itself. This in turn means that all electrons in the cascade emit coherently. The effect was first observed in 2000 at SLAC. Recently the Askaryan effect has been clearly confirmed and characterized for ice as the medium, as part of the pre-flight calibration of the ANITA-1 payload. The Askaryan effect can be seen only at high energies $E > 10^{17-18}$ eV. Experiments using this effect benefit from the absence of atmospheric neutrino flux at such high energies, but they also have to look over a huge effective volume in order to see tiny neutrino fluxes at highest energies.

Experiments that used this effect to search for UHE neutrinos are FORTE [33], RICE [34], and ANITA [35, 36]. FORTE is a satellite experiment, which, in particular, looked over the Greenland ice. Unfortunately, the threshold of this experiment was very high, $E_\nu > 10^{22}$ eV, so it could test only exotic top-down models. RICE was an array of radio antennas located in the ice at the South Pole at the same place as the AMANDA experiment. This experiment presented its final results in 2006. Finally, the most advanced for the moment of this kind of experiment is the ANtarctic Impulsive Transient Array (ANITA) radio balloon experiment, see Fig. 30. In the left panel one can see an array of radio antennas in the balloon. In the right panel one can see a schematic map of flight over the Antarctic at a height of 37 km.

3.3 Search for cosmogenic neutrinos

As discussed in Section 1.3 UHECR protons lose their energy in interactions with CMB photons and produce pions at energies above threshold $E > 6 \cdot 10^{19}$ eV. This GZK threshold was found in 1966 [9]. As long ago as 1969 Berezinsky and Zatsepin suggested that one can try to observe secondary neutrinos from pion decays and called them cosmogenic neutrinos [37]. Recently the ANITA Collaboration proposed to call such neutrinos Berezinsky–Zatsepin neutrinos, or BZ neutrinos [36]. Below we follow this suggestion.

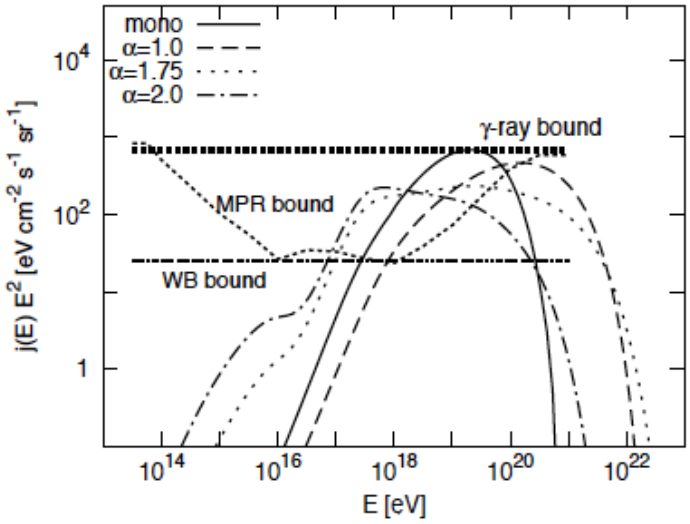


Fig. 31: Predictions of cosmogenic neutrino fluxes and theoretical bounds on them [38, 39]

One can calculate the flux of BZ neutrinos theoretically, after fitting the corresponding proton spectrum to the experimental flux above some energy. The absolute limit for neutrino flux comes from the fact that gamma rays unavoidably produced from π^0 decays and from electrons $\pi^{\pm}$ decays cascade down to GeV energies and the maximum flux of such gamma rays cannot overshoot the EGRET measuremen shown in Fig. 16. This bound on the BZ neutrino flux is called "gamma-ray bound" in Fig. 31. Note that there are many additional ways to create photons in the EGRET energy range, including electron–positron pair production discussed in the previous section, so the real BZ neutrino flux is always lower than this region.

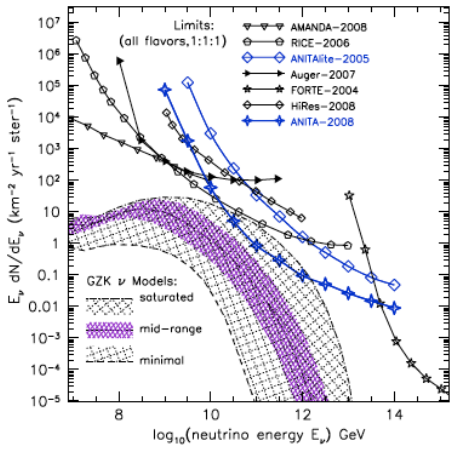


Fig. 32: Experimental limits on cosmogenic neutrino flux. Best up-to-date ANITA-1 limits based on no surviving candidates for 18 days of live time shown as ANITA-2008 [36]. Also limits from Auger [40], HiRes [41], FORTE [33], Anita prototype ANITAlite [35], RICE [34], and AMANDA II [42] are shown.

Also in Fig. 31 we plot two theoretical limits derived under a set of theoretical assumptions. One is called the Waxman–Bahcall (WB) bound and the other the MPR bound. On the same figure we show several examples of theoretical neutrino fluxes which violate both WB and MPR bounds, but all of them are consistent with the experimental gamma-ray bound.

In Fig. 32 we show present-day experimental bounds confronting theoretical predictions for BZ neutrinos from Ref. [36]. One can see that the best up-to-date experimental bounds come from the ANITA experiment. ANITA-1 was able to view a volume of ice of ~ 1.6 Mkm^3 during 17.3 days, however, volumetric acceptance to a diffuse neutrino flux, accounting for the small solid angle of acceptance for any given volume element, is several hundred km^3 water-equivalent steradians at $E_\nu = 10^{19}$ eV. This allowed them for the first time a tough theoretically interesting region, excluding part of the parameter space with highest neutrino fluxes.

On the same figure one can see existing limits on diffuse neutrino flux from the Auger [40], HiRes [41], FORTE [33], Anita prototype ANITAlite [35], RICE [34], and AMANDA II [42] experiments.

Let us note also that in Fig. 32 the composition is assumed to be proton-dominated. If recent Auger results presented in Fig. 9 are confirmed, theoretical expectations for neutrino flux in Fig. 32 will be strongly reduced. This will make observations of the diffused flux of UHE neutrinos an even more complicated issue. However, at lower energies one still can have a hope of seeing point sources with neutrinos, as will be discussed in the next section.

3.4 Point sources of UHE neutrinos

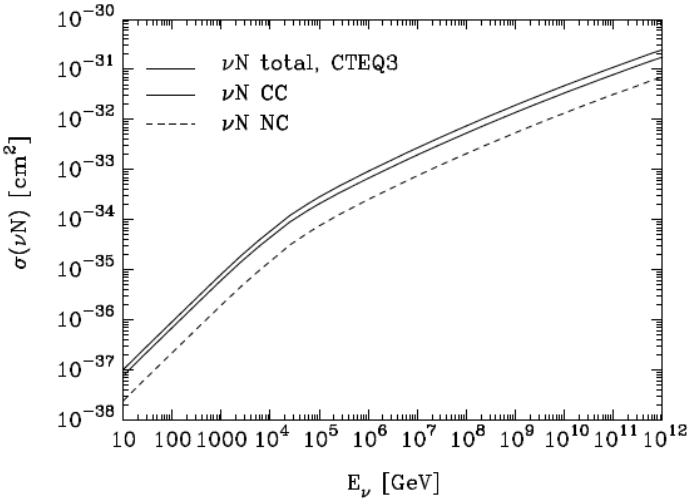


Fig. 33: Neutrino–nucleon cross-section as a function of the neutrino energy. Charge-current and neutral-current contributions to the cross-section are shown with thin solid and dashed lines. The total cross-section is presented by a thick solid line. See Ref. [43] for details.

At highest energies the neutrino flux is too low to detect one single source of neutrinos, but at lower energies $E < 1000$ TeV the flux from a single source can be high enough to detect it. Indeed, in Fig. 33 the neutrino–nucleon cross-section is shown as a function of energy. This cross-section is proportional to E at low energies $E < 1$ TeV and to $E^{0.4}$ at high energies $E > 10^6$ GeV. Good candidates for neutrino sources in the Galaxy are objects emitting TeV gamma rays. They can produce neutrinos in the proton–proton collisions in objects in the case of binary systems and in the interaction with molecular clouds in the Galaxy. In the 10 TeV energy range

$$\sigma_{p\nu}(10 \text{ TeV}) = 10^{-34} \text{ cm}^2 \, . \tag{20}$$

In the IceCube detector only a small fraction of neutrinos will produce a signal:

$$\tau_\nu = \sigma_{p\nu} n_{ICE} R \sim \ 10^{-5} \, , \tag{21}$$

where $n_{ICE} \sim 10^{24}/\text{cm}^3$ is the density of the ice and $R = 1$ km is the height of the IceCube detector.

The expected flux of neutrinos produced in the proton–proton collisions in the Galactic sources is

$$F_\nu \sim F_\gamma = 10^{-12} \frac{1}{\text{cm}^2\text{s}} \approx 3 \cdot 10^5 \frac{1}{\text{km}^2\text{yr}} \,. \tag{22}$$

Thus in the IceCube detector one can expect three events per year for a 10 TeV neutrino flux.

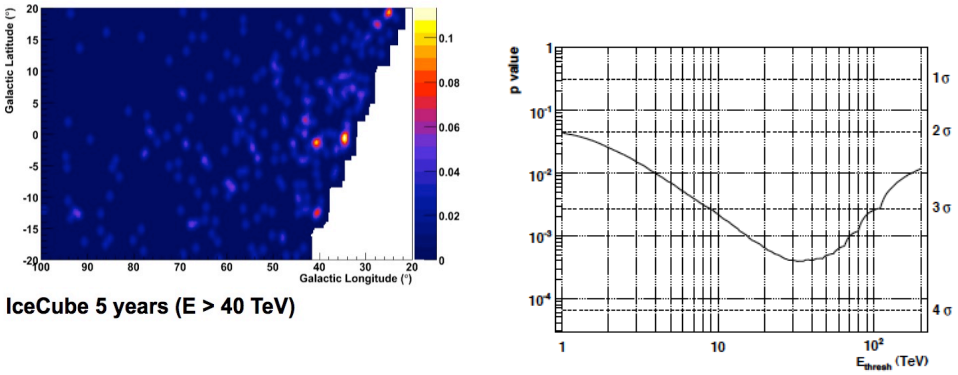


Fig. 34: Left: Simulated detection of Milagro TeV galactic sources by IceCube. **Right:** Significance of Milagro hotspots after five years of observation of IceCube.

In Fig. 34 one can see a simulation of Milagro sources from Fig. 26 after five years of working of the IceCube detector.

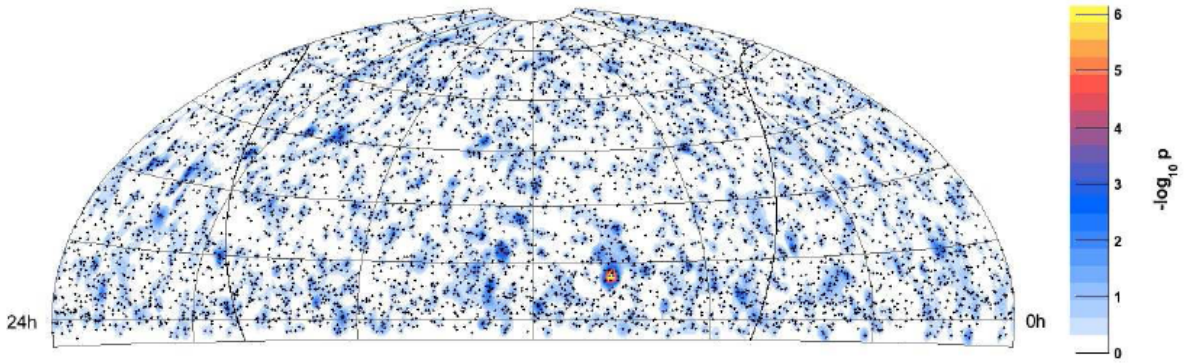


Fig. 35: Equatorial sky-map of events (points) and pre-trial significances (p-value) of the all-sky point source search in the 22-string IceCube detector [44]. The solid curve is the galactic plane. The most significant spot arrives in a random sky with probability $P \sim 1\%$.

We now present recent results for point-source searches using data recorded during 2007–08 with 22 strings of IceCube (1/4 of the detector). An all-sky search within the declination range -5° to $+85^\circ$ found the most significant deviation from the background at 153.4° r.a., 11.4° dec. Accounting for all trials in the point-source search, the final p-value for this result is 1.34%, consistent with the null hypothesis of background-only events at the 2.2σ level. No obvious source candidates are near this location, and an analysis of the timing of the events did not find any evidence of a burst in time. The

location can be added to the a priori source candidate list for analysis using future IceCube data, in which case a similar excess would be identified with much higher significance [44].

3.5 Summary

IceCube is half-complete. If it observes first sources, a new field of astroparticle physics will be started: neutrino astrophysics. If not, much bigger detectors are needed with a size of at least 10 km^3. Secondary neutrino flux from UHECR protons can be detected by future radio experiments, like ANITA. Neutrinos from some bright galactic sources can be detected by IceCube. Extragalactic sources can be observed during bright flair activity. In order to detect continuous flux from sources like Cen A one needs detectors much larger than 1 km^3. Galactic SN can be detected with neutrinos at low and high energies. Cubic-kilometre water detectors will be constructed if IceCube gives positive results.

Acknowledgements

I would like to thank the Organizing Committee of the 5th CERN Latin American School for giving me the opportunity to present lectures there and for the excellent organization of the School.

References

[1] E. S. Seo *et al.*, Measurement of cosmic-ray proton and helium spectra during the 1987 solar minimum, Astrophys. J. **378**, 763 (1991); M. Nagano *et al.*, Energy spectrum of primary cosmic rays between $10^{14.5}$ and 10^{18} eV, J. Phys. **G10**, 1295 (1984); M. Nagano *et al.*, Energy spectrum of primary cosmic rays above 10^{17} eV determined from extensive air shower experiments at Akeno, J. Phys. **G18**, 423 (1992); N. L. Grigorov *et al.*, Energy spectrum of primary cosmic rays in the $10^{11} - 10^{15}$ eV according to the data of Proton-4 measurements, Proceedings 12th ICRC, **1**, 1760 (1971).

[2] R. Abbasi *et al.* [HiRes Collaboration], Phys. Rev. Lett. **100**, 101101 (2008) [arXiv:astro-ph/0703099].

[3] M. Nagano and A. A. Watson, Rev. Mod. Phys. **72**, 689 (2000).

[4] J. Abraham *et al.* [The Pierre Auger Collaboration], The Cosmic Ray Energy Spectrum and Related Measurements with the Pierre Auger Observatory, arXiv:0906.2189.

[5] M. Kachelriess, Lecture Notes on High Energy Cosmic Rays, 2008, arXiv:0801.4376 [astro-ph].

[6] A. M. Hillas, Annu. Rev. Astron. Astrophys. **22**, 425 (1984).

[7] P. Bhattacharjee and G. Sigl, Phys. Rep. **327**, 109 (2000) [arXiv:astro-ph/9811011].

[8] M. Kachelriess, E. Parizot and D. V. Semikoz, JETP Lett. **88**, 553 (2009) [arXiv:0711.3635 [astro-ph]].

[9] K. Greisen, Phys. Rev. Lett. **16**, 748 (1966). G. T. Zatsepin and V. A. Kuzmin, JETP Lett. **4**, 78 [Pisma Zh. Eksp. Teor. Fiz. **4**, 114 (1966)].

[10] F. W. Stecker, Phys. Rev. **180**, 1264 (1969).

[11] M. Kachelriess, P. D. Serpico and M. Teshima, Astropart. Phys. **26**, 378 (2006) [arXiv:astro-ph/0510444].

[12] K. Dolag, D. Grasso, V. Springel and I. Tkachev, JCAP **0501**, 009 (2005) [arXiv:astro-ph/0410419].

[13] G. Sigl, F. Miniati and T. A. Ensslin, Phys. Rev. D **70**, 043007 (2004) [arXiv:astro-ph/0401084].

[14] J. Abraham *et al.* [The Pierre Auger Collaboration], Operations of and Future Plans for the Pierre Auger Observatory, arXiv:0906.2354.

[15] J. Abraham *et al.* [The Pierre Auger Collaboration], Studies of Cosmic Ray Composition and Air Shower Structure with the Pierre Auger Observatory, arXiv:0906.2319.

[16] M. Unger, Study of the Cosmic Ray Composition with the PAO, talk at conference Searching for the Origins of Cosmic Rays, Trondheim, Norway, 2009, http://web.phys.ntnu.no/∼mika/unger2.pdf.

[17] T. Sako *et al.*, Current status and plan of the LHCf experiment, Proceedings of the 31st ICRC, Lodz, 2009.

[18] M. Kachelrieß and D. V. Semikoz, Clustering of ultra-high energy cosmic ray arrival directions on medium scales, Astropart. Phys. **26**, 10 (2006) [arXiv:astro-ph/0512498].

[19] S. Mollerach and the Pierre Auger Collaboration, Nucl. Phys. Proc. Suppl. **190**, 198 (2009) [arXiv:0901.4699 [astro-ph.HE]].

[20] J. Abraham *et al.* [The Pierre Auger Collaboration], arXiv:0906.2347 [astro-ph.HE].

[21] J. Abraham *et al.* [Pierre Auger Collaboration], Science **318**, 938 (2007) [arXiv:0711.2256 [astro-ph]].

[22] O. E. Kalashev, D. V. Semikoz and G. Sigl, Phys. Rev. D **79**, 063005 (2009) [arXiv:0704.2463 [astro-ph]].

[23] G. Gelmini, O. Kalashev and D. V. Semikoz, J. Exp. Theor. Phys. **106**, 1061 (2008) [arXiv:astro-ph/0506128]. Astropart. Phys. **28**, 390 (2007) [arXiv:astro-ph/0702464].

[24] F. Aharonian, J. Buckley, T. Kifune and G. Sinnis, Rep. Prog. Phys. **71**, 096901 (2008).

[25] A. Neronov and D. V. Semikoz, JETP Lett. **85**, 473 (2007) [arXiv:astro-ph/0604607].

[26] A. Elyiv, A. Neronov and D. V. Semikoz, Phys. Rev. D **80**, 023010 (2009) [arXiv:0903.3649 [astro-ph.CO]].

[27] K. Dolag, M. Kachelriess, S. Ostapchenko and R. Tomas, Astrophys. J. **703**, 1078 (2009) [arXiv:0903.2842 [astro-ph.HE]].

[28] R. Plaga, Nature **374**, 430 (1995).

[29] K. Murase, K. Takahashi, S. Inoue, K. Ichiki and S. Nagataki, Astrophys. J. **686** L67 (2008) [arXiv:0806.2829 [astro-ph]].

[30] A. Neronov and D. Semikoz, Sensitivity of gamma-ray telescopes for detection of magnetic fields in arXiv:0910.1920 [astro-ph.CO]. To appear in Physical Review D.

[31] T. DeYoung [for the IceCube Collaboration], Recent Results from IceCube and AMANDA, arXiv:0910.3644 [astro-ph.HE].

[32] F. Halzen, A. Kappes and A. O'Murchadha, Phys. Rev. D **78**, 063004 (2008) [arXiv:0803.0314 [astro-ph]].

[33] N. G. Lehtinen, P. W. Gorham, A. R. Jacobson and R. A. Roussel-Dupre, Phys. Rev. D **69**, 013008 (2004) [arXiv:astro-ph/0309656].

[34] I. Kravchenko *et al.*, Phys. Rev. D **73**, 082002 (2006) [arXiv:astro-ph/0601148].

[35] S. W. Barwick *et al.* [ANITA Collaboration], Phys. Rev. Lett. **96**, 171101 (2006) [arXiv:astro-ph/0512265].

[36] P. W. Gorham *et al.* [ANITA collaboration], Phys. Rev. Lett. **103**, 051103 (2009) [arXiv:0812.2715 [astro-ph]].

[37] V. S. Berezinsky and G. T. Zatsepin, Phys. Lett. B **28**, 423 (1969).

[38] O. E. Kalashev, V. A. Kuzmin, D. V. Semikoz and G. Sigl, Phys. Rev. D **66**, 063004 (2002) [arXiv:hep-ph/0205050].

[39] D. V. Semikoz and G. Sigl, JCAP **0404**, 003 (2004) [arXiv:hep-ph/0309328].

[40] J. Abraham *et al.* [The Pierre Auger Collaboration], Phys. Rev. Lett. **100**, 211101 (2008) [arXiv:0712.1909 [astro-ph]].

[41] R. U. Abbasi *et al.*, An upper limit on the electron-neutrino flux from the HiRes detector, arXiv:0803.0554 [astro-ph].

[42] M. Ackermann *et al.* [IceCube Collaboration], Astrophys. J. **675**, 1014 (2008) [arXiv:0711.3022

[astro-ph]].

[43] R. Gandhi, C. Quigg, M. H. Reno and I. Sarcevic, Astropart. Phys. **5**, 81 (1996) [arXiv:hep-ph/9512364].

[44] R. Abbasi *et al.* [IceCube Collaboration], Astrophys. J. **701**, L47 (2009) [arXiv:0905.2253 [astro-ph.HE]].

Relativistic heavy-ion physics

*G. Herrera Corral**
CERN, Geneva, Switzerland

Abstract
The study of relativistic heavy-ion collisions is an important part of the LHC research programme at CERN. This emerging field of research focuses on the study of matter under extreme conditions of temperature, density, and pressure. Here we present an introduction to the general aspects of relativistic heavy-ion physics. Afterwards we give an overview of the accelerator facility at CERN and then a quick look at the ALICE project as a dedicated experiment for heavy-ion collisions.

1 Introduction

The study of relativistic heavy-ion collisions started in the 1970s at the Bevalac, Lawrence Berkeley National Laboratory, where a transport line was built to bring heavy ions from Hilac (Heavy ion linear accelerator) to the Bevatron. The Bevatron at LBNL is best known for the antiproton, discovered there in the 1955 by O. Chamberlain and E. Segré. The so-called Bevalac accelerated nuclei at about 1 *A* GeV/*c*. The demonstration that excited nuclear matter could be studied gave birth to research programmes at Brookhaven National Laboratory and at the European Organization for Nuclear Research (CERN). The Alternating Gradient Synchrotron (AGS) at Brookhaven National Laboratory in the United States accelerated silicon ions up to 15 *A* GeV. In Europe the Super Proton Synchrotron (SPS, CERN) produced a 60 *A* GeV beam of oxygen and then increased the energy to 200 *A* GeV.

Nowadays, research is conducted at the Relativistic Heavy Ion Collider (RHIC). This accelerator was completed in 1999 at Brookhaven National Laboratory in the United States. RHIC collides nuclear beams at 100 *A* GeV, i.e., at ten times more energy than at the SPS. At RHIC, four experiments are taking and analysing data: BRAHMS, PHENIX, PHOBOS, and STAR.

High-energy heavy-ion collisions involve large amounts of energy. RHIC accelerates gold nuclei at 100 GeV/nucleon, which means that each nucleus carries energy

$$100\ \text{GeV} \times 197\ \ \text{nucleons} = 19.7\ \ \text{TeV}.$$

In the centre of mass, these interacting ions deliver 39.4 TeV. The Large Hadron Collider at CERN will reach

$$\sqrt{s} = 1200 \quad \text{TeV}$$

in lead–lead interactions.

In high-energy collisions of protons and/or electrons, the energy available in the beam goes into a point interaction. In heavy-ion interactions, however, an enormous amount of energy is deposited in a small region of space and in a very short time. In this region the density of energy is so large that it may favour the appearance of new forms of matter. The search for these new forms of matter is the central objective of heavy-ion physics.

The energy density of nuclei with atomic number A in normal conditions is given by

$$\varepsilon = \frac{A \times nucleon_{mass}}{V_{nuclear}}, \qquad \text{where} \qquad V_{nuclear} = \frac{4}{3}\pi (r_0 A^{1/3})^3 .$$

* On sabbatical leave from Physics Department, CINVESTAV, Mexico City, Mexico.

A typical value of energy density for nuclear matter is $\varepsilon = 0.14$ GeV/fm^3. The energy densities reached at relativistic heavy-ion collisions are above 1 GeV /fm^3, i. e., 10 times larger than normal nuclei densities.

The future of these studies is now moving to CERN where the ALICE experiment is being prepared to study relativistic heavy-ion collisions at the highest energy ever. As mentioned above, the LHC will provide beams of lead at energies 30 times greater than at RHIC. The CMS and ATLAS experiments at the LHC will also study heavy-ion interactions in addition to their rich programme on proton–proton collisions.

Here we shall give an introduction to the new and exciting field of relativistic heavy-ion collisions. We take a quick historical look at Hagedorn's first predictions. We quickly go through Glauber's model to understand the way phenomena are experimentally evaluated and measured. We then explain the concept of energy density.

In order to introduce the QCD phase space diagram, we shall study the MIT bag model. This model provides an easy way to grasp general ideas before a more formal approach can be taken. With these tools we can discuss some of the probes and signatures that will uncover the appearance of a quark–gluon plasma. Finally we shall comment on the Large Hadron Collider as well as on the ALICE experiment which is dedicated to the study of ion–ion collisions at CERN.

2 Hagedorn limiting temperature

Rolf Hagedorn was the first to point out the possibility of a transition of ordinary matter into a plasma of quarks and gluons. He developed statistical physics methods and applied them to particle production in high-energy collisions. He observed that the measured density of hadron states grows exponentially, i.e.,

$$\frac{d\rho}{dm} \approx m^a e^{m/m_0} \tag{1}$$

where m represents the mass of the observed hadrons and a is a parameter [1]. In 1965 Hagedorn showed that this exponential behaviour implies a limiting temperature which he understood as a melting point of hadrons. Indeed, the number of states with energy in an interval between E and $E + dE$ can be written [2] as

$$dn(E) \approx dE \int_0^E pE dm \frac{d\rho}{dm} e^{-E/kT} .$$

Introducing here the expression given in Eq. (1), and using $p^2 = E^2 - m^2$, one obtains

$$dn(E) \approx dE \int_0^E m^a e^{m/m_0} e^{-E/kT} \sqrt{E^2 - m^2}\, E dm .$$

Henceforth,

$$= E^{a+3} dE \int_0^1 z^a e^{zE/m_0} e^{-E/kT} \sqrt{1 - z^2}\, dz$$

with $m = zE$. Substituting $z = \cos(\varphi)$

$$= E^{a+3} e^{-E/kT} dE \int_0^{\pi/2} \cos^a(\varphi) \sin^2(\varphi) e^{E\cos(\varphi)/m_0} d\varphi \tag{2}$$

assuming $E/m_0 >> 1$, we may approximate Eq. (2) by

$$\approx E^{a+3} e^{-E/kT} dE \int_0^{\pi} \cos^a(\varphi) \sin^2(\varphi) e^{E\cos(\varphi)/m_0} d\varphi.$$

So that the integral can be calculated,

$$= E^{a+3} e^{-E/kT} dE \sqrt{\pi} \frac{2m_0}{E} \frac{\sqrt{\pi}}{2} \frac{e^{E/m_0}}{\sqrt{2\pi E/m_0}}$$

and then simplified to

$$= E^{a+3} dE \sqrt{\frac{\pi m_0^3}{2E^3}} e^{(E/m_0 - E/kT)} \quad .$$

Henceforth, the total energy density $\int_0^{\infty} E dn(E)$ diverges for $kT > kT_0 = m_0$.

The conclusion is therefore that no higher temperatures are possible or some new physics must become relevant.

Figure 1 shows the mass spectrum $\rho_{\exp}(m) = \sum \nu_i \delta(m - m_i)$ with $\nu_i = (2J_i + 1)(2I_i + 1)2^{\lambda_i}$ where J is the spin, I the isospin, and $\lambda = 1$, when particles are different from antiparticles and $\lambda = 0$ when particles are identical to their antiparticles. Figure 1 is then a comparison of the logarithmic smoothed mass spectrum for the hadronic particles known today and previously. One can see that the new hadron resonances improve the exponential behaviour predicted by Hagedorn.

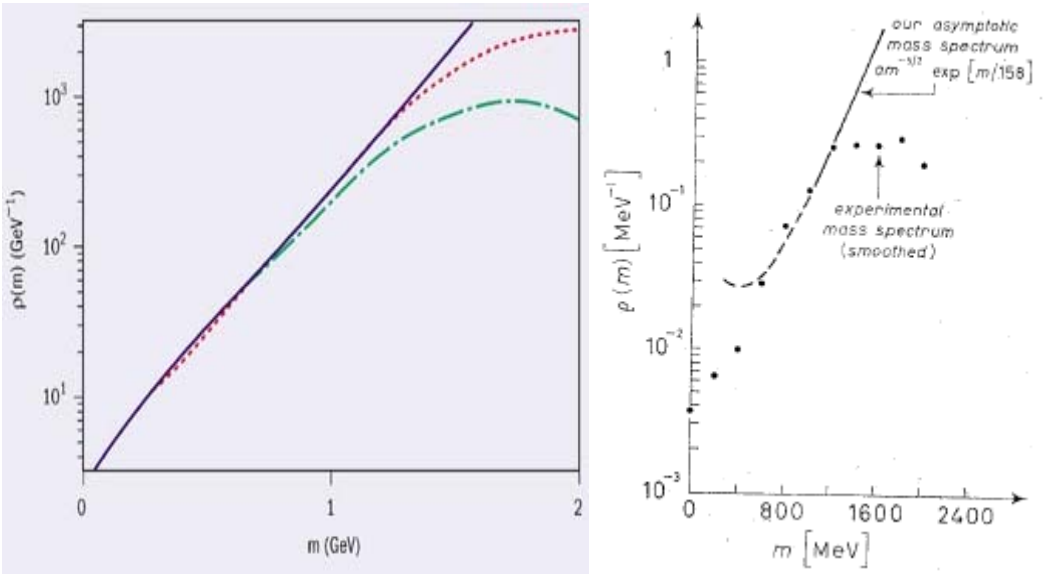


Fig. 1: On the left, a picture extracted from Ref. [3]; the solid blue line is the exponential fit to the smoothed hadron mass spectrum with present day information (represented here by the short-dashed red line). The long dashed green line corresponds to the Hagedorn spectrum obtained in 1967. On the right a similar picture extracted from the paper of Hagedorn from 1965 (see bibliography).

3 The Glauber model

The Glauber model [4], describes the interaction of two nuclei in terms of the interaction of the constituent nucleons. The model assumes the movement of the nucleus in a straight line and pictures the collision between the nuclei with a given impact parameter. In that sense it is a classical model of the interaction. It is widely used in heavy-ion collisions to describe interaction processes.

Figure 2 shows the geometry of a collision between nucleus *B* and nucleus *A*. The probability of finding a baryon in the volume element $d\vec{b}_B dz_B$ of nucleus *B* is $\rho(\vec{b}_B, z_B) d\vec{b}_B dz_B$.

A similar expression for a nucleus *A* can be written. With this in mind, the probability element for having a baryon–baryon interaction when ions *A* and *B* collide with an impact parameter $\vec{b}$ is

$$dP = \underbrace{\rho_A(\vec{b}_A, z_A) d\vec{b}_A dz_A}\underbrace{\rho_B(\vec{b}_B, z_B) d\vec{b}_B dz_B}\underbrace{t(\vec{b} - \vec{b}_A - \vec{b}_B)\sigma_{in}}$$

probability for finding a baryon in A *in B. Probability for an inelastic collision.*

We define the thickness functions $T_A(\vec{b}_A) = \int dz_A \rho_A(\vec{b}_A, z_A)$ for nucleus *A* and correspondingly $T_B(\vec{b}_B) = \int dz_B \rho_B(\vec{b}_B, z_B)$.

So we can write

$$T(\vec{b}) = \int d\vec{b}_A d\vec{b}_B T_A(\vec{b}_A) T_B(\vec{b}_B) t(\vec{b} - \vec{b}_A - \vec{b}_B) . \qquad (3)$$

With this, we can now write the probability for the occurrence of n inelastic interactions when two nuclei A and B collide with an impact parameter $\vec{b}$:

$$P(n,b) = \binom{AB}{n} \left[T(b)\sigma_{in}\right]^n \left[1 - T(b)\sigma_{in}\right]^{AB-n}.$$

The total probability of having an inelastic event in the collision of A and B is therefore

$$\frac{d\sigma_{inel}^{AB}}{db} = \sum_{n=1}^{AB} P(n,b) = 1 - \left[1 - T(b)\sigma_{in}\right]^{AB}. \tag{4}$$

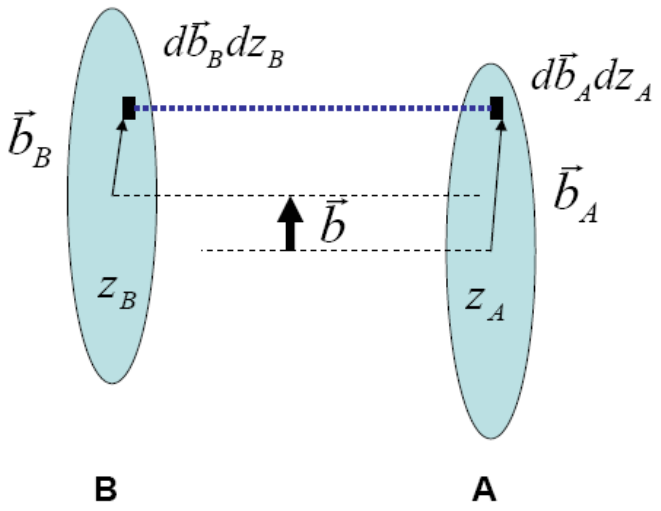


Fig. 2: Collision of nucleus A with nucleus B at impact parameter $\vec{b}$

From Eq. (4) one can see that the total inelastic cross section σ_{in}^{AB} is

$$\sigma_{inel}^{AB} = \int db \left\{1 - \left[1 - T(b)\sigma_{in}\right]^{AB}\right\}. \tag{5}$$

One may approximate the thickness function t in Eq. (3) with a Gaussian. For nuclei with small atomic number the density function can also be approximated by a Gaussian so that $T(b)$ in Eq. 3 can be written

$$T(b) = \exp(-b^2 / 2\beta^2) / 2\pi\beta^2 \cdot$$

The total inelastic cross-section is then

$$\sigma_{in}^{AB} = -2\pi\beta^2 \sum_{n=1}^{AB} \binom{AB}{n} \left(-\frac{\sigma_{in}^{n}}{n(2\pi\beta^2)^n} \right).$$

The simplest case of a proton–proton collisions with *n=1* is fulfilled in this approximation.

4 Energy density

The larger the number of nucleon–nucleon inelastic collisions the larger the energy deposited in the volume where those collisions take place.

Figure 3 shows two colliding ions A' and B'. The overlap area in the transverse direction is denoted with A. The volume formed by this area and a thickness length Δz is then $A\Delta z$. The number density of particles produced in that volume at $z = 0$ and at the time at which a quark–gluon plasma may form is given by

$$\frac{\Delta N}{A\Delta z} = \frac{1}{A}\frac{dN}{dy}\frac{dy}{dz}\bigg|_{y=0} = \frac{1}{A}\frac{dN}{dy}\frac{1}{\tau_0 \cosh y}\bigg|_{y=0}. \qquad (6)$$

Since $z = \tau \sinh y$ with $\tau = \sqrt{t^2 - z^2}$, and where τ is the proper time. We evaluate dy/dz in Eq. (6). This relation connects energy density and rapidity density. It was derived by Bjorken [5].

Considering that $E = m_T \cosh y$, with E being the average energy per produced particle, the energy density at the moment of the collision is

$$\varepsilon_0 = \frac{\Delta N}{A\Delta z} m_T \cosh y. \qquad (7)$$

In this context, produced particles means everything appearing at rapidities intermediate between those of the original incoming nuclei.

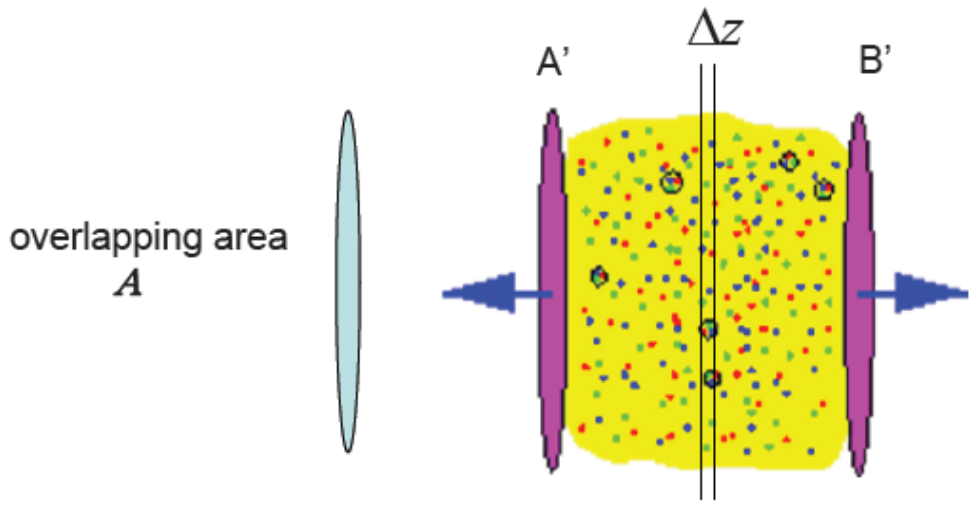


Fig. 3: Two colliding ions A' and B'

Using Eq. (6) in Eq. (7) we obtain

$$\varepsilon_0 = \frac{m_T}{\tau_0 A}\frac{dN}{dy}\bigg|_{y=0},$$

where τ_0 is unknown. Bjorken estimated $\tau_0 = 1 fm/c$, however, the determination of the time scale at which the QGP is formed requires a knowledge of the dynamics behind.

5 The quantum chromodynamics phase diagram

In the MIT bag model [6], hadrons are thought of as closed containers of massless quarks which can be described by the Dirac equation. In the space time representation

$$(i\gamma^{\mu} p_{\mu} - m)\varphi = 0\,.$$

With $m = 0$, the equation becomes

$$\gamma p \varphi = 0,$$

i. e. ,

$$(\gamma^0 p^0 - \vec{\gamma}\cdot\vec{p})\varphi = 0.$$

In the Dirac representation of the γ matrices,

$$\gamma^0 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} \qquad \gamma = \begin{pmatrix} 0 & \sigma \\ -\sigma & 0 \end{pmatrix}$$

the equation above can be written as

$$\begin{pmatrix} p^0 & -\vec{\sigma}\cdot\vec{p} \\ \vec{\sigma}\cdot\vec{p} & -p^0 \end{pmatrix}\begin{pmatrix} \varphi_+ \\ \varphi_- \end{pmatrix} = 0 .$$

These coupled equations $p^0\varphi_+ - \vec{\sigma}\cdot\vec{p}\varphi_- = 0$ and $\vec{\sigma}\cdot\vec{p}\varphi_+ - p_0\varphi_- = 0$ can be solved analytically. The lowest energy solution is

$$\varphi_+(\vec{r},t) = Ne^{-ip^0t} j_0(p^0 r)\chi_+ \qquad \varphi_-(\vec{r},t) = Ne^{-ip^0t}\vec{\sigma}\cdot\hat{r}j_1(p^0 r)\chi_- ,$$

in terms of the spherical Bessel functions j_0 and j_1 . Confinement can now be imposed by requiring the current flux through the spherical bag surface to be zero. This means that the normal component of the current $J_\mu = \bar{\varphi}\gamma_\mu\varphi$ is equal to zero, i.e., $n^\mu\bar{\varphi}\gamma_\mu\varphi = 0$, therefore $\bar{\varphi}\varphi = 0$.

This confinement condition means that

$$\bar{\varphi}\varphi\big|_{r=R} = \left[j_0(p^0 R)\right]^2 - \hat{\sigma}\cdot\hat{r}\hat{\sigma}\cdot\hat{r}\left[j_1(p^0 R)\right]^2 = 0 .$$

That condition can be fulfilled if $p^0 R = 2.04$, which means that the energy of the quarks in the bag will be $E = 2.04N/R$. For a bag under an external pressure B , the energy of the quarks inside becomes

$$E = \frac{2.04N}{R} + \frac{4\pi}{3}R^3 B.$$

The bag will be in equilibrium when

$$\frac{\partial E}{\partial R} = 0 , \qquad \text{i.e. ,} \qquad 4\pi R^2 B - \frac{2.04N}{R^2} = 0 .$$

Henceforth, a proton with three quarks (*N=3*) and radius *r* = 0.8 fm, will have external pressure $B^{1/4} = 1044 \times 197.3$, ($\hbar c = 197$ MeV fm), i.e. ,

$$B^{1/4} = 206 \quad \text{MeV}.$$

Let us now see what happens with a gas of quarks (fermions) and gluons (bosons) in thermal equilibrium. The total pressure of an ideal gas of quarks and gluons would be given by

$$P = \left[g_g + \frac{7}{8}(g_q + g_{\bar{q}}) \right]\frac{\pi^2}{90}T^4 \qquad (8)$$

where $g_g = 8 \times 2$ is the gluon degeneracy determined by the number of gluons and the two possible states. For the quarks we shall have $g_q = g_{\bar{q}} = N_{colors} \times N_{spin} \times N_{flavor}$. The pressure can then be written

$$P = 37\frac{\pi^2}{90}T^4 .$$

When the pressure equals the bag pressure, i.e., $P = B$, the equation would give us the critical temperature at which the bag would break:

$$T_c = \left(\frac{90}{37\pi^2} \right)^{1/4} B^{1/4}. \tag{9}$$

Figure 4 shows T_c as obtained from Eq. (9) with $B^{1/4} = 206$ MeV.

In a similar way one could estimate a critical density and see that deconfinement may happen even at temperature *T = 0.* The number of quarks in a volume *V* with momentum *p* in the interval *dp* is

$$N_q = \frac{g_q V}{(2\pi)^3} \int_0^{\mu_q} 4\pi p^2 dp = \frac{g_q V}{6\pi^2} \mu_q^3$$

so that the number density is

$$n_q = \frac{N_q}{V} = \frac{g_q}{6\pi^2} \mu_q^3.$$

The energy of these quarks in a volume V is

$$E_q = \frac{g_q V}{(2\pi)^3} \int_0^{\mu_q} 4\pi p^3 dp = \frac{V g_q}{8\pi^2} \mu_q^4.$$

From the relation between pressure and energy $P_q = \frac{1}{3}\frac{E}{V}$, we obtain

$$P_q = \frac{g_q}{24\pi^2} \mu_q^4.$$

A change of state will occur when the pressure equals the bag pressure, i.e., $P_q = B$, this corresponds to a critical quark number density

$$n_q = 4 \left(\frac{g_q}{24\pi^2} \right)^{1/4} B^{3/4},$$

and taking baryons as groups of three quarks, the critical baryon number is therefore

$$N_q = \frac{4}{3} \left(\frac{g_q}{24\pi^2} \right)^{1/4} B^{3/4}.$$

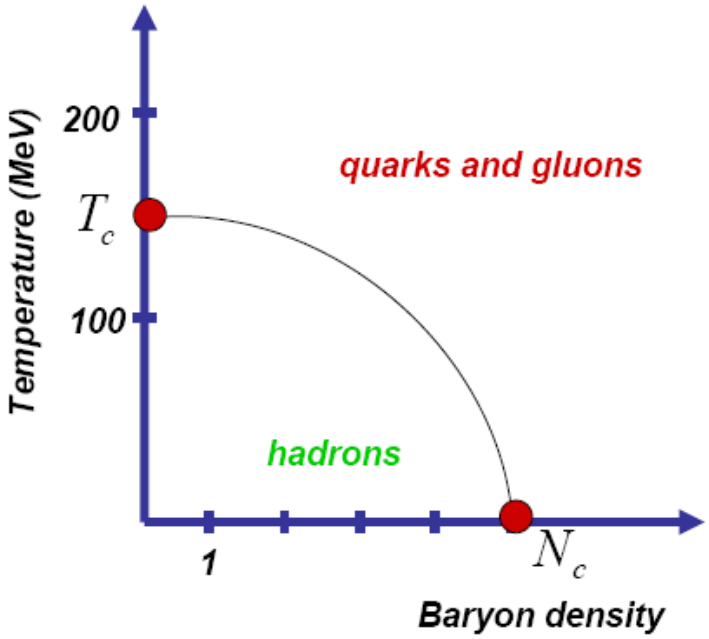


Fig. 4: QCD phase diagram. For high temperature, as discussed in the text, there is a critical temperature T_c. Beyond this temperature, the bag would break releasing quarks and gluons. Similarly a baryon density beyond N_c would produce a phase transition.

We now take ordinary nuclear matter composed of *u* and *d* quarks only so that $g_q =$ $= 3_{colors} \times 2_{spin} \times 2_{flavor}$ and a bag pressure $B^{1/4} = 206$ MeV, the critical baryon number density at temperature $T = 0$, is

$$N_c = 0.72 / \text{fm}^3,$$

which corresponds to 5 times (see Fig 4) the normal nuclear density ($\varepsilon = 0.14$ GeV/fm^3, estimated in the introduction to this article.

6 Quark–gluon plasma probes and signatures

In order to know if a plasma of quarks and gluons has been created in the collision of ultrarelativistic heavy ions, we need observables. There are a number of ideas on what to look at to disentangle the short existence of a new state of matter. For lack of space, we shall not review all the probes and signatures considered by experiments nowadays, we shall only comment on some of them. The interested reader can then expand his knowledge from the bibliography recommended at the end of this article.

Bose–Einstein correlations

Two-particle correlations are among the most promising observables of the heavy-ion reaction to reveal the spacetime evolution (Fig. 5).

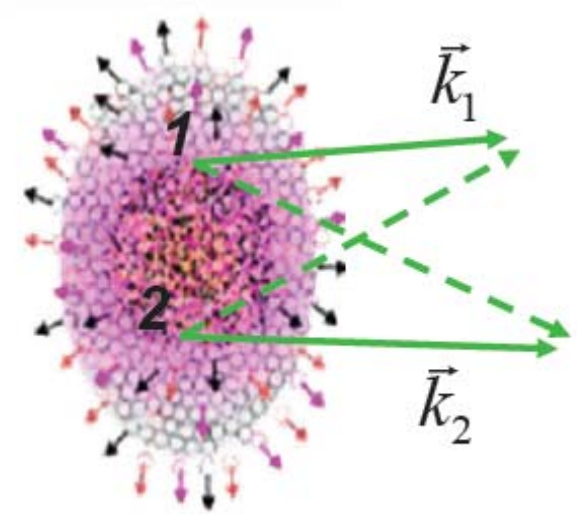


Fig. 5: Two identical particles are produced at different spacetime points 1 and 2, with momentum k_1 and k_2. Identical bosons obey Bose–Einstein statistics, so that quantum correlations are present and modify the phase space of the produced particles.

The wave function of the two particles produced at points $\vec{r}_1$ and $\vec{r}_2$ with momentum $\vec{k}_1$ and $\vec{k}_2$ can be written as

$$A_{\pi\pi} = e^{i\vec{k}_1\vec{r}_1 + i\vec{k}_2\vec{r}_2} + e^{i\vec{k}_1\vec{r}_2 + \vec{k}_2\vec{r}_1} \quad .$$

The amplitude for the process is then

$$\left|A_{\pi\pi}\right|^2 = \left|e^{i\vec{k}_1\vec{r}_1 + i\vec{k}_2\vec{r}_2} + e^{i\vec{k}_1\vec{r}_2 + \vec{k}_2\vec{r}_1}\right|^2$$

$$\left|A_{\pi\pi}\right|^2 = 1 + 1 + e^{i(\vec{k}_1 - \vec{k}_2)\vec{r}_1} + e^{i(\vec{k}_2 - \vec{k}_1)\vec{r}_2}$$

$$\left|A_{\pi\pi}\right|^2 = 2 + 2\cos(qr)$$

where $q = \left|\vec{k}_1 - \vec{k}_2\right|$ and $r = \left|\vec{r}_1 - \vec{r}_2\right|$. Henceforth the probability of having two identical bosons, say pions, produced at two points in spacetime $\vec{r}_1$, $\vec{r}_2$ and with two momenta, $\vec{k}_1$, $\vec{k}_2$ divided by the probability of producing bosons independently, is given by

$$\frac{|A_{\pi\pi}|^2}{|A_\pi||A_\pi|} = 1 + \cos(qr) \quad .$$

We may introduce a probability density $\rho(x)$ for the pions to be produced at different points in spacetime. This amplitude would modify as follows:

$$A_{\pi\pi} = \rho(x) e^{i\vec{k}_1\vec{r}_1 + i\vec{k}_2\vec{r}_2} + \rho(x) e^{i\vec{k}_1\vec{r}_2 + \vec{k}_2\vec{r}_1}$$

with

$$\int \rho(x) dx = 1 \, .$$

In this case the ratio of amplitudes would contain the Fourier transform of the probability density $\rho(x)$, i.e.,

$$\frac{|A_{\pi\pi}|^2}{|A_\pi||A_\pi|} = 1 + |F(\rho)|^2 \, .$$

This ratio of amplitudes is the so-called two-particle correlation function.

$$C(k_1, k_2) = \frac{P(k_1, k_2)}{P(k_1)P(k_2)} = 1 + |F(\rho)|^2 \, .$$

Taking for example

$$\rho(x, y, z) = \frac{N}{4\pi^2 R_x R_y R_z \sigma_t} e^{-\left(\frac{x^2}{2R_x^2} + \frac{y^2}{2R_y^2} + \frac{z^2}{2R_z^2} + \frac{t^2}{2\sigma_t^2}\right)} \, ,$$

the correlation function would be

$$C(k_1 k_2) = 1 + F[\rho](q) = 1 + N' \exp\left[-\frac{1}{2}\left(R_x^2 q_x^2 + R_y^2 q_y^2 + R_z^2 q_z^2 + \sigma_t^2 q_t^2 \right) \right].$$

By studying the two-particle correlation function one can measure the geometry of the particle production system.

Along these lines, one can use more sophisticated parametrizations. Figure 6 shows a common parametrization for the heavy-ion particle production environment and the results obtained by the PHENIX Collaboration using this particular geometry [7].

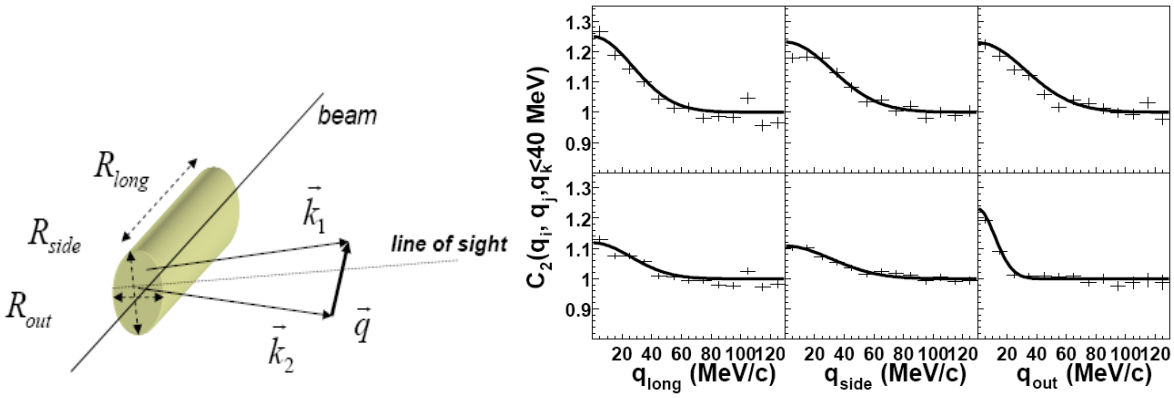


Fig. 6: On the left, a parametrization in terms of R_{long} along the beam direction, R_{out} along the *line of sight,* and R_{side} perpendicular to the *line of sight*. The momenta of the pions are $\vec{k}_1$ and $\vec{k}_2$. On the right the experimental results by the PHENIX Collaboration [7] in terms of these parameters for pion pairs in the laboratory (top) and the pair centre-of-mass frame (bottom). The data is plotted as a function of one variable keeping the other two below 40 MeV/*c*.

On the other hand, if a quark–gluon plasma is produced, it will hadronize, populating the central rapidity region.

Considering S_{QGP}, the entropy of the quark–gluon plasma, and S_{had}, the entropy of the hadronization phase, then by the second law of thermodynamics

$$volume_{QGP} \times S_{QGP} \quad \leq \quad volume_{had} \times S_{had}.$$

Since $S_{had} < S_{QGP}$ then $V_{had} > V_{QGP}$.

Measuring the volume of the hadronization region by means of the two-particle correlation function one may say something about the production of a new state of matter. This is just an example of the ideas that have been considered in the frame of data coming from RHIC experiments. The source size extracted by fitting the correlation function to data grows with the event multiplicity and decreases with transverse momentum. However, the size and time of emission are anomalously large with respect to what has been suggested as signals for quark–gluon plasma formation. A better understanding of models and data is necessary.

J/Ψ suppression

The suppression of *J/Ψ* meson production was proposed in 1986 as a signature of a quark–gluon plasma [8]. It should be the manifestation of colour screening that would hinder *c* and anti-*c* quarks from binding to form a *J/Ψ* meson.

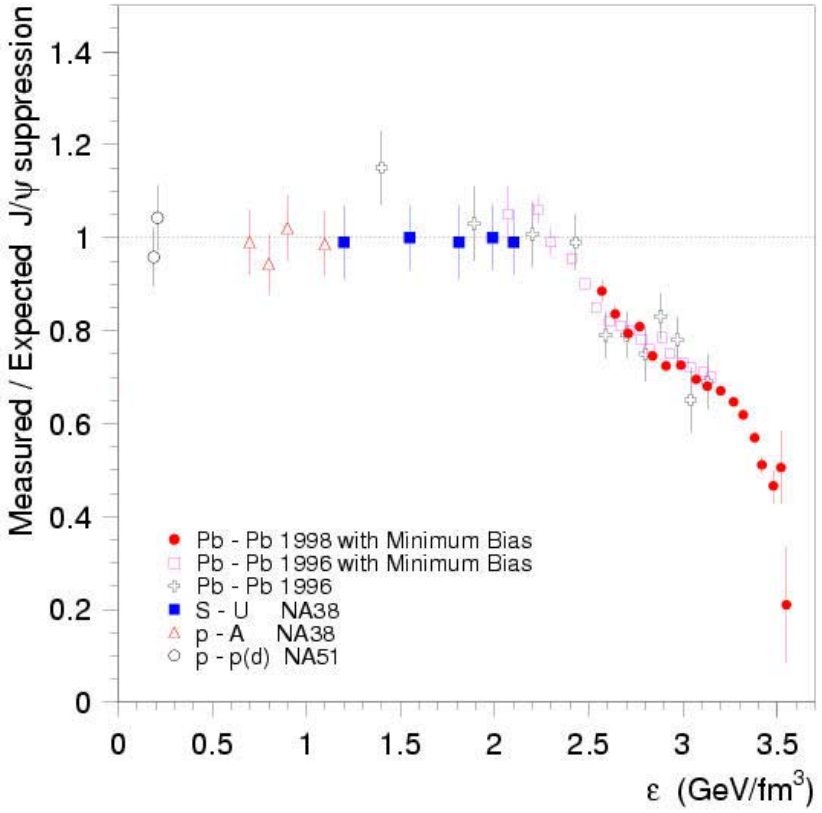


Fig. 7: Ratio of J/ψ charmed mesons produced and expected, in several reactions and as a function of the energy density obtained in the reaction. Figure extracted from Ref. [9].

The first observation in experiment NA50 [9] at the SPS was explained as the result of inelastic interactions of these mesons with dense hadronic matter created in the collision. However, an anomalous suppression was observed later on by the NA50 and NA60 experiments.

The suppression has been a subject of study since then. A number of explanations like multiple scattering, gluon distribution changes, excited-state decays, heavy quark/gluon energy loss etc. have been provided. Further and more careful studies are needed.

Figure 7 shows the production of *J/Ψ* mesons measured by several experiments in various reactions and as a function of the energy density reached in the collision. The measured cross-section for *J/Ψ* production were divided by the values expected from nuclear absorption. One can see that in lead–lead interactions the production is suppressed according to the expected nuclear absorption for energy densities below $2.2\ \mathrm{GeV/fm}^3$. As higher energy densities are obtained, the suppression starts to become important. This may be the result of charmonium melting, i.e., a manifestation of QGP appearance.

Jet quenching

The phenomenon of jet quenching was proposed in 1982 by J. D. Bjorken [10] as the result of energy loss of quarks propagating through a quark–gluon plasma. In his paper [10] Bjorken says:

High energy quarks and gluons propagating through a quark gluon plasma suffer differential energy loss via elastic scattering from quanta in the plasma. An interesting signature may be events in which

the hard collision occurs near the edge of the overlap region, with one jet escaping without absorption and the other fully absorbed.

First evidence of parton energy loss has been observed at RHIC [11]. Observation of high p_T hadron spectra and jet production in central Au–Au collisions and d–Au collisions confirmed the prediction of jet quenching.

Figure 8 show the azimuthal dependence of jets of particles. One sees clearly the presence of two jets in opposite directions in *proton–proton* and d–Au interactions. In Au–Au central collisions, however, one of the jets disappears. To obtain the plot in Fig. 8, one takes the highest transverse momentum track, which is between 4 and 6 GeV and then plots the tracks with transverse momentum in the interval $2\ \text{GeV} < p_T < p_T^{trigger}$ associated with the azimuth $\Delta\phi$.

The strong suppression of pion production at p_T up to 20 GeV has been observed at PHENIX [12] while direct photons which do not carry colour charge are not suppressed. The pions are generated by a fragmenting quark which does interact with the surrounding via its colour charge.

The magnitude of the measured suppression at high p_T and jet-like angular correlations in central Au–Au collisions suggest that the initial energy density of the created medium is significantly larger than normal nuclear density.

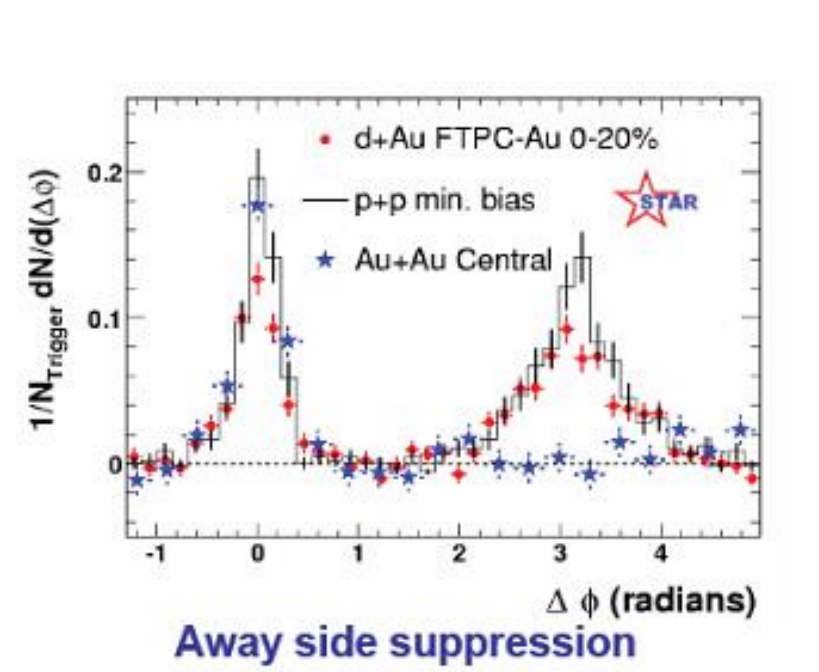


Fig. 8: The azimuthal correlations of charged particles relative to a high p_T trigger particle [11]. The jet outgoing in opposite direction for central Au–Au collisions (blue stars) is suppressed compared to the proton–proton (black histogram) and d–Au (red dots).

7 The Large Hadron Collider

The Large Hadron Collider (LHC) [13] accelerates protons in a 27 km long tunnel located at the European Organization for Nuclear Research (CERN) in Geneva, Switzerland. The LHC will also accelerate lead ions to make them collide at the highest energy ever.

The acceleration process starts in Linac 2 for protons and Linac 3 for lead ions. The protons accelerated in Linac 2 are injected into a Proton Synchrotron Booster with an energy of 50 MeV. In the synchrotron, protons reach an energy of 1.4 GeV. The Super Proton Synchrotron (SPS) has been modified to deliver a high-brightness proton beam required by the LHC. The SPS takes 26 GeV protons from the Proton Synchrotron (PS) and brings them to 450 GeV before extraction.

The Linac 3 produces 4.2 MeV/u lead ions. Linac 3 was commissioned in 1994 by an international collaboration and upgraded in 2007 for the LHC. The Low Energy Ion Ring (LEIR) is used as a storage and cooler unit providing ions to the (PS) with an energy of 72 MeV/nucleon. Ions will be further accelerated by the PS and the SPS before they are injected into the LHC where they reach an energy of 2.76 TeV/nucleon.

The LHC consists of 1232 superconducting dipole magnets with double aperture that operate at up to 9 Tesla magnetic field. The accelerator also includes more than 500 quadrupole magnets and more than 4000 corrector magnets of many types.

Ions are obtained from purified lead that is heated to 550 $^{\circ}$C. The lead vapour is then ionized with an electric current that produces various charge states. The Pb^{27+} ions are then selected with magnetic fields. This process takes place in an Electron Cyclotron Resonance (ECR) source (Fig. 9).

The ECR lead source is equipped with an hexapole permanent magnet. The plasma chamber is immersed in a solenoidal magnetic field. Pulsed beam currents produce Pb^{27+} ions that are then extracted to the Linac.

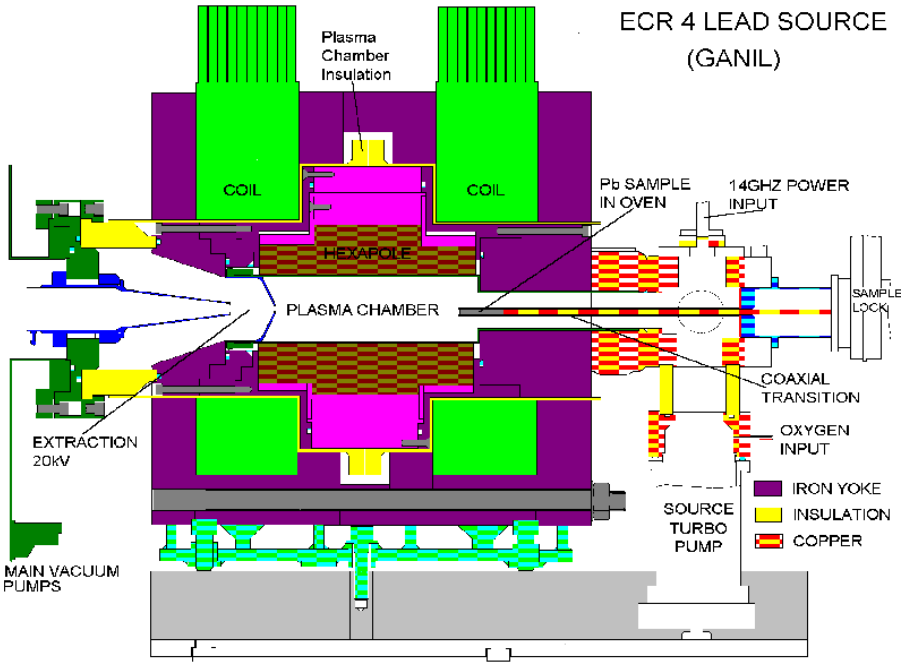


Fig. 9: Electron Cyclotron Resonance (ECR) ion source.

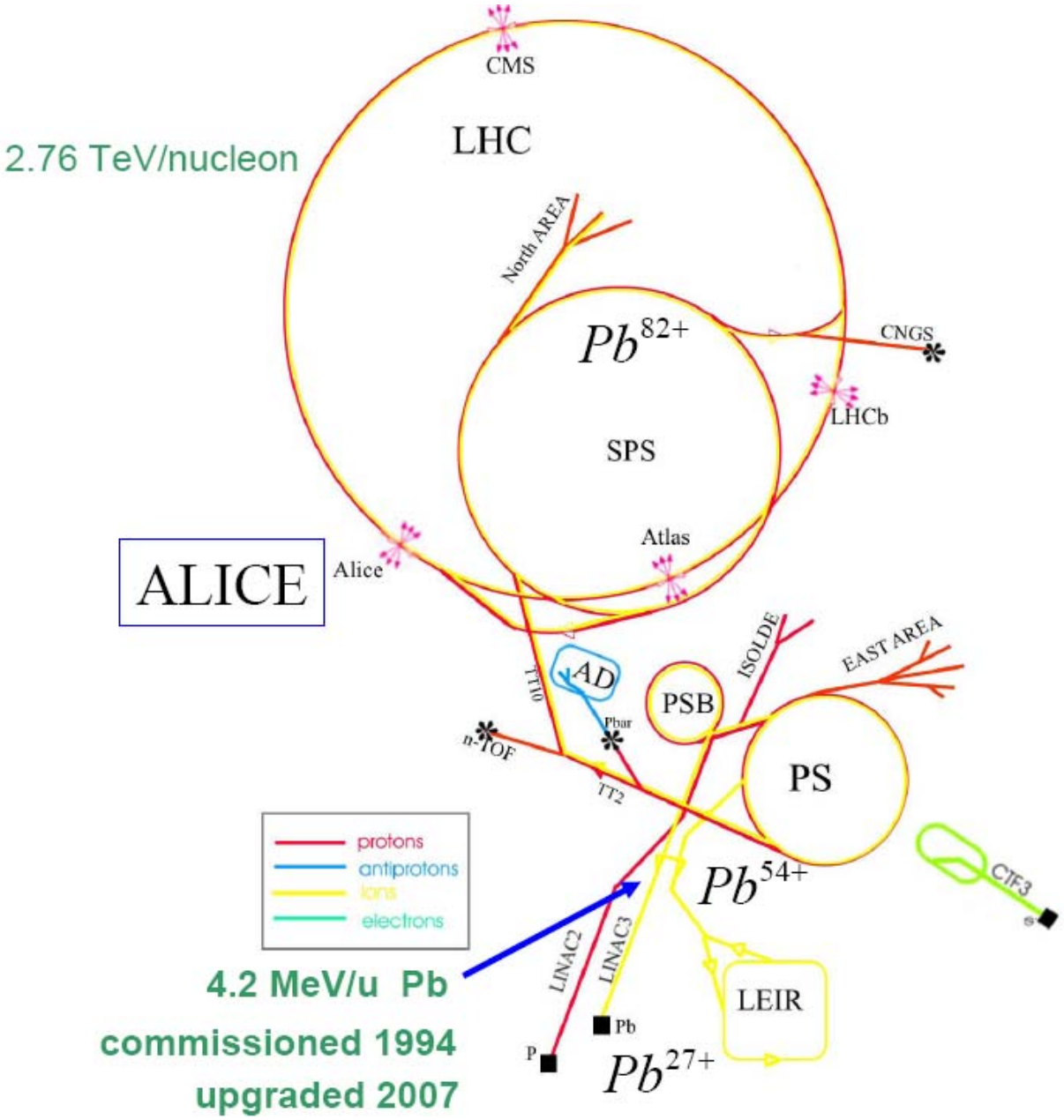


Fig. 10: Accelerators at CERN. The process of acceleration starts in Linac 2 and Linac 3 for protons and ions respectively.

After acceleration, the lead ions go through a carbon foil that strips them to Pb^{54+}, which are accumulated in the Low Energy Ion Ring (LEIR). LEIR is a circular machine which transforms the long pulses of Linac 3 into high-density bunches needed by the LHC. LEIR injects bunches of ions to the PS.

At the SPS, ions go once more through a thin aluminium foil which strips them to Pb^{82+}. The thickness of the stripper foil has to be chosen carefully to reduce contamination of lower charge states and keep emittance low. Foils of 0.5 to 1 mm thickness have been studied. In this way, fully stripped lead ions are obtained for the LHC.

Figure 10 shows the accelerators at CERN that are in use for the LHC.

The total cross-section of proton–proton interaction at 7 TeV could be inferred from hadronic cross-section measurements at lower energy [14]. It would be around 110 mbarn and correspond to about 60 mbarn of inelastic-scattering cross-section. The accelerator, at its design level, will reach a luminosity of $10^{34}\ \mathrm{s}^{-1}\ \mathrm{cm}^{-1}$ which means that the interaction rate will be

$$rate = 10^{34}\,(1/\mathrm{cm}^2\,\mathrm{s}) \times 60 \times 10^{-3}\,\mathrm{barn} \times 10^{-24}\ \mathrm{cm}^2/\mathrm{barn} = 600 \times 10^6\ \mathrm{collisions/s}\,.$$

A 25 ns interval between bunches gives a 40 MHz crossing rate. On average 19 inelastic events will occur each time bunches cross. Since there will be gaps in the beam structure, an average crossing rate of 31.6 MHz will be reached. Detectors at the LHC must be designed to cope with these frequencies. However, ALICE will run at a modest 300 kHz interaction rate in proton–proton mode and 10 kHz in Pb–Pb.

During autumn 2009, bunches of protons will be injected into the LHC ring. During the start-up phase, first collisions with protons at 3.5 TeV will take place. An increase of the proton beam energy in a second phase is foreseen. By the end of the run with protons in year 2010, lead-ion collisions will be produced.

The ALICE experiment is ready to take data on all the phases of the accelerator operation.

8 A Large Ion Collider Experiment

The ALICE experiment has been designed to observe the transition of ordinary matter into a plasma of quarks and gluons [15]. At the energies achieved by the LHC, the density, the size, and the lifetime of the excited quark matter will be high enough to allow a careful investigation of the properties of this new state of matter. The temperature will exceed by far the critical value predicted for the transition to take place.

ALICE has been optimized to study global event features. The number of colliding nucleons will provide information on the energy density achieved. The measurement of elliptic flow patterns will provide information about thermalization on the partonic level and the equation of state of the system in the high-density phase. Particle ratios in the final state are connected to chemical equilibration and provide a landmark on the trajectory of the system in the phase diagram. The spacetime evolution of the system can be investigated via particle interferometry, complemented by the study of resonaces. Moreover, important information about the system properties can be obtained by the study of hard probes, which will be produced abundantly at the LHC. Deconfinement may be reflected in the abundancies of J/ψ and Upsilon. The study of jet production on an event-by-event basis will allow one to investigate the transport properties of hard-scattered partons in the medium, which are expected to be strongly modified if a quark–gluon plasma is formed.

ALICE is also well suited for studies of proton–proton and photon–photon reactions. Photon–photon reactions include QED and QCD processes that go from lepton-pair to hadron and jet

production. As for proton–proton interactions, diffractive physics would be an exciting area of research.

The ALICE detector will have a tracking system over a wide range of transverse momentum which goes from 100 MeV/*c* to 100 GeV/*c* as well as particle identification able to separate pions, kaons, protons, muons, electrons, and photons.

A longitudinal view of the ALICE detector is shown in Fig. 11. A detailed description of the ALICE detector can be found in Ref. [16].

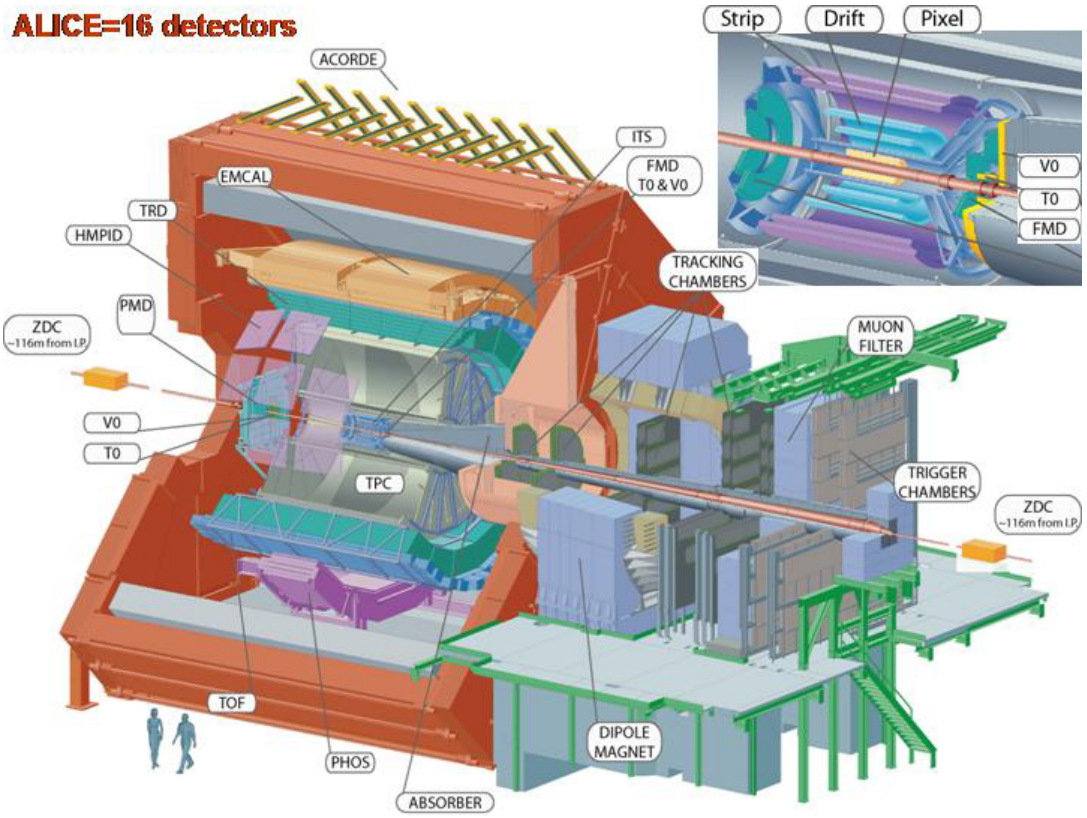


Fig. 11: The ALICE experiment consists of 16 detector subsystems. It combines particle identification, tracking, calorimetry, and trigger detectors.

In the forward direction a set of tracking chambers inside a dipole magnet will measure muons. An absorber will stop all the products of the interaction except for the muons which travel across and reach the tracking chambers that form the muon arm.

The central part of the ALICE detector is located inside a solenoid that provides a magnetic field of 0.5 T. The central tracking and particle identification system cover $-0.9 < \eta < 0.9$.

Electrons and photons are measured in the central region: photons will be measured in PHOS, a high-resolution calorimeter 5 m below the interaction point. The PHOS is built from $PBWO_4$ crystals which have a high light output.

The track measurement is performed with a set of six barrels of silicon detectors and a large Time Projection Chamber (TPC). The TPC has an effective volume of 88 m^3. It is the largest TPC ever built. These detectors will make available information on the energy loss allowing particle identification too. In addition to this, a Transition Radiation Detector (TRD) and a Time-of-Flight system will provide excellent particle separation at intermediate momentum. The Time-of-Flight system (TOF) uses Multi-gap Resistive Plate Chambers (MRPCs) with a total of 160 000 readout channels. A Ring Imaging Cherenkov detector will extend the particle identification capability to higher momentum particles. It covers 15% of the acceptance in the central area and will separate pions from kaons with momenta up to 3 GeV/*c* and kaons from protons with momenta up to 5 GeV/*c*.

A Forward Multiplicity Detector (FMD) consisting of silicon strip detectors and a Zero Degree Calorimeter (ZDC) will cover the very forward region providing information on the charge multiplicity and energy flow. A honeycomb proportional counter for photon multiplicity (PMD) measurements is located in the forward direction on one side of the ALICE detector.

The trigger system is complemented by a high level trigger (HLT) system which makes use of a computer farm to select events after read-out. In addition, the HLT system provides a data quality monitoring.

The V0 system is formed by two scintillation counters on each side of the interaction point. The system will be used as the main interaction trigger. In the top of the magnet, A Cosmic Ray Detector (ACORDE) will signal the arrival of cosmic muons. We briefly describe these two systems as examples of devices now in operation in the ALICE detector.

8.1 The V0 detector

The V0 system consists of two detectors: V0A and V0C, located in the central part of ALICE. The V0A is installed at a distance of 328 cm from the interaction point as shown in Fig. 12, mounted in two rigid half-boxes around the beam pipe. Each detector is an array of 32 cells of plastic scintillator, distributed in 4 rings forming a disc with 8 sectors. For the V0C, the cells of rings 3 and 4 are divided into two identical pieces that will be read with a single photomultiplier. This is done to achieve uniformity of detection and a small time fluctuation.

In proton–proton mode the mean number of charged particles within 0.5 units of rapidity is about 3. Each ring covers approximately 0.5 units of rapidity. The particles coming from the main vertex will interact with other components of the detector generating secondary particles. In general, each cell of the V0 detector will, on average, register one hit. For this reason the detector should have a very high efficiency. In Pb–Pb collisions the number of particles in a similar pseudo-rapidity range could be up to 4000 once secondary particles are included. Comparing the number of hits in the detector for proton–proton versus Pb–Pb mode, we can see that the required dynamic range will be 1–500 minimum-ionizing particles.

The Hamamatsu photomultiplier tubes (PMT) are installed inside the magnet not far from the detector. In order to tolerate the magnetic field, fine mesh tubes have been chosen. The segments of the V0A detector were constructed with a megatile technique [17]. This technique consists of machining the plastic scintillator and filling the grooves with TiO_2 loaded epoxy in order to separate one sector from the other.

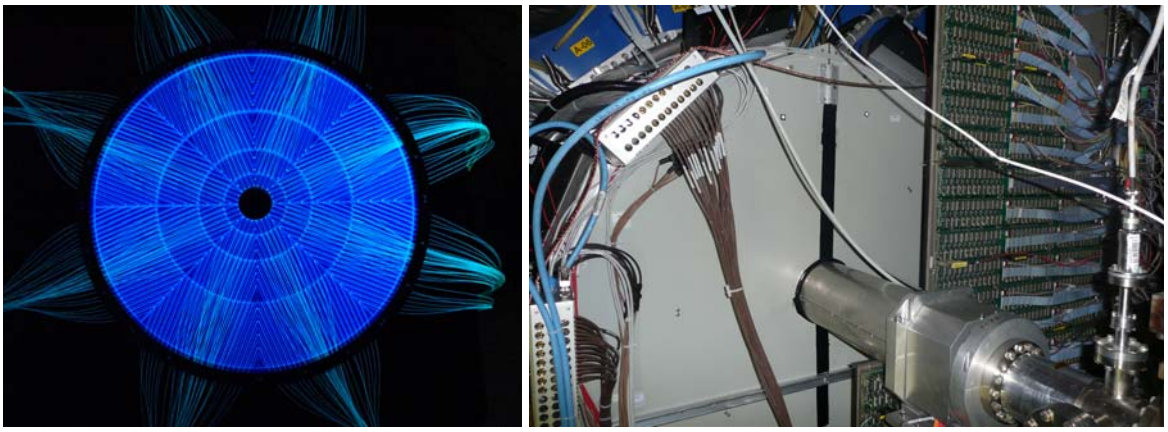


Fig.12: The V0A before optical isolation (left). The segmentation and the optical fibres are visible. On the right side, V0A in its box placed in the final position around the beam pipe. One half of the PMD just in front of the V0A can be seen in this picture.

A detailed description of the V0 system can be found in Ref. [18]. Figure 12 shows the V0A detector in its mechanical structure.

8.2 A Cosmic-Ray Detector ACORDE

The cosmic-ray detector consists of an array of 60 scintillator counters located in the upper part of the ALICE magnet [19]. The plastic used for the construction of the detector was part of the DELPHI detector. The material was carefully studied and the design of the detector was made according to the capabilities of the plastic available. The material was transported to Mexico where the construction was done.

Each module has a sensitive area of 1.9×0.195 m^2 and is built with two superimposed plastics. The doublet has an efficiency around 90% along the module.

The cosmic-ray detector

- Generates a single muon trigger to calibrate the Time Projection Chamber and other components of ALICE.
- Generates a multi-muon trigger to study cosmic rays with the help of tracking systems like the ITS and the TPC.
- Provides a wake-up signal for the Transition Radiation Detector.

The geometry is shown in Fig. 13. Modules on the far ends of the inner and outer faces of the magnet were moved to the centre of the upper face in order to have a much better efficiency for single muons.

Figure 13 shows a real cosmic-ray event reconstructed with the Time Projection Chamber and projected to ACORDE on the top of the magnet. This event contains 52 muons that fired 38 modules of ACORDE. It was recorded during the cosmic data run in October 2008.

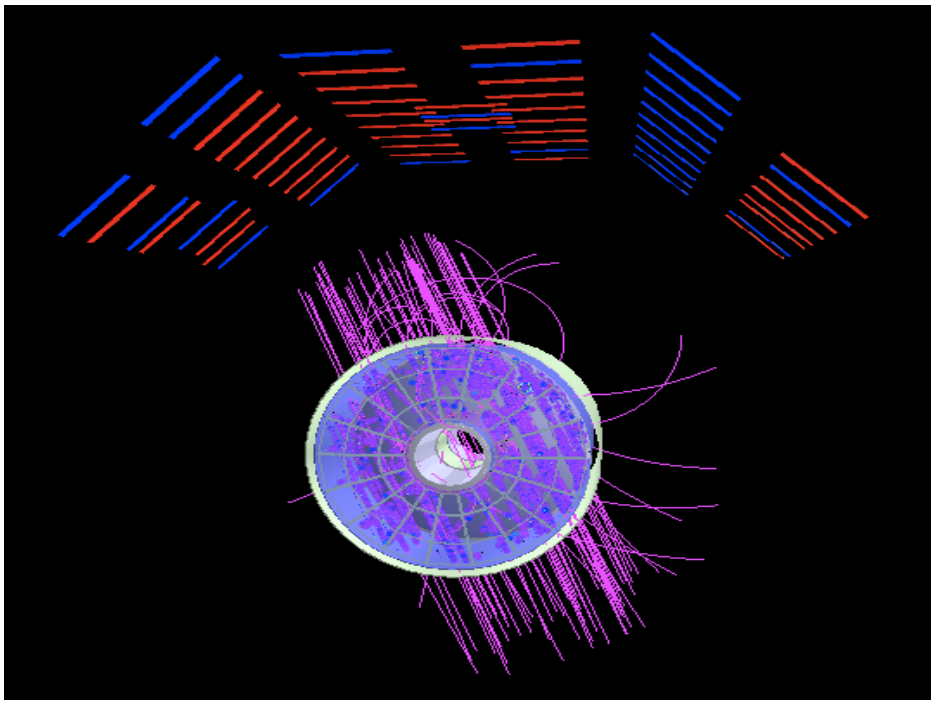


Fig.13: ACORDE modules can be seen in the top. This event was taken during a cosmic data run in October 2008. The cosmic-ray detector triggered the TPC to register 52 muons in one single event.

In 2009 a period of two months of cosmic studies will be conducted. The cosmic-ray detector will play a crucial role in triggering interesting events like the one shown here. The commissioning of several systems will be done during this period but interesting physics could be a bonus before accelerator activities start later on this year.

Acknowledgement

I want to thank the organizers of the School for the kind invitation to deliver these lectures. In particular I would like to thank Danielle Metral-Lillestol and Egil Lillestol for their kind support and warm hospitality.

References

[1] R. Hagedorn, *Suppl. Nuovo Cim.* **3** (1965) 147.

[2] W. Greiner, S. Schramm, and E. Stein, *Quantum Chromodynamics*, 2nd edition (Springer Verlag, Berlin, 2002) ISBN 3-540-66610-9.

[3] T. Ericson and J. Rafelski, *CERN Courier*, September 2003 also available in the web site: http://cerncourier.com/cws/article/cern/28919.

[4] P. Shukla, arXiv:nucl-th/0112039v1, 13 December 2001.

[5] J. D. Bjorken, *Phys. Rev. D* **27** (1983) 140.

[6] A. Chodos, R.L. Jaffe, K. Johnson, C.B. Thorn, and V. Weisskopf, *Phys. Rev. D* **9** (1974) 3471,

A. Chodos, R.L. Jaffe, K. Johnson and C.B. Thorn, *Phys. Rev. D* ***10*** (1974) 2599,

T. DeGrand, R.L. Jaffe, K. Johnson, and J. Kiskis, *Phys. Rev. D* **12** (1975) 2060.

[7] G. Herrera and A. Gago, *Mod. Phys. Lett. A* **10** (1995) 1435,

PHENIX Collaboration, *Phys Rev. Lett.* **88** (2002) 192302.

[8] T. Matsui and H. Satz, *Phys. Lett. B* **178** (1986) 416.

[9] http://na50.web.cern.ch/NA50/

[10] J. D. Bjorken, Energy loss of energetic partons in quark gluon plasma: possible extinction of high–pt jets in hadron–hadron collisions, FERMILAB-Pub-82/59 –THY, August 1982.

[11] PHENIX Collaboration, *Phys. Rev. Lett.* **88** (2002) 022301,

STAR Collaboration, *Phys. Rev. Lett.* **90** (2003) 082302.

[12] S. Adler *et al.*, PHENIX Collaboration, *Phys. Rev. Lett.* **91** (2003) 072301.

[13] L. Evans, *Eur. Phys. J. C* **34** (2004) 57,

Proceedings of the ECFA-CERN Workshop on the Large Hadron Collider in the LEP tunnel, CERN 84-10 (1984).

[14] Particle Data Group, *Phys. Lett. B* **667** (2008) 1,

J.R. Cudell, *et al.* (COMPETE Collaboration), *Phys. Rev. D* **65** (2002) 074024.

[15] ALICE Collaboration, *J. Phys. Nucl. Part. Phys. G* **30** (2004) 1517–1763,

J. Phys. Nucl. Part. Phys. G **32** (2006) 1295–2040.

[16] ALICE : Technical Proposal for A Large Ion Collider Experiment at the CERN LHC,

N. Ahmad *et al.,* CERN/LHCC/95-71 (1995).

[17] S. Kim, *Nucl. Instrum. Methods Phys Res. A* **360** (1995) 206.

[18] ALICE Technical Design Report, CERN-LHCC-2004-025 ALICE TDR 011 (10 September 2004),

R. Alfaro *et al.*, ALICE Internal Note, ALICE-INT-2003-040, version 1.0,

R. Alfaro-Molina *et al.*, ALICE Internal Note, ALICE-INT-2006-018.

[19] A. Fernández *et al.*, *Nucl. Instrum. Methods Phys. Res. A* **572** (2007) 102,

Czech. J. Phys. **55** (2005) *B* 801–*B* 807.

Bibliography

M. Kliemant, R. Sahoo, T. Schuster and R. Stock, Global properties of nucleus nucleus collisions. arXiv:0809.2482[nucl-ex].

Cheuk–Yin Wong, *Introduction to High Energy Heavy-Ion Collisions* (World Scientific, Singapore,1994) ISBN 9810202636.

J. Lettessier and J. Rafelski, *Hadrons and Quark–Gluon Plasma,* Cambridge Monogaphs on Particle Physics, Nuclear Physics and Cosmology (Cambridge University Press, 2005) ISBN 0 521 01823 4.

Trigger and data acquisition

N. Ellis
CERN, Geneva, Switzerland

Abstract
The lectures address some of the issues of triggering and data acquisition in large high-energy physics experiments. Emphasis is placed on hadron-collider experiments that present a particularly challenging environment for event selection and data collection. However, the lectures also explain how T/DAQ systems have evolved over the years to meet new challenges. Some examples are given from early experience with LHC T/DAQ systems during the 2008 single-beam operations.

1 Introduction

These lectures concentrate on experiments at high-energy particle colliders, especially the general-purpose experiments at the Large Hadron Collider (LHC) [1]. These experiments represent a very challenging case that illustrates well the problems that have to be addressed in state-of-the-art high-energy physics (HEP) trigger and data-acquisition (T/DAQ) systems. This is also the area in which the author is working (on the trigger for the ATLAS experiment at LHC) and so is the example that he knows best. However, the lectures start with a more general discussion, building up to some examples from LEP [2] that had complementary challenges to those of the LHC. The LEP examples are a good reference point to see how HEP T/DAQ systems have evolved in the last few years.

Students at this school come from various backgrounds — phenomenology, experimental data analysis in running experiments, and preparing for future experiments (including working on T/DAQ systems in some cases). These lectures try to strike a balance between making the presentation accessible to all, and going into some details for those already familiar with T/DAQ systems.

1.1 Definition and scope of trigger and data acquisition

T/DAQ is the online system that selects particle interactions of potential interest for physics analysis (trigger), and that takes care of collecting the corresponding data from the detectors, putting them into a suitable format and recording them on permanent storage (DAQ). Special modes of operation need to be considered, e.g., the need to calibrate different detectors in parallel outside of normal data-taking periods. T/DAQ is often taken to include associated tasks, e.g., run control, monitoring, clock distribution and book-keeping, all of which are essential for efficient collection and subsequent offline analysis of the data.

1.2 Basic trigger requirements

As introduced above, the trigger is responsible for selecting interactions that are of potential interest for physics analysis. These interactions should be selected with high efficiency, the efficiency should be precisely known (since it enters in the calculation of cross-sections), and there should not be biases that affect the physics results. At the same time, a large reduction of rate from unwanted high-rate processes may be needed to match the capabilities of the DAQ system and the offline computing system. High-rate processes that need to be rejected may be instrumental backgrounds or high-rate physics processes that are not relevant for the analyses that one wants to make. The trigger system must also be affordable, which implies limited computing power. As a consequence, algorithms that need to be executed at high rate must be fast. Note that it is not always easy to achieve the above requirements (high efficiency for signal, strong background rejection and fast algorithms) simultaneously.

Trigger systems typically select events[1] according to a 'trigger menu', i.e., a list of selection criteria — an event is selected if one or more of the criteria are met. Different criteria may correspond to different signatures for the same physics process — redundant selections lead to high selection efficiency and allow the efficiency of the trigger to be measured from the data. Different criteria may also reflect the wish to concurrently select events for a wide range of physics studies — HEP 'experiments' (especially those with large general-purpose 'detectors' or, more precisely, detector systems) are really experimental facilities. Note that the menu has to cover the physics channels to be studied, plus additional data samples required to complete the analysis (e.g., measure backgrounds, and check the detector calibration and alignment).

1.3 Basic data-acquisition requirements

The DAQ system is responsible for the collection of data from detector digitization systems, storing the data pending the trigger decision, and recording data from the selected events in a suitable format. In doing so it must avoid corruption or loss of data, and it must introduce as little dead-time as possible ('dead-time' refers to periods when interesting interactions cannot be selected — see below). The DAQ system must, of course, also be affordable which, for example, places limitations on the amount of data that can be read out from the detectors.

2 Design of a trigger and data-acquisition system

In the following a very simple example is used to illustrate some of the main issues for designing a T/DAQ system. An attempt is made to omit all the detail and concentrate only on the essentials — examples from real experiments will be discussed later.

Before proceeding to the issue of T/DAQ system design, the concept of dead-time, which will be an important element in what follows, is introduced. 'Dead-time' is generally defined as the fraction or percentage of total time where valid interactions could not be recorded for various reasons. For example, there is typically a minimum period between triggers — after each trigger the experiment is dead for a short time.

Dead-time can arise from a number of sources, with a typical total of up to $\mathcal{O}(10\%)$. Sources include readout and trigger dead-time, which are addressed in detail below, operational dead-time (e.g., time to start/stop data-taking runs), T/DAQ downtime (e.g., following a computer failure), and detector downtime (e.g., following a high-voltage trip). Given the huge investment in the accelerators and the detectors for a modern HEP experiment, it is clearly very important to keep dead-time to a minimum.

In the following, the design issues for a T/DAQ system are illustrated using a very simple example. Consider an experiment that makes a time-of-flight measurement using a scintillation-counter telescope, read out with time-to-digital converters (TDCs), as shown in Fig. 1. Each plane of the telescope is viewed by a photomultiplier tube (PMT) and the resulting electronic signal is passed to a 'discriminator' circuit that gives a digital pulse with a sharp leading edge when a charged particle passes through the detector. The leading edge of the pulse appears a fixed time after the particle traverses the counter. (The PMTs and discriminators are not shown in the figure.)

Two of the telescope planes are mounted close together, while the third is located a considerable distance downstream giving a measurable flight time that can be used to determine the particle's velocity. The trigger is formed by requiring a coincidence (logical AND) of the signals from the first two planes, avoiding triggers due to random noise in the photomultipliers — it is very unlikely for there to be noise pulses simultaneously from both PMTs. The time of arrival of the particle at the three telescope planes is measured, relative to the trigger signal, using three channels of a TDC. The pulses going to the TDC from each of the three planes have to be delayed so that the trigger signal, used to start the TDC (analogous to starting a stop-watch), gets there first.

[1] The term 'event' will be discussed in Section 3 — for now, it may be taken to mean the record of an interaction.

The trigger signal is also sent to the DAQ computer, telling it to initiate the readout. Not shown in Fig.1 is logic that prevents further triggers until the data from the TDC have been read out into the computer — the so-called dead-time logic.

2.1 Traditional approach to trigger and data acquisition

The following discussion starts by presenting a 'traditional' approach to T/DAQ (as might be implemented using, for example, NIM and CAMAC electronics modules[2], plus a DAQ computer). Note that this approach is still widely used in small test set-ups. The limitations of this model are described and ways of improving on it are presented. Of course, a big HEP experiment has an enormous number of sensor channels [up to $\mathcal{O}(10^8)$ at LHC], compared to just three in the example. However, the principles are the same, as will be shown later.

Limitations of the T/DAQ system shown in Fig. 1 are as follows:

1. The trigger decision has to be made very quickly because the TDCs require a 'start' signal that arrives before the signals that are to be digitized (a TDC module is essentially a multichannel digital stop-watch). The situation is similar with traditional analog-to-digital converters (ADCs) that digitize the magnitude of a signal arriving during a 'gate' period, e.g., the electric charge in an analog pulse — the gate has to start before the pulse arrives.
2. The readout of the TDCs by the computer may be quite slow, which implies a significant dead-time if the trigger rate is high. This limitation becomes much more important in larger systems, where many channels have to be read out for each event. For example, if 1000 channels have to be read out with a readout time of 1 µs per channel (as in CAMAC), the readout time per event is 1 ms which excludes event rates above 1 kHz.

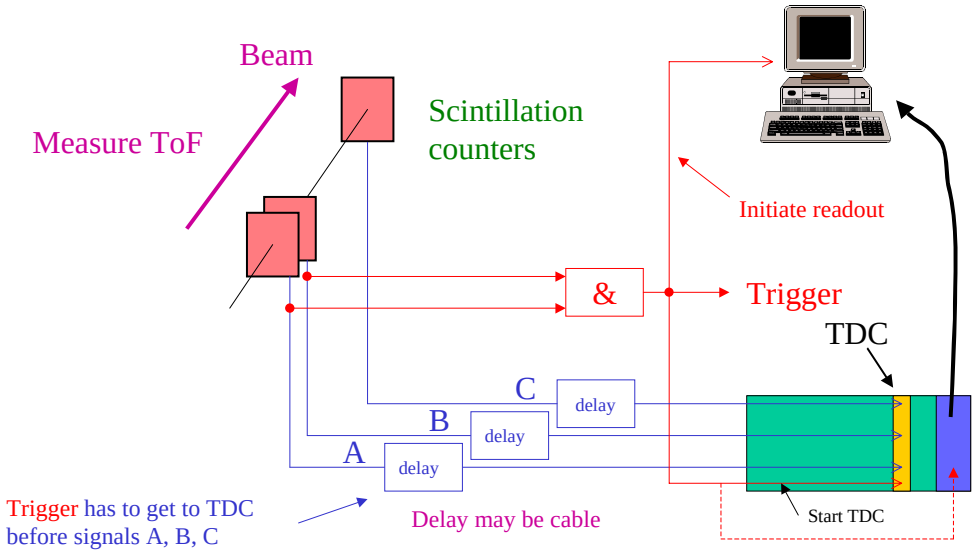


Fig. 1: Example of a simple experiment with its T/DAQ system

The 'readout model' of this traditional approach to T/DAQ is illustrated in Fig. 2, which shows the sequence of actions — arrival of the trigger, arrival of the detector signals (followed by digitization and storage in a data register in the TDC), and readout into the DAQ computer. Since no new trigger can be accepted until the readout is complete, the readout dead-time is given by the product of the trigger rate and the readout time.

[2]NIM [3] and CAMAC [4] modules are electronic modules that conform to agreed standards — modules for many functions needed in a T/DAQ system are available commercially.

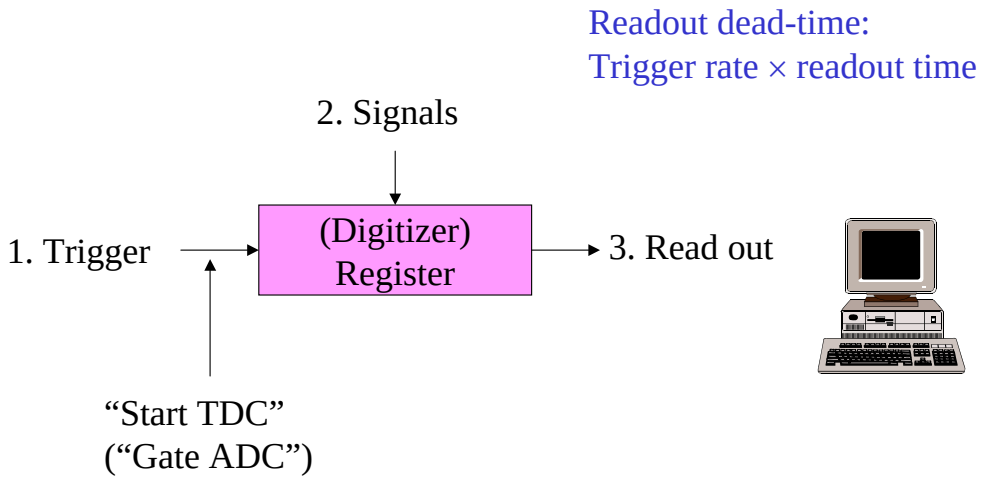


Fig. 2: Readout model in the 'traditional' approach

2.2 Local buffer

The traditional approach described above can be improved by adding a local 'buffer' memory into which the data are moved rapidly following a trigger, as illustrated in Fig. 3. This fast readout reduces the dead-time, which is now given by the product of the trigger rate and the local readout time. This approach is particularly useful in large systems, where the transfer of data can proceed in parallel with many local buffers (e.g., one local buffer for each crate of electronics) — local readout can remain fast even in a large system. Also, the data may be moved more quickly into the local buffer within the crate than into the DAQ computer. Note that the dashed line in the bottom, right-hand part of Fig. 1 indicates this extension to the traditional approach — the trigger signal is used to initiate the local readout within the crate.

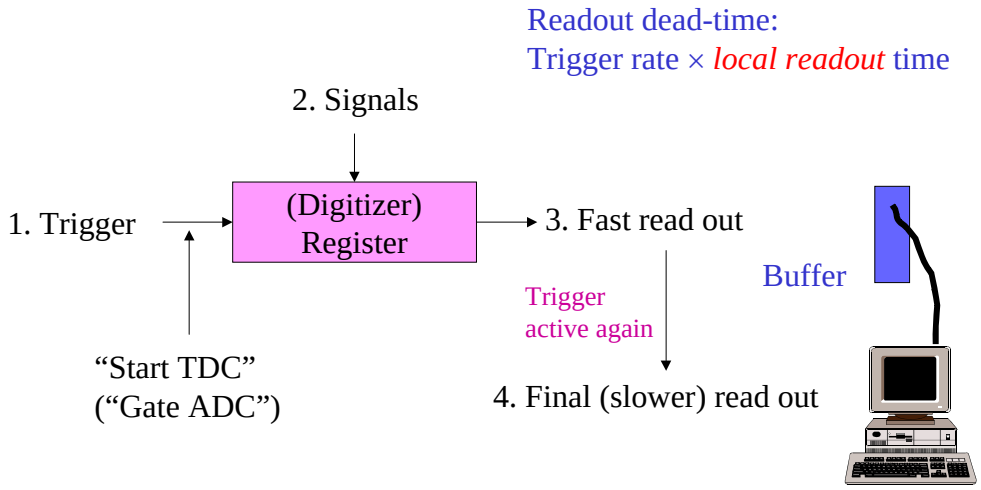


Fig. 3: Readout system with local buffer memory

The addition of a local buffer reduces the effective readout time, but the requirement of a fast trigger still remains. Signals have to be delayed until the trigger decision is available at the digitizers. This is not easy to achieve, even with very simple trigger logic — typically one relies on using fast (air-core) cables for trigger signals with the shortest possible routing so that the trigger signals arrive before the rest of the signals (which follow a longer routing on slower cables). It is not possible to apply complex selection criteria on this time-scale.

2.3 Multi-level triggers

It is not always possible to simultaneously meet the physics requirements (high efficiency, high background rejection) and achieve an extremely short trigger 'latency' (time to form the trigger decision and distribute it to the digitizers). A solution is to introduce the concept of multi-level triggers, where the first level has a short latency and maintains high efficiency, but only has a modest rejection power. Further background rejection comes from the higher trigger levels that can be slower. Sometimes the very fast first stage of the trigger is called the 'pre-trigger' — it may be sufficient to signal the presence of minimal activity in the detectors at this stage.

The use of a pre-trigger is illustrated in Fig. 4. Here the pre-trigger is used to provide the start signal to the TDCs (and to gate ADCs, etc.), while the main trigger (which can come later and can therefore be based on more complex calculations) is used to initiate the readout. In cases where the pre-trigger is not confirmed by the main trigger, a 'fast clear' is used to re-activate the digitizers (TDCs, ADCs, etc.).

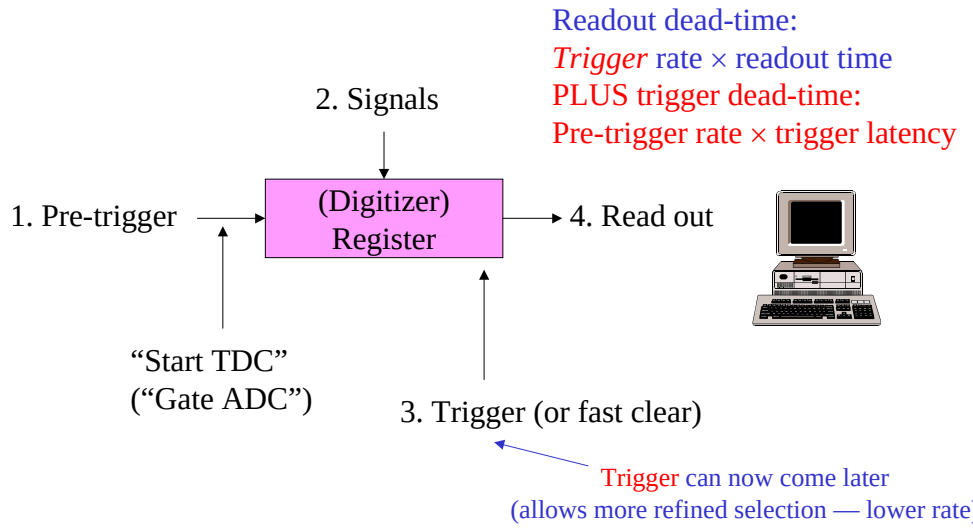


Fig. 4: Readout system with pre-trigger and fast clear

Using a pre-trigger (but without using a local buffer for now), the dead-time has two components. Following each pre-trigger there is a dead period until the trigger or fast clear is issued (defined here as the trigger latency). For the subset of pre-triggers that give rise to a trigger, there is an additional dead period given by the readout time. Hence, the total dead-time is given by the product of the pre-trigger rate and the trigger latency, added to the product of the trigger rate and the readout time.

The two ingredients — use of a local buffer and use of a pre-trigger with fast clear — can be combined as shown in Fig. 5, further reducing the dead-time. Here the total dead-time is given by the product of the pre-trigger rate and the trigger latency, added to the product of the trigger rate and the local readout time.

2.4 Further improvements

The idea of multi-level triggers can be extended beyond having two levels (pre-trigger and main trigger). One can have a series of trigger levels that progressively reduce the rate. The efficiency for the desired physics must be kept high at all levels since rejected events are lost forever. The initial levels can have modest rejection power, but they must be fast since they see a high input rate. The final levels must have strong rejection power, but they can be slower because they see a much lower rate (thanks to the rejection from the earlier levels).

In a multi-level trigger system, the total dead-time can be written as the sum of two parts: the trigger dead-time summed over trigger levels, and the readout dead-time. For a system with N levels, this can be written

$$(\sum_{i=2}^{N} R_{i-1} \times L_i) + R_N \times T_{\mathrm{LRO}}$$

where R_i is the rate after the i^{th} trigger level, L_i is the latency of the i^{th} trigger level, and T_{LRO} is the local readout time. Note that R_1 corresponds to the pre-trigger rate.

In the above, two implicit assumptions have been made: (1) that all trigger levels are completed before the readout starts, and (2) that the pre-trigger (i.e., the lowest-level trigger) is available by the time the first signals from the detector arrive at the digitizers. The first assumption results in a long dead period for some events — those that survive the first (fast) levels of selection. The dead-time can be reduced by moving the data into intermediate storage after the initial stages of trigger selection, after which further low-level triggers can be accepted (in parallel with the execution of the later stages of trigger selection on the first event). The second assumption can also be avoided, e.g., in collider experiments with bunched beams as discussed below.

In the next section, aspects of particle colliders that affect the design of T/DAQ systems are introduced. Afterwards, the discussion returns to readout models and dead-time, considering the example of LEP experiments.

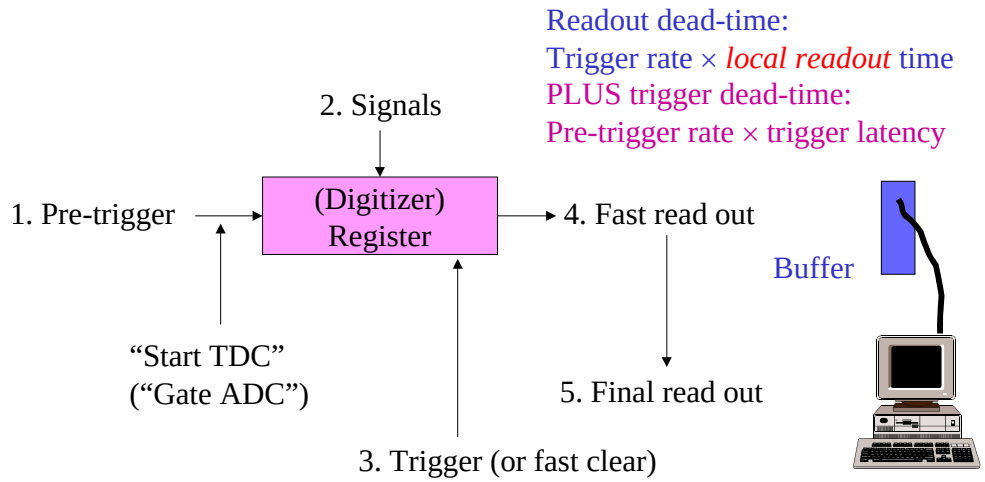


Fig. 5: Readout system using both pre-trigger and local buffer

3 Collider experiments

In high-energy particle colliders (HERA, LEP, LHC, Tevatron), the particles in the counter-rotating beams are bunched. Bunches of particles cross at regular intervals and interactions occur only during the bunch crossings. Here the trigger has the job of selecting the *bunch crossings* of interest for physics analysis, i.e., those containing interactions of interest.

In the following notes, the term 'event' is used to refer to the record of all the products from a given bunch crossing (plus any activity from other bunch crossings that gets recorded along with this). Be aware (and beware!) — the term 'event' is not uniquely defined! Some people use the term 'event' for the products of a single interaction between incident particles. Note that many people use 'event' interchangeably to mean different things.

In $\mathrm{e}^+\,\mathrm{e}^-$ colliders, the interaction rate is very small compared to the bunch-crossing rate (because of the low $\mathrm{e}^+\,\mathrm{e}^-$ cross-section). Generally, selected events contain just one interaction — i.e., the event

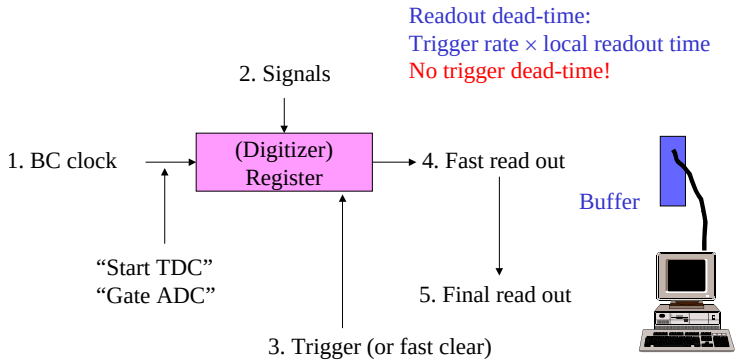


Fig. 6: Readout system using bunch-crossing (BC) clock and fast clear

is generally a single interaction. This was the case at LEP and also at the e–p collider HERA. In contrast, at LHC with the design luminosity $\mathscr{L}$ of $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ for proton beams, each bunch crossing will contain on average about 25 interactions as discussed below. This means that an interaction of interest, e.g., one that produced $\mathrm{H} \to \mathrm{ZZ} \to \mathrm{e}^+\mathrm{e}^-\mathrm{e}^+\mathrm{e}^-$, will be recorded together with 25 other proton–proton interactions that occurred in the same bunch crossing. The interactions that make up the 'underlying event' are often called 'minimum-bias' interactions because they are the ones that would be selected by a trigger that selects interactions in an unbiased way. The presence of additional interactions that are recorded together with the one of interest is sometimes referred to as 'pile-up'.

A further complication is that particle detectors do not have an infinitely fast response time — this is analogous to the exposure time of a camera. If the 'exposure time' is shorter than the bunch-crossing period, the event will contain only information from the selected bunch crossing. Otherwise, the event will contain, in addition, any activity from neighbouring bunches. In $\mathrm{e}^+\,\mathrm{e}^-$ colliders (e.g., LEP) it is very unlikely for there to be any activity in nearby bunch crossings, which allows the use of slow detectors such as the time projection chamber (TPC). This is also true at HERA and in the ALICE experiment [5] at LHC that will study heavy-ion collisions at much lower luminosities than in the proton–proton case.

The bunch-crossing period for proton–proton collisions at LHC will be only 25 ns (corresponding to a 40 MHz rate). At the design luminosity the interaction rate will be $\mathcal{O}(10^9)$ Hz and, even with the short bunch-crossing period, there will be an average of about 25 interactions per bunch crossing. Some detectors, for example the ATLAS silicon tracker, achieve an exposure time of less than 25 ns, but many do not. For example, pulses from the ATLAS liquid-argon calorimeter extend over many bunch crossings.

The instrumentation for the LHC experiments is described in the lecture notes of Jordan Nash from this School [6]. The Particle Data Group's Review of Particle Physics [7] includes much useful information, including summaries of the parameters of various particle colliders.

4 Design of a trigger and data-acquisition system for LEP

Let us now return to the discussion of designing a T/DAQ system, considering the case of experiments at LEP (ALEPH [8], DELPHI [9], L3 [10], and OPAL [11]), and building on the model developed in Section 2.

4.1 Using the bunch-crossing signal as a 'pre-trigger'

If the time between bunch crossings (BCs) is reasonably long, one can use the clock that signals when bunches of particles cross as the pre-trigger. The first-level trigger can then use the time between bunch crossings to make a decision, as shown in Fig. 6. For most crossings the trigger will reject the event by issuing a fast clear — in such cases no dead-time is introduced. Following an 'accept' signal, dead-time

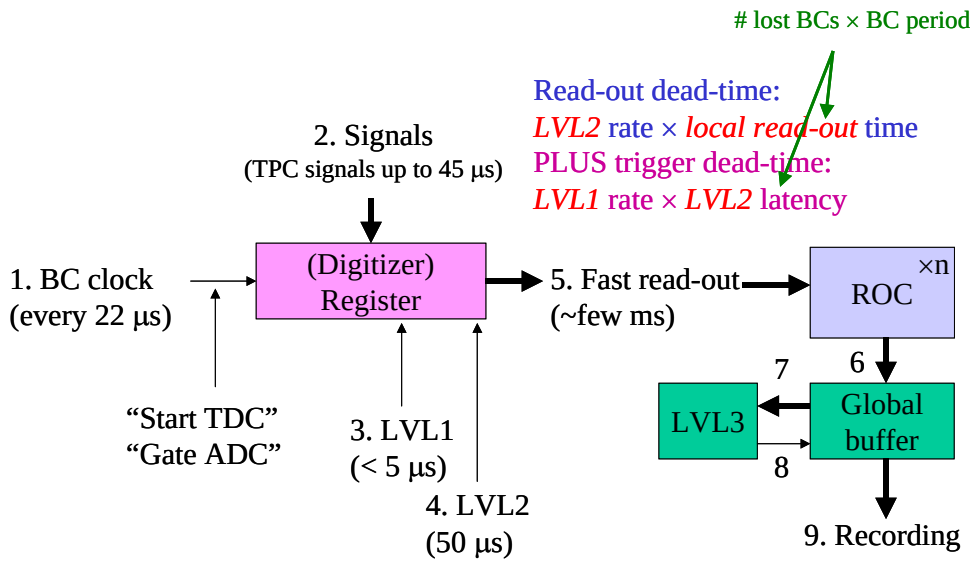


Fig. 7: LEP readout model (ALEPH)

will be introduced until the data have been read out (or until the event has been rejected by a higher-level trigger). This is the basis of the model that was used at LEP, where the bunch-crossing interval of 22 μs (11 μs in eight-bunch mode) allowed comparatively complicated trigger processing (latency of a few microseconds). Note that there is no first-level trigger dead-time because the decision is made during the interval between bunch crossings where no interactions occur. As discussed below, the trigger rates were reasonably low (very much less than the BC rate), giving acceptable dead-time due to the second-level trigger latency and the readout.

In the following, the readout model used at LEP is illustrated by concentrating on the example of ALEPH [8][3]. Figure 7 shows the readout model, using the same kind of block diagram as presented in Section 2. The BC clock is used to start the TDCs and generate the gate for the ADCs, and a first-level (LVL1) trigger decision arrives in less than 5 μs so that the fast clear can be completed prior to the next bunch crossing. For events retained by LVL1, a more sophisticated second-level (LVL2) trigger decision is made after a total of about 50 μs. Events retained by LVL2 are read out to local buffer memory (within the readout controllers or 'ROCs'), and then passed to a global buffer. There is a final level of selection (LVL3) before recording the data on permanent storage for offline analysis.

For readout systems of the type shown in Fig. 7, the total dead-time is given by the sum of two components — the trigger dead-time and the readout dead-time.

The trigger dead-time is evaluated by counting the number of BCs that are lost following each LVL1 trigger, then calculating the product of the LVL1 trigger rate, the number of lost BCs and the BC period. Note that the effective LVL2 latency, given by the number of lost BCs and the BC period, is less than (or equal to) the true LVL2 latency.

The readout dead-time is given by the product of the LVL2 trigger rate and the time taken to perform local readout into the ROCs. Strictly speaking, one should also express this dead-time in terms of the number of BCs lost after the LVL2 trigger, but since the readout time is much longer than the BC period the difference is unimportant. Note that, as long as the buffers in the ROCs and the global buffers do not fill up, no additional dead-time is introduced by the final readout and the LVL3 trigger.

Let us now look quantitatively at the example of the DELPHI experiment for which the T/DAQ

[3]The author was not involved in any of the LEP experiments. In these lectures the example of ALEPH is used to illustrate how triggers and data acquisition were implemented at LEP; some numbers from DELPHI are also presented. The T/DAQ systems in all of the LEP experiments were conceptually similar.

Table 1: Typical T/DAQ parameters for the DELPHI experiment at LEP-II

Quantity	Value
LVL1 rate	~ 500–1000 Hz (instrumental background)
LVL2 rate	6–8 Hz
LVL3 rate	4–6 Hz
LVL2 latency	38 μs (1 lost BC ⇒ 22 μs effective)
Local readout time	~ 2.5 ms
Readout dead-time	~ 7 Hz × 2.5 · 10^{-3} s = 1.8%
Trigger dead-time	~ 750 Hz × 22 · 10^{-6} s = 1.7%
Total dead-time	~ 3–4%

system was similar to that described above for ALEPH. Typical numbers for LEP-II[4] are shown in Table 1 [9].

4.2 Data acquisition at LEP

Let us now continue our examination of the example of the ALEPH T/DAQ system. Following a LVL2 trigger, events were read out locally and in parallel within the many readout crates — once the data had been transferred within each crate to the ROC, further LVL1 and LVL2 triggers could be accepted. Subsequently, the data from the readout crates were collected by the main readout computer, 'building' a complete event. As shown in Fig. 8, event building was performed in two stages: an event contained a number of sub-events, each of which was composed of several ROC data blocks. Once a complete event was in the main readout computer, the LVL3 trigger made a final selection before the data were recorded.

The DAQ system used a hierarchy of computers — the local ROCs in each crate; event builders (EBs) for sub-events; the main EB; the main readout computer. The ROCs performed some data processing (e.g., applying calibration algorithms to convert ADC values to energies) in addition to reading out the data from ADCs, TDCs, etc. (Zero suppression was already performed at the level of the digitizers where appropriate.) The first layer of EBs combined data read out from the ROCs of individual sub-detectors into sub-events; then the main EB combined the sub-events for the different sub-detectors. Finally, the main readout computer received full events from the main EB, performed the LVL3 trigger selection, and recorded selected events for subsequent analysis.

As indicated in Fig. 9, event building was bus based — each ROC collected data over a bus from the digitizing electronics; each sub-detector EB collected data from several ROCs over a bus; the main EB collected data from the sub-detector EBs over a bus. As a consequence, the main EB and the main readout computer saw the full data rate prior to the final LVL3 selection. At LEP this was fine — with an event rate after LVL2 of a few hertz and an event size of 100 kbytes, the data rate was a few hundred kilobytes per second, much less than the available bandwidth (e.g., 40 Mbytes/s maximum on VME bus [12]).

4.3 Triggers at LEP

The triggers at LEP aimed to select any $e^+ e^-$ annihilation event with a visible final state, including events with little visible energy, plus some fraction of two-photon events, plus Bhabha scattering events. Furthermore, they aimed to select most events by multiple, independent signatures so as to maximize the trigger efficiency and to allow the measurement of the efficiency from the data. The probability for an event to pass trigger A or trigger B is $\sim 1-\delta_A\delta_B$, where δ_A and δ_B are the individual trigger inefficiencies, which is very close to unity for small δ. Starting from a sample of events selected with trigger A, the

[4] LEP-II refers to the period when LEP operated at high energy, after the upgrade of the RF system.

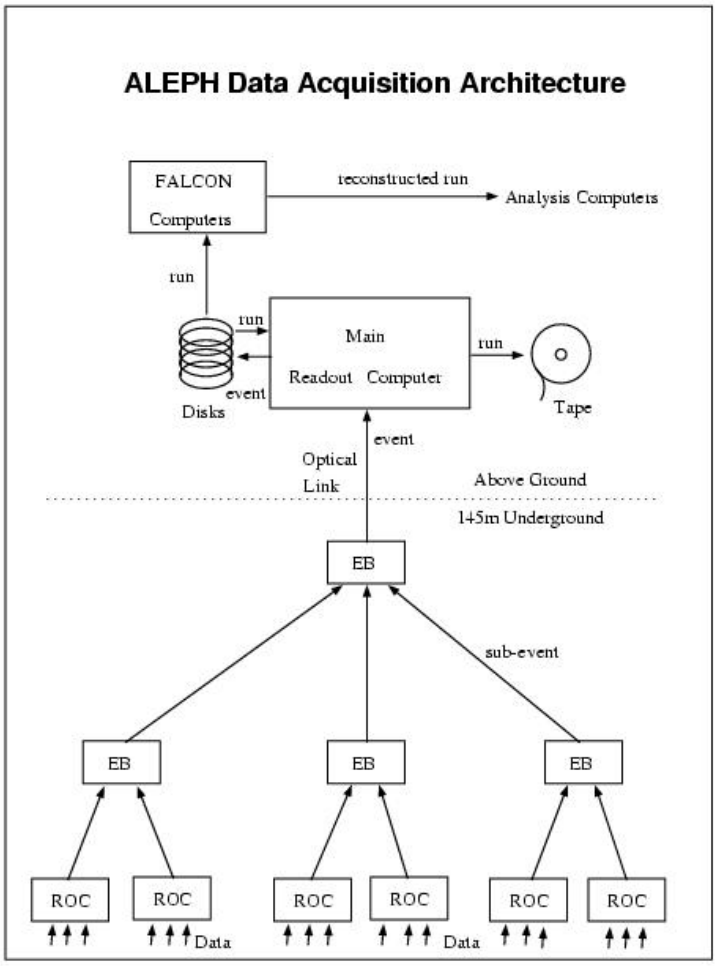


Fig. 8: ALEPH data-acquisition architecture

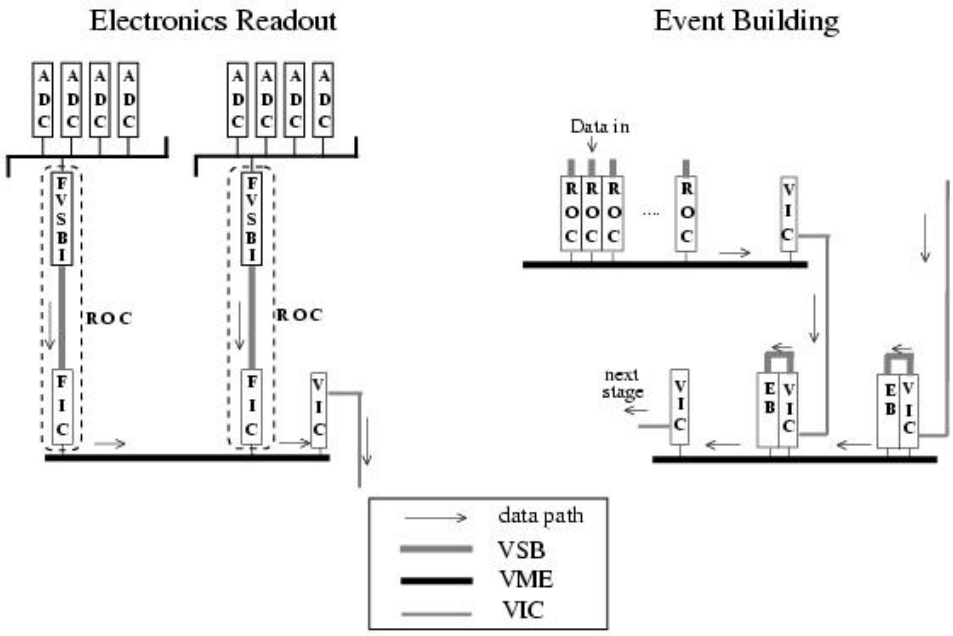


Fig. 9: Event building in ALEPH

efficiency of trigger B can be estimated as the fraction of events passing trigger B in addition. Note that in the actual calculations small corrections were applied for correlations between the trigger efficiencies.

5 Towards the LHC

In some experiments it is not practical to make a trigger in the time between bunch crossings because of the short BC period — the BC interval is 396 ns at Tevatron-II[5], 96 ns at HERA and 25 ns at LHC. In such cases the concept of 'pipelined' readout has to be introduced (also pipelined LVL1 trigger processing). Furthermore, in experiments at high-luminosity hadron colliders the data rates after the LVL1 trigger selection are very high, and new ideas have to be introduced for the high-level triggers (HLTs) and DAQ — in particular, event building has to be based on data networks and switches rather than data buses.

5.1 Pipelined readout

In pipelined readout systems (see Fig. 10), the information from each BC, for each detector element, is retained during the latency of the LVL1 trigger (several μs). The information may be retained in several forms — analog levels (held on capacitors); digital values (e.g., ADC results); binary values (i.e., hit or no hit). This is done using a logical 'pipeline', which may be implemented using a first-in, first-out (FIFO) memory circuit. Data reaching the end of the pipeline are either discarded or, in the case of a trigger accept decision, moved to a secondary buffer memory (small fraction of BCs).

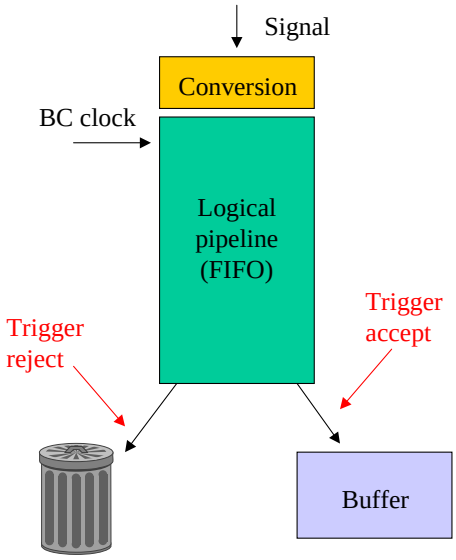


Fig. 10: Example of pipelined readout

Pipelined readout systems will be used in the LHC experiments (they have already been used in experiments at HERA [13, 14] and the Tevatron [15, 16], but the demands at LHC are even greater because of the short BC period). A typical LHC pipelined readout system is illustrated in Fig. 11, where the digitizer and pipeline are driven by the 40 MHz BC clock. A LVL1 trigger decision is made for each bunch crossing (i.e., every 25 ns), although the LVL1 latency is several microseconds — the LVL1 trigger must concurrently process many events (this is achieved by using pipelined trigger processing as discussed below).

The data for events that are selected by the LVL1 trigger are transferred into a 'derandomizer' — a memory that can accept the high instantaneous input rate (i.e., one word per 25 ns) while being read out

[5]Tevatron-II refers to the Tevatron collider after the luminosity upgrade.

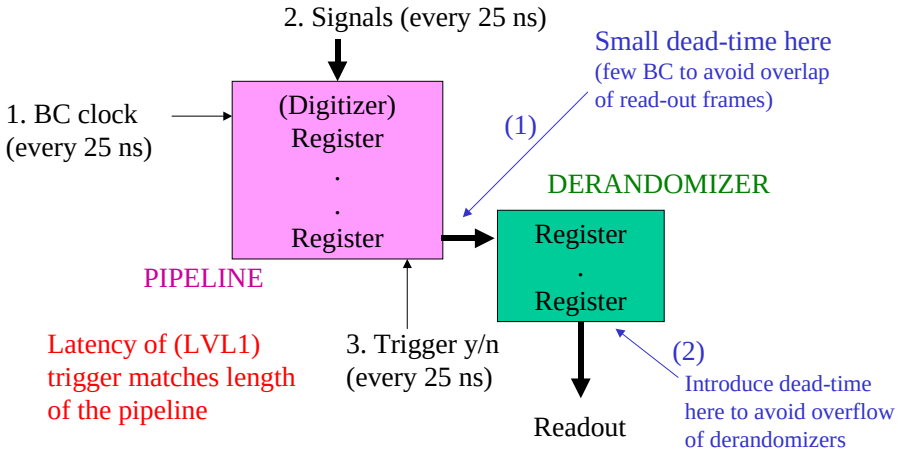


Fig. 11: Pipelined readout with derandomizer at the LHC

at the much lower average data rate (determined by the LVL1 trigger rate rather than the BC rate). In principle no dead-time needs to be introduced in such a system. However, in practice, data are retained for a few BCs around the one that gave rise to the trigger, and a dead period of a few BCs is introduced to ensure that the same data do not have to be accessed for more than one trigger. Dead-time must also be introduced to prevent the derandomizers from overflowing, e.g., where, due to a statistical fluctuation, many LVL1 triggers arrive in quick succession. The dead-time from the first of these sources can be estimated as follows (numbers from ATLAS): taking a LVL1 trigger rate of 75 kHz and 4 dead BCs following each LVL1 trigger gives $75\text{ kHz} \times 4 \times 25\text{ ns} = 0.75\%$. The dead-time from the second source depends on the size of the derandomizer and the speed with which it can be emptied — in ATLAS the requirements are $< 1\%$ dead-time for a LVL1 rate of 75 kHz ($< 6\%$ for 100 kHz).

Some of the elements of the readout chain in the LHC experiments have to be mounted on the detectors (and hence are totally inaccessible during running of the machine and are in an environment with high radiation levels). This is shown for the case of CMS in Fig. 12.

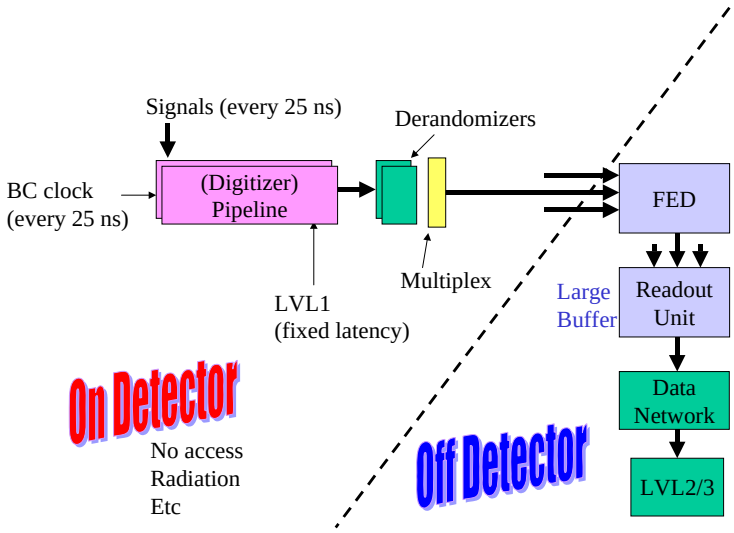


Fig. 12: Location of readout components in CMS

There are a variety of options for the placement of digitization in the readout chain, and the optimum choice depends on the characteristics of the detector in question. Digitization may be performed on

the detector at 40 MHz rate, prior to a digital pipeline (e.g., CMS calorimeter). Alternatively, it may be done on the detector after multiplexing signals from several analog pipelines (e.g., ATLAS EM calorimeter) — here the digitization rate can be lower, given by the LVL1 trigger rate multiplied by the number of signals to be digitized per trigger. Another alternative (e.g., CMS tracker) is to multiplex analog signals from the pipelines over analog links, and then to perform the digitization off-detector.

5.2 Pipelined LVL1 trigger

As discussed above, the LVL1 trigger has to deliver a new decision every BC, although the trigger latency is much longer than the BC period; the LVL1 trigger must concurrently process many events. This can be achieved by 'pipelining' the processing in custom trigger processors built using modern digital electronics. The key ingredients in this approach are to break the processing down into a series of steps, each of which can be performed within a single BC period, and to perform many operations in parallel by having separate processing logic for each calculation. Note that in such a system the latency of the LVL1 trigger is fixed — it is determined by the number of steps in the calculation, plus the time taken to move signals and data to, from and between the components of the trigger system (e.g., propagation delays on cables).

Pipelined trigger processing is illustrated in Fig. 13 — as will be seen later, this example corresponds to a (very small) part of the ATLAS LVL1 calorimeter trigger processor. The drawing on the left of Fig. 13 depicts the EM calorimeter as a grid of 'towers' in η–ϕ space (η is pseudorapidity, ϕ is azimuth angle). The logic shown on the right determines if the energy deposited in a horizontal or vertical pair of towers in the region [A, B, C] exceeds a threshold. In each 25 ns period, data from one layer of 'latches' (memory registers) are processed through the next step in the processing 'pipe', and the results are captured in the next layer of latches. Note that, in the real system, such logic has to be performed in parallel for ~3500 positions of the reference tower; the tower 'A' could be at any position in the calorimeter. In practice, modern electronics is capable of doing more than a simple add or compare operation in 25 ns, so there is more logic between the latches than in this illustration.

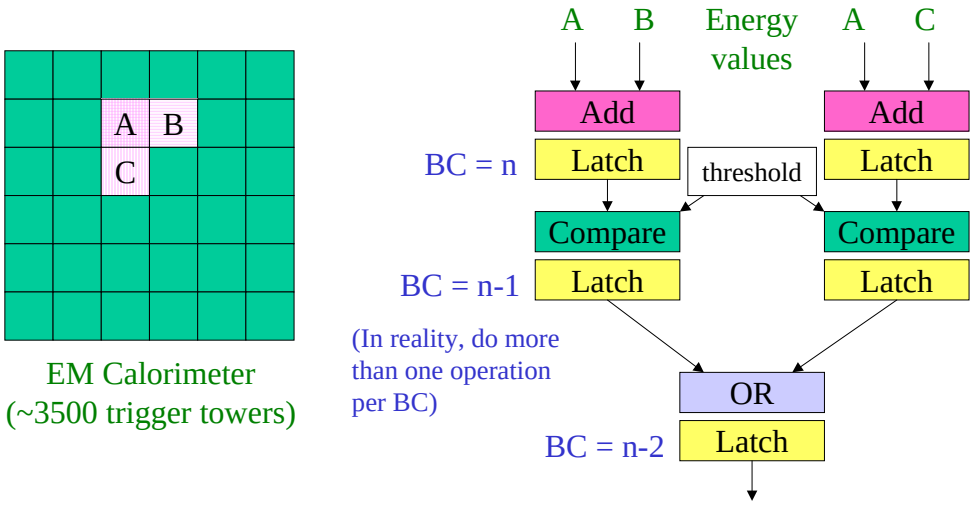


Fig. 13: Illustration of pipelined processing

The amount of data to be handled varies with depth in the processing pipeline, as indicated in Fig. 14. Initially the amount of data expands compared to the raw digitization level since each datum typically participates in several operations — the input data need to be 'fanned out' to several processing elements. Subsequently the amount of data decreases as one moves further down the processing tree. The final trigger decision can be represented by a single bit of information for each BC — yes or no (binary 1 or 0). Note that, in addition to the trigger decision, the LVL1 processors produce a lot of data

for use in monitoring the system and to guide the higher levels of selection.

Although they have not been discussed in these lectures because of time limitations, some fixed-target experiments have very challenging T/DAQ requirements. Some examples can be found in Refs. [17, 18].

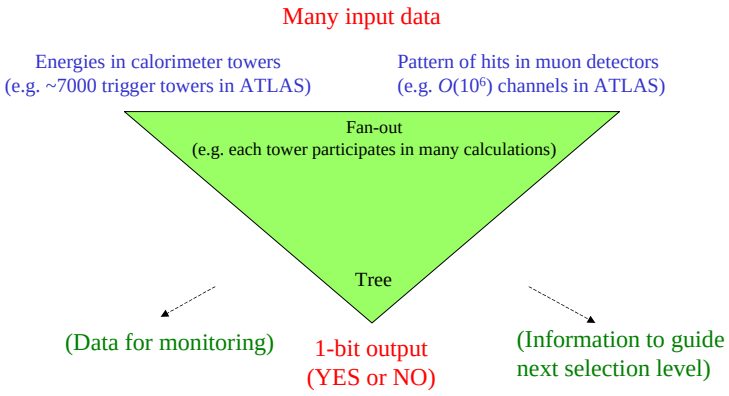


Fig. 14: LVL1 data flow

6 High-level triggers and data acquisition at the LHC

In the LHC experiments, data are transferred after a LVL1 trigger accept decision to large buffer memories — in normal operation the subsequent stages should not introduce further dead-time. At this point in the readout chain, the data rates are still massive. An event size of ~ 1 Mbyte (after zero suppression or data compression) at ~ 100 kHz event rate gives a total bandwidth of ~ 100 Gbytes/s (i.e., ~ 800 Gbits/s). This is far beyond the capacity of the bus-based event building of LEP. Such high data rates will be dealt with by using network-based event building and by only moving a subset of the data.

Network-based event building is illustrated in Fig. 15 for the example of CMS. Data are stored in the readout systems until they have been transferred to the filter systems [associated with high-level trigger (HLT) processing], or until the event is rejected. Note that no node in the system sees the full data rate — each readout system covers only a part of the detector and each filter system deals with only a fraction of the events.

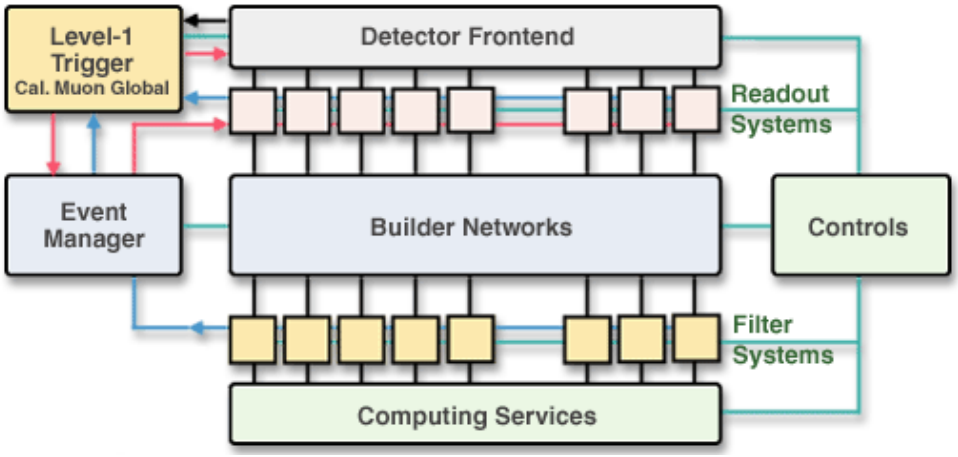


Fig. 15: CMS event builder

The LVL2 trigger decision can be made without accessing or processing all of the data. Substantial rejection can be made with respect to LVL1 without accessing the inner-tracking detectors — calorimeter

triggers can be refined using the full-precision, full-granularity calorimeter information; muon triggers can be refined using the high-precision readout from the muon detectors. It is therefore only necessary to access the inner-tracking data for the subset of events that pass this initial selection. ATLAS and CMS both use this sequential selection strategy. Nevertheless, the massive data rates pose problems even for network-based event building, and different solutions have been adopted in ATLAS and CMS to address this.

In CMS the event building is factorized into a number of 'slices', each of which sees only a fraction of the total rate (see Fig. 16). This still requires a large total network bandwidth (which has implications for the cost), but it avoids the need for a very big central network switch. An additional advantage of this approach is that the size of the system can be scaled, starting with a few slices and adding more later (e.g., as additional funding becomes available).

Eight slices:
Each slice sees
only 1/8th of
the events

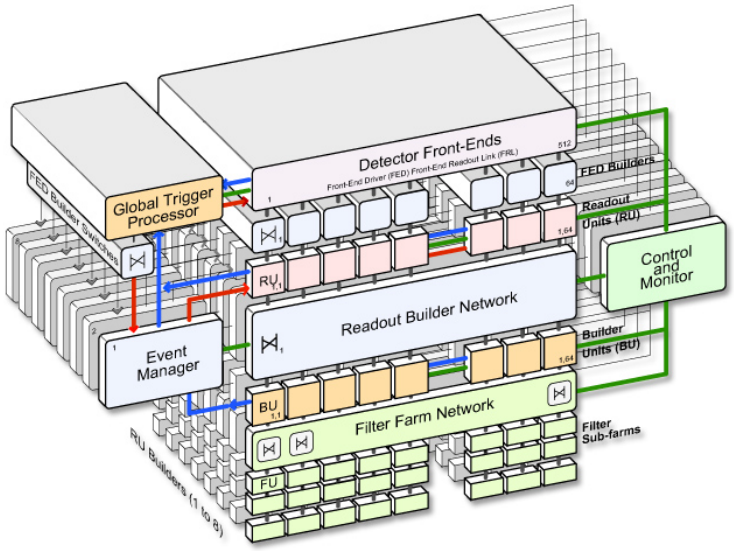


Fig. 16: The CMS slicing concept

In ATLAS the amount of data to be moved is reduced by using the region-of-interest (RoI) mechanism (see Fig. 17). Here, the LVL1 trigger indicates the geographical location in the detector of candidate objects. LVL2 then only needs to access data from the RoIs, a small fraction of the total, even for the calorimeter and muon detectors that participated in the LVL1 selection. This requires relatively complicated mechanisms to serve the data selectively to the LVL2 trigger processors.

In the example shown in Fig. 17, two muons are identified by LVL1. It can be seen that only a small fraction of the detector has to be accessed to validate the muons. In a first step only the data from the muon detectors are accessed and processed, and many events will be rejected where the more detailed analysis does not confirm the comparatively crude LVL1 selection (e.g., sharper p_{T} cut). For those events that remain, the inner-tracker data will be accessed within the RoIs, allowing further rejection (e.g., of muons from decays in flight of charged pions and kaons). In a last step, calorimeter information may be accessed within the RoIs to select isolated muons (e.g., to reduce the high rate of events with muons from bottom and charm decays, while retaining those from W and Z decays).

Concerning hardware implementation, the computer industry is putting on the market technologies that can be used to build much of the HLT/DAQ systems at the LHC. Computer network products now offer high performance at affordable cost. Personal computers (PCs) provide exceptional value for money in processing power, with high-speed network interfaces as standard items. Nevertheless, custom hardware is needed in the parts of the system that see the full LVL1 trigger output rate (~ 100 kHz). This concerns the readout systems that receive the detector data following a positive LVL1 trigger decision,

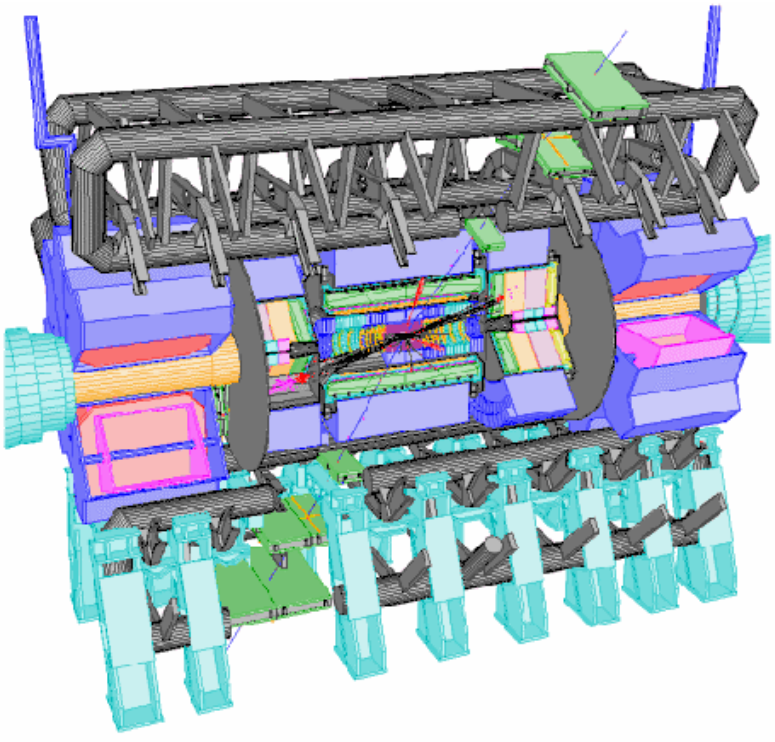


Fig. 17: The ATLAS region-of-interest concept — example of a dimuon event (see text)

and (in ATLAS) the interface to the LVL1 trigger that receives the RoI pointers. Of course, this is in addition to the specialized front-end electronics associated with the detectors that was discussed earlier (digitization, pipelines, derandomizers, etc.).

7 Physics requirements — two examples

In the following, the physics requirements on the T/DAQ systems at LEP and at the LHC are examined. These are complementary cases — at LEP precision physics was the main emphasis, at the LHC discovery physics will be the main issue. Precision physics at LEP needed accurate determination of the absolute cross-section (e.g., in the determination of the number of light-neutrino species). Discovery physics at the LHC will require sensitivity to a huge range of predicted processes with diverse signatures (with very low signal rates expected in some cases), aiming to be as sensitive as possible to new physics that has not been predicted (by using inclusive signatures). This has to be achieved in the presence of an enormous rate of Standard Model physics backgrounds (the rate of proton–proton collisions at the LHC will be $\mathcal{O}(10^9)$ Hz — $\sigma \sim 100$ mb, $\mathscr{L} \sim 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$).

7.1 Physics requirements at LEP

Triggers at LEP aimed to identify all events coming from $\mathrm{e}^+\,\mathrm{e}^-$ annihilations with visible final states. At LEP-I, operating with $\sqrt{s} \sim m_{\mathrm{Z}}$, this included $\mathrm{Z} \to$ hadrons, $\mathrm{Z} \to \mathrm{e}^+\mathrm{e}^-$, $\mathrm{Z} \to \mu^+\mu^-$, and $\mathrm{Z} \to \tau^+\tau^-$; at LEP-II, operating above the WW threshold, this included WW, ZZ and single-boson events. Sensitivity was required even in cases where there was little visible energy, e.g., in the Standard Model for $\mathrm{e}^+\mathrm{e}^- \to \mathrm{Z}\gamma$, with $\mathrm{Z} \to \nu\nu$, and in new-particle searches such as $\mathrm{e}^+\mathrm{e}^- \to \tilde{\chi}^+\tilde{\chi}^-$ for the case of small $\tilde{\chi}^\pm - \tilde{\chi}^0$ mass difference that gives only low-energy visible particles ($\tilde{\chi}^0$ is the lightest supersymmetric particle). In addition, the triggers had to retain some fraction of two-photon collision events (used for QCD studies), and identify Bhabha scatters (needed for precise luminosity determination).

The triggers could retain events with any significant activity in the detector. Even when running at

the Z peak, the rate of Z decays was only $\mathcal{O}(1)$ Hz — physics rate was not an issue. The challenge was in maximizing the efficiency (and acceptance) of the trigger, and making sure that the small inefficiencies were very well understood. The determination of absolute cross-section depends on knowing the integrated luminosity and the experimental efficiency to select the process in question (i.e., the efficiency to trigger on the specific physics process). Precise determination of the integrated luminosity required excellent understanding of the trigger efficiency for Bhabha-scattering events (luminosity determined from the rate of Bhabha scatters within a given angular range). A major achievement at LEP was to reach 'per mil' precision.

The trigger rates (events per second) and the DAQ rates (bytes per second) at LEP were modest as discussed in Section 4.

7.2 Physics requirements at the LHC

Triggers in the general-purpose proton–proton experiments at the LHC (ATLAS [19, 20] and CMS [21, 22]) will have to retain as high as possible a fraction of the events of interest for the diverse physics programmes of these experiments. Higgs searches in and beyond the Standard Model will include looking for $\mathrm{H} \rightarrow \mathrm{ZZ} \rightarrow$ leptons and also $\mathrm{H} \rightarrow \mathrm{b\overline{b}}$. Supersymmetry (SUSY) searches will be performed with and without the assumption of R-parity conservation. One will search for other new physics using inclusive triggers that one hopes will be sensitive to unpredicted processes. In parallel with the searches for new physics, the LHC experiments aim to do precision physics, such as measuring the W mass and some B-physics studies, especially in the early phases of LHC running when the luminosity is expected to be comparatively low.

In contrast to the experiments at LEP, the LHC trigger systems have a hard job to reduce the physics event rate to a manageable level for data recording and offline analysis. As discussed above, the design luminosity $\mathscr{L} \sim 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, together with $\sigma \sim 100$ mb, implies an $\mathcal{O}(10^9)$ Hz interaction rate. Even the rate of events containing leptonic decays of W and Z bosons is $\mathcal{O}(100)$ Hz. Furthermore, the size of the events is very large, $\mathcal{O}(1)$ Mbyte, reflecting the huge number of detector channels and the high particle multiplicity in each event. Recording and subsequently processing offline $\mathcal{O}(100)$ Hz event rate per experiment with an $\mathcal{O}(1)$ Mbyte event size is considered feasible, but it implies major computing resources [23]. Hence, only a tiny fraction of proton–proton collisions can be selected — taking the order-of-magnitude numbers given above, the maximum fraction of interactions that can be selected is $\mathcal{O}(10^{-7})$. Note that the general-purpose LHC experiments have to balance the needs of maximizing physics coverage and reaching acceptable (i.e., affordable) recording rates.

The LHCb experiment [24], which is dedicated to studying B-physics, faces similar challenges to ATLAS and CMS. It will operate at a comparatively low luminosity ($\mathscr{L} \sim 10^{32}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$), giving an overall proton–proton interaction rate of ~ 20 MHz — chosen to maximize the rate of single-interaction bunch crossings. The event size will be comparatively small (~ 100 kbytes) as a result of having fewer detector channels and of the lower occupancy of the detector (due to the lower luminosity with less pile-up). However, there will be a very high rate of beauty production in LHCb — taking $\sigma \sim$ 500 μb, the production rate will be ~ 100 kHz — and the trigger must search for specific B-decay modes that are of interest for physics analysis, with the aim of recording an event rate of only ~ 200 Hz.

The heavy-ion experiment ALICE [5] is also very demanding, particularly from the DAQ point of view. The total interaction rate will be much smaller than in the proton–proton experiments — $\mathscr{L} \sim 10^{27}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ is predicted to give a rate ~ 8000 Hz for Pb–Pb collisions. However, the event size will be huge due to the high final-state multiplicity in Pb–Pb interactions at LHC energy. Up to $\mathcal{O}(10^4)$ charged particles will be produced in the central region, giving an event size of up to ~ 40 Mbytes when the full detector is read out. The ALICE trigger will select 'minimum-bias' and 'central' events (rates scaled down to a total of about 40 Hz), and events with dileptons (~ 1 kHz with only part of the detector read out). Even compared to the other LHC experiments, the volume of data to be stored and subsequently processed offline will be massive, with a data rate to storage of ~ 1 Gbytes/s (considered to be

about the maximum affordable rate).

8 Signatures of different types of particle

The generic signatures for different types of particle are illustrated in Fig. 18. Moving away from the interaction point (shown as a star on the left-hand side of Fig. 18), one finds the inner tracking detector (IDET), the electromagnetic calorimeter (ECAL), the hadronic calorimeter (HCAL) and the muon detectors (MuDET). Charged particles (electrons, muons and charged hadrons) leave tracks in the IDET. Electrons and photons shower in the ECAL, giving localized clusters of energy without activity in the HCAL. Hadrons produce larger showers that may start in the ECAL but extend into the HCAL. Muons traverse the calorimeters with minimal energy loss and are detected in the MuDET.

The momenta of charged particles are measured from the radii of curvature of their tracks in the IDET which is embedded in a magnetic field. A further measurement of the momenta of muons may be made in the MuDET using a second magnet system. The energies of electrons, photons and hadrons are measured in the calorimeters. Although neutrinos leave the detector system without interaction, one can infer their presence from the momentum imbalance in the event (sometimes referred to as 'missing energy'). Hadronic jets contain a mixture of particles, including neutral pions that decay almost immediately into photon pairs that are then detected in the ECAL. The jets appear as broad clusters of energy in the calorimeters where the individual particles will sometimes not be resolved.

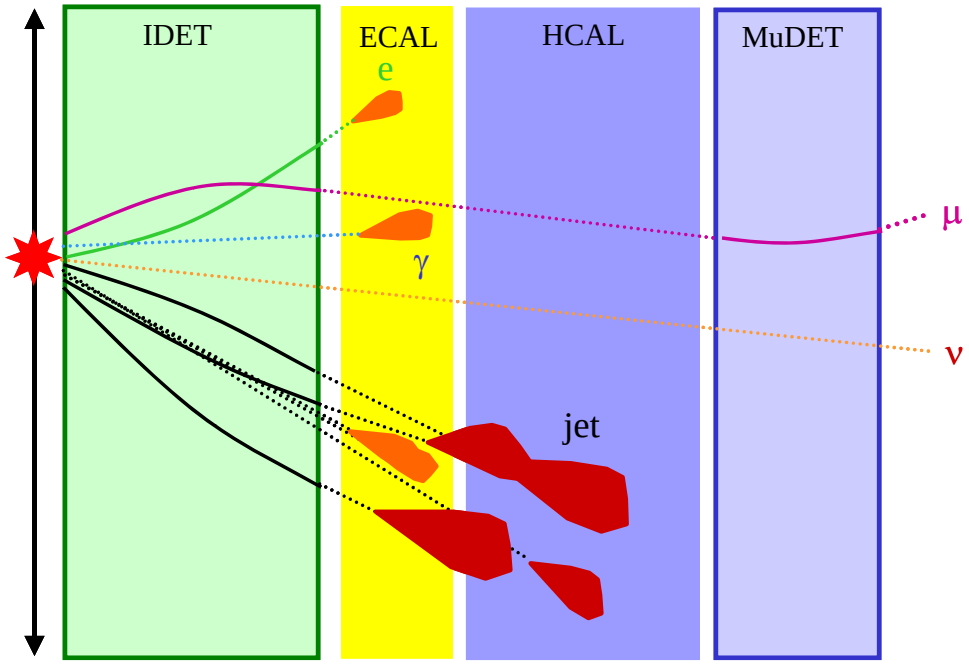


Fig. 18: Signatures of different types of particle in a generic detector

9 Selection criteria and trigger implementations at LEP

The details of the selection criteria and trigger implementations at LEP varied from experiment to experiment [8–11]. Discussion of the example of ALEPH is continued with the aim of giving a reasonably in-depth view of one system. For triggering purposes, the detector was divided into segments with a total of 60 regions in θ, ϕ (θ is polar angle and ϕ is azimuth with respect to the beam axis). Within these segments, the following trigger objects were identified:

1. muon — requiring a track penetrating the hadron calorimeter and seen in the inner tracker;

2. charged electromagnetic (EM) energy — requiring an EM calorimeter cluster and a track in the inner tracker;
3. neutral EM energy — requiring an EM calorimeter cluster (with higher thresholds than in (2) to limit the rate to acceptable levels).

In addition to the above local triggers, there were total-energy triggers (applying thresholds on energies summed over large regions — the barrel or a full endcap), a back-to-back tracks trigger, and triggers for Bhabha scattering (luminosity monitor).

The LVL1 triggers were implemented using a combination of analog and digital electronics. The calorimeter triggers were implemented using analog electronics to sum signals before applying thresholds on the sums. The LVL1 tracking trigger looked for patterns of hits in the inner-tracking chamber (ITC) consistent with a track with $p_{\mathrm{T}} > 1$ GeV [6] — at LVL2 the Time Projection Chamber (TPC) was used instead. The final decision was made by combining digital information from calorimeter and tracking triggers, making local combinations within segments of the detector, and then making a global combination (logical OR of conditions).

10 Selection criteria at LHC

Features that distinguish new physics from the bulk of the cross-section for Standard Model processes at hadron colliders are generally the presence of high-p_{T} particles (or jets). For example, these may be the products of the decays of new heavy particles. In contrast, most of the particles produced in minimum-bias interactions are soft ($p_{\mathrm{T}} \sim 1$ GeV or less). More specific signatures are the presence of high-p_{T} leptons (e, μ, τ), photons and/or neutrinos. For example, these may be the products (directly or indirectly) of new heavy particles. Charged leptons, photons and neutrinos give a particularly clean signature (c.f. low-p_{T} hadrons in minimum-bias events), especially if they are 'isolated' (i.e., not inside jets). The presence of heavy particles such as W and Z bosons can be another signature for new physics — e.g., they may be produced in Higgs decays. Leptonic W and Z decays give a very clean signature that can be used in the trigger. Of course it is interesting to study W and Z boson production $perse$, and such events can be very useful for detector studies (e.g., calibration of the EM calorimeters).

In view of the above, LVL1 triggers at hadron colliders search for the following signatures (see Fig. 18).

- High-p_{T} muons — these can be identified as charged particles that penetrate beyond the calorimeters; a p_{T} cut is needed to control the rate of muons from $\pi^{\pm} \to \mu^{\pm}\nu$ and $\mathrm{K}^{\pm} \to \mu^{\pm}\nu$ decays in flight, as well as those from semi-muonic beauty and charm decays.
- High-p_{T} photons — these can be identified as narrow clusters in the EM calorimeter; cuts are made on transverse energy ($E_{\mathrm{T}} >$ threshold), and isolation and associated hadronic transverse energy ($E_{\mathrm{T}} <$ threshold), to reduce the rate due to misidentified high-p_{T} jets.
- High-p_{T} electrons — identified in a similar way to photons, although some experiments require a matching track as early as LVL1.
- High-p_{T} taus — identified as narrow clusters in the calorimeters (EM and hadronic energy combined).
- High-p_{T} jets — identified as wider clusters in the calorimeters (EM and hadronic energy combined); note that one needs to cut at very high p_{T} to get acceptable rates given that jets are the dominant high-p_{T} process.
- Large missing E_{T} or scalar E_{T}.

Some experiments also search for tracks from displaced secondary vertices at an early stage in the trigger selection.

[6] Here, p_{T} is transverse momentum (measured with respect to the beam axis); similarly, E_{T} is transverse energy.

The trigger selection criteria are typically expressed as a list of conditions that should be satisfied — if any of the conditions is met, a trigger is generated (subject to dead-time requirements, etc.). In these notes, the list of conditions is referred to as the 'trigger menu', although the name varies from experiment to experiment. An illustrative example of a LVL1 trigger menu for high-luminosity running at LHC includes the following (rates [19] are given for the case of ATLAS at $\mathscr{L} \sim 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$):

- one or more muons with $p_{\mathrm{T}} > 20$ GeV (rate ~ 11 kHz);
- two or more muons each with $p_{\mathrm{T}} > 6$ GeV (rate ~ 1 kHz);
- one or more e/γ with $E_{\mathrm{T}} > 30$ GeV (rate ~ 22 kHz);
- two or more e/γ each with $E_{\mathrm{T}} > 20$ GeV (rate ~ 5 kHz);
- one or more jets with $E_{\mathrm{T}} > 290$ GeV (rate ~ 200 Hz);
- one or more jets with $E_{\mathrm{T}} > 100$ GeV and missing-$E_{\mathrm{T}} > 100$ GeV (rate ~ 500 Hz);
- three or more jets with $E_{\mathrm{T}} > 130$ GeV (rate ~ 200 Hz);
- four or more jets with $E_{\mathrm{T}} > 90$ GeV (rate ~ 200 Hz).

The above list represents an extract from a LVL1 trigger menu, indicating some of the most important trigger requirements — the full menu would include many items in addition (typically more than 100 items in total). The additional items are expected to include the following:

- τ (or isolated single-hadron) candidates;
- combinations of objects of different types (e.g., muon *and* e/γ);
- pre-scaled[7] triggers with lower thresholds;
- triggers needed for technical studies and to aid understanding of the data from the main triggers (e.g., trigger on bunch crossings at random to collect an unbiased data sample).

As for the LVL1 trigger, the HLT has a trigger menu that describes which events should be selected. This is illustrated in Table 2 for the example of CMS, assuming a luminosity for early running of $\mathscr{L} \sim 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. The total rate of ~ 100 Hz contains a large fraction of events that are useful for physics analysis. Lower thresholds would be desirable, but the physics coverage has to be balanced against considerations of the offline computing cost. Note that there are large uncertainties on the rate calculations.

Table 2: Estimated high-level trigger rates for $\mathscr{L} \sim 2 \times 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ (CMS numbers from Ref. [21])

Trigger configuration	**Rate**
One or more electrons with $p_{\mathrm{T}} > 29$ GeV, or two or more electrons with $p_{\mathrm{T}} > 17$ GeV	~ 34 Hz
One or more photons with $p_{\mathrm{T}} > 80$ GeV, or two or more photons with $p_{\mathrm{T}} > 40, 25$ GeV	~ 9 Hz
One or more muons with $p_{\mathrm{T}} > 19$ GeV, or two or more muons with $p_{\mathrm{T}} > 7$ GeV	~ 29 Hz
One or more taus with $p_{\mathrm{T}} > 86$ GeV, or two or more taus with $p_{\mathrm{T}} > 59$ GeV	~ 4 Hz
One or more jets with $p_{\mathrm{T}} > 180$ GeV *and* missing-$E_{\mathrm{T}} > 123$ GeV	~ 5 Hz
One or more jets with $p_{\mathrm{T}} > 657$ GeV, or three or more jets with $p_{\mathrm{T}} > 247$ GeV, or four or more jets with $p_{\mathrm{T}} > 113$ GeV	~ 9 Hz
Others (electron and jet, b-jets, etc.)	~ 7 Hz

[7] Some triggers may be 'pre-scaled' — this means that only every N^{th} event satisfying the relevant criteria is recorded, where N is a parameter called the pre-scale factor; this is useful for collecting samples of high-rate triggers without swamping the T/DAQ system.

A major challenge lies in the HLT/DAQ software. The event-selection algorithms for the HLT can be subdivided, at least logically, into LVL2 and LVL3 trigger stages. These might be performed by two separate processor systems (e.g., ATLAS), or in two distinct processing steps within the same processor system (e.g., CMS). The algorithms have to be supported by a software framework that manages the flow of data, supervising an event from when it arrives at the HLT/DAQ system until it is either rejected, or accepted and recorded on permanent storage. This includes software for efficient transfer of data to the algorithms. In addition to the above, there is a large amount of associated online software (run control, databases, book-keeping, etc.).

11 LVL1 trigger design for the LHC

A number of design goals must be kept in mind for the LVL1 triggers at the LHC. It is essential to achieve a very large reduction in the physics rate, otherwise the HLT/DAQ system will be swamped and the dead-time will become unacceptable. In practice, the interaction rate, $\mathcal{O}(10^9)$ Hz, must be reduced to less than 100 kHz in ATLAS and CMS. Complex algorithms are needed to reject the background while keeping the signal events.

Another important constraint is to achieve a short latency — information from all detector elements ($\mathcal{O}(10^7$–$10^8)$ channels!) has to be held on the detector pending the LVL1 decision. The pipeline memories that do this are typically implemented in ASICs (application-specific integrated circuits), and memory size contributes to the cost. Typical LVL1 latency values are a few microseconds (e.g., less than 2.5 μs in ATLAS and less than 3.2 μs in CMS).

A third requirement is to have flexibility to react to changing conditions (e.g., a wide range of luminosities) and — it is hoped — to new physics! The algorithms must be programmable, at least at the level of parameters (thresholds, etc.).

11.1 Case study — ATLAS e/γ trigger

The ATLAS e/γ trigger algorithm can be used to illustrate the techniques used in LVL1 trigger systems at LHC. It is based on 4×4 'overlapping, sliding windows' of trigger towers as illustrated in Fig. 19. Each trigger tower has a lateral extent of 0.1×0.1 in η, ϕ space, where η is pseudorapidity and ϕ is azimuth. There are about 3500 such towers in each of the EM and hadronic calorimeters. Note that each tower participates in calculations for 16 windows. The algorithm requires a local maximum in the EM calorimeter to define the η–ϕ position of the cluster and to avoid double counting of extended clusters (so-called 'declustering'). It can also require that the cluster be isolated, i.e., little energy surrounding the cluster in the EM calorimeter or the hadronic calorimeter.

The implementation of the ATLAS LVL1 calorimeter trigger [25] is sketched in Fig. 20. Analog electronics on the detector sums signals from individual calorimeter cells to form trigger-tower signals. After transmission to the 'pre-processor' (PPr), which is located in an underground room close to the detector and shielded against radiation, the tower signals are received and digitized; then the digital data are processed to obtain estimates of E_{T} per trigger tower for each BC. At this point in the processing chain (i.e., at the output of the PPr), there is an 'η–ϕ matrix' of the E_{T} per tower in each of the EM and hadronic calorimeters that gets updated every 25 ns.

The tower data from the PPr are transmitted to the cluster processor (CP). Note that the CP is implemented with very dense electronics so that there are only four crates in total. This minimizes the number of towers that need to be transmitted ('fanned out') to more than one crate. Fan out is required for towers that contribute to windows for which the algorithmic processing is implemented in more than one crate. Also, within each CP crate, trigger-tower data need to be fanned out between electronic modules, and then between processing elements within each module. Considerations of connectivity and data-movement drive the design.

In parallel with the CP, a jet/energy processor (JEP) searches for jet candidates and calculates

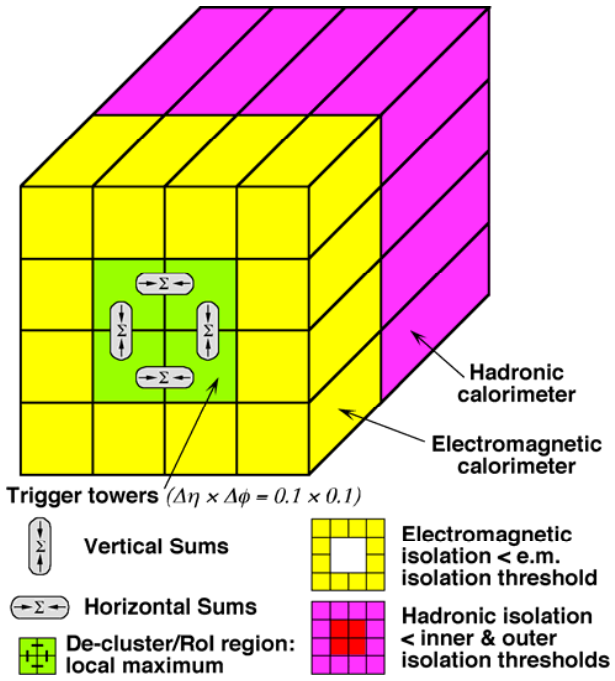


Fig. 19: ATLAS e/γ trigger algorithm

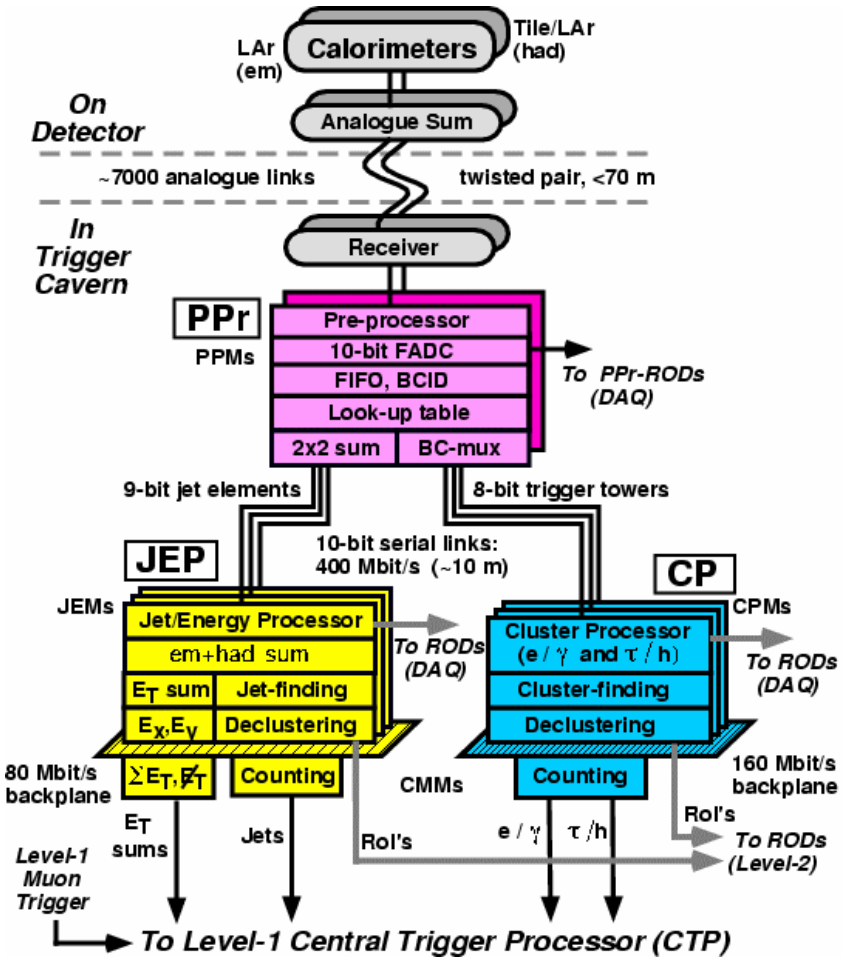


Fig. 20: Overview of the ATLAS LVL1 calorimeter trigger

missing-E_{T} and scalar-E_{T} sums. This is not described further here.

A very important consideration in designing the LVL1 trigger is the need to identify uniquely the BC that produced the interaction of interest. This is not trivial, especially given that the calorimeter signals extend over many BCs. In order to assign observed energy deposits to a given BC, information has to be combined from a sequence of measurements. Figure 21 illustrates how this is done within the PPr (the logic is repeated ~7000 times so that this is done in parallel for all towers). The raw data for a given tower move along a pipeline that is clocked by the 40 MHz BC signal. The multipliers together with the adder tree implement a finite-impulse-response filter whose output is passed to a peak finder (a peak indicates that the energy was deposited in the BC currently being examined) and to a look-up table that converts the peak amplitude to an E_{T} value. Special care is taken to avoid BC misidentification for very large pulses that may get distorted in the analog electronics, since such signals could correspond to the most interesting events. The functionality shown in Fig. 21 is implemented in ASICs (four channels per ASIC).

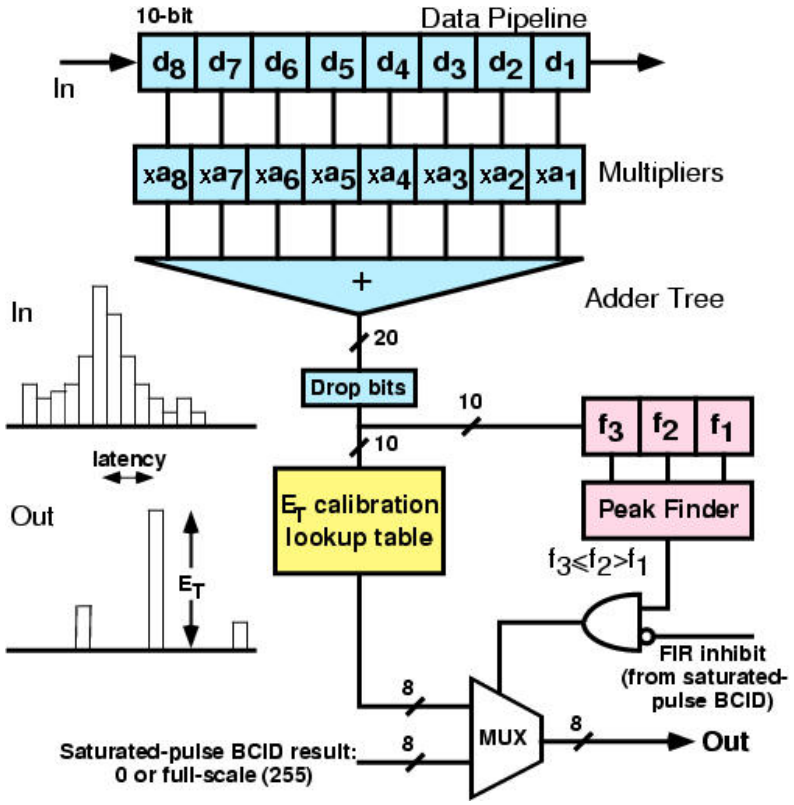


Fig. 21: Bunch-crossing identification

The transmission of the data (i.e., the E_{T} matrices) from the PPr to the CP is performed using a total of 5000 digital links each operating at 400 Mbits/s (each link carries data from two towers using a technique called BC multiplexing [25]). Where fan out is required, the corresponding links are duplicated with the data being sent to two different CP crates. Within each CP crate, data are shared between neighbouring modules over a very high density crate back-plane (~800 pins per slot in a 9U crate; data rate of 160 Mbits/s per signal pin using point-to-point connections). On each of the modules, data are passed to eight large field-programmable gate arrays (FPGAs) that perform the algorithmic processing, fanning out signals to more than one FPGA where required.

As an exercise, it is suggested that students make an order-of-magnitude estimate of the total bandwidth between the PPr and the CP, considering what this corresponds to in terms of an equivalent

number of simultaneous telephone calls[8].

The e/γ (together with the τ/h) algorithms are implemented using FPGAs. This has only become feasible thanks to recent advances in FPGA technology since very large and very fast devices are needed. Each FPGA handles an area of 4×2 windows, requiring data from 7×5 towers in each of the EM and hadronic calorimeters. The algorithm is described in a programming language (e.g., VHDL) that can be converted into the FPGA configuration file. This gives flexibility to adapt algorithms in the light of experience — the FPGAs can be reconfigured *in situ*. Note that parameters of the algorithms can be changed easily and quickly, e.g., as the luminosity falls during the course of a coast of the beams in the LHC machine, since they are held in registers inside the FPGAs that can be modified at run time (i.e., there is no need to change the 'program' in the FPGA).

12 High-level trigger algorithms

There was not time in the lectures for a detailed discussion of the algorithms that are used in the HLT. However, it is useful to consider the case of the electron selection that follows after the first-level trigger. The LVL1 e/γ trigger is already very selective, so it is necessary to use complex algorithms and full-granularity, full-precision detector data in the HLT.

A calorimeter selection is made applying a sharper E_{T} cut (better resolution than at LVL1) and shower-shape variables that distinguish between the electromagnetic showers of an electron or photon on one hand, and activity from jets on the other hand. The shower-shape variables use both lateral and depth profile information. Then, for electrons, a requirement is made of an associated track in the inner detector, matching the calorimeter cluster in space, and with consistent momentum and energy measurements from the inner detector and calorimeter respectively.

Much work is going on to develop the algorithms and tune their many parameters to optimize their signal efficiency and background rejection. So far this has been done with simulated data, but further optimization will be required once samples of electrons are available from offline reconstruction of real data. It is worth noting that the efficiency value depends on the signal definition as shown in Fig. 22, an example of a study taken from Ref. [26]. Here the trigger efficiency is shown, as a function of electron transverse energy, relative to three different offline selections. With a loose offline selection, the trigger is comparatively inefficient, whereas it performs much better relative to the tighter offline cuts. This is related to the optimization of the trigger both for signal efficiency (where loose cuts are preferable) and for background rejection (where tighter cuts are required).

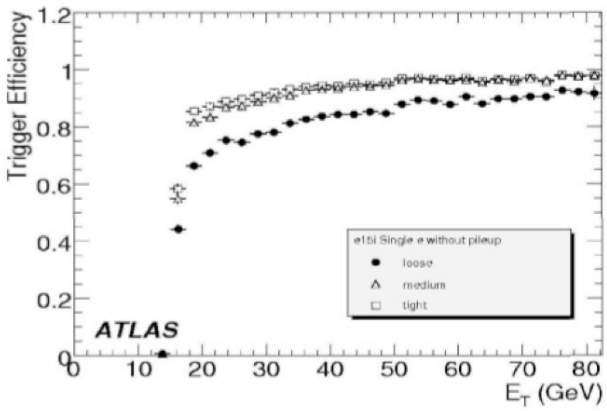


Fig. 22: Trigger efficiency versus electron E_{T} for three different offline selections of the reference sample

[8] One may assume an order-of-magnitude data rate for voice calls of 10 kbits/s — for example, the GSM mobile-phone standard uses a 9600 bit/s digital link to transmit the encoded voice signal.

13 Commissioning of the T/DAQ systems at LHC

Much more detail on the general commissioning of the LHC experiments can be found in the lectures of Andreas Hoecker at this School [27]. Here an attempt is made to describe how commissioning of the T/DAQ systems started in September 2008.

On 10 September 2008 the first beams passed around the LHC in both the clockwise and anti-clockwise directions, but with only one beam at a time (so there was no possibility of observing proton–proton collisions). The energy of the protons was 450 GeV which is the injection energy prior to acceleration; acceleration to higher energies was not attempted.

As a first step, the beams were brought around the machine and stopped on collimators such as those upstream of the ATLAS experiment. Given the huge number of protons per bunch, as well as the sizeable beam energy, extremely large numbers of secondary particles were produced, including muons that traversed the experiment depositing energy in all of the detector systems.

Next, the collimators were removed and the beams were allowed to circulate around the machine for a few turns and, after some tuning, for a few tens of turns. Subsequently, the beams were captured by the radio-frequency system of the LHC and circulated for periods of tens of minutes.

The first day of LHC operations was very exciting for all the people working on the experiments. There was a very large amount of media interest, with television broadcasts from various control rooms around the CERN site. It was a particularly challenging time for those working on the T/DAQ systems who were anxious to see if the first beam-related events would be identified and recorded successfully. Much to the relief of the author, the online event display of ATLAS soon showed a spectacular beam-splash event produced when the beam particles hit the collimator upstream of the experiment. The first ATLAS event is shown in Fig. 23; similar events were seen by the other experiments.

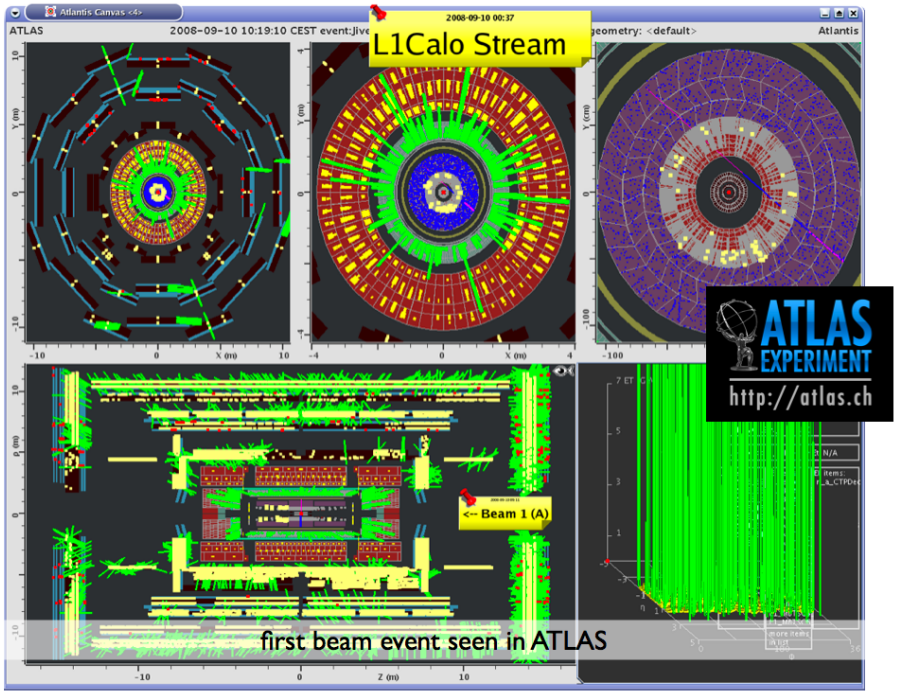


Fig. 23: The first beam-splash event in ATLAS

Analysis of the beam-spash events provided much useful information for commissioning the detectors and also the trigger. For example, the relative timing of different detector elements could be measured allowing the adjustment of programmable delays to the correct settings. The very large amount

of activity in the events had the advantage that signals were seen in an unusually large fraction of the detector channels.

An example of a very early study done with beam-splash events in shown in Fig. 24 which plots E_{T} versus η and ϕ for the ATLAS LVL1 calorimeter trigger readout. The E_{T} values are colour coded; η is along the x-axis and ϕ is along the y-axis. The eight-fold ϕ structure of the ATLAS magnets can be seen, as well as the effects of the tunnel floor and heavy mechanical support structures that reduced the flux of particles reaching the calorimeters in the bottom part of the detector ($\phi \approx 270$ degrees). The difference in absolute scale between the left-hand and right-hand sides of the plot is attributed to the fact that timing of the left-hand side was actually one bunch-crossing away from ideal when the data were collected; the timing calibration was subsequently adjusted as a result of these observations.

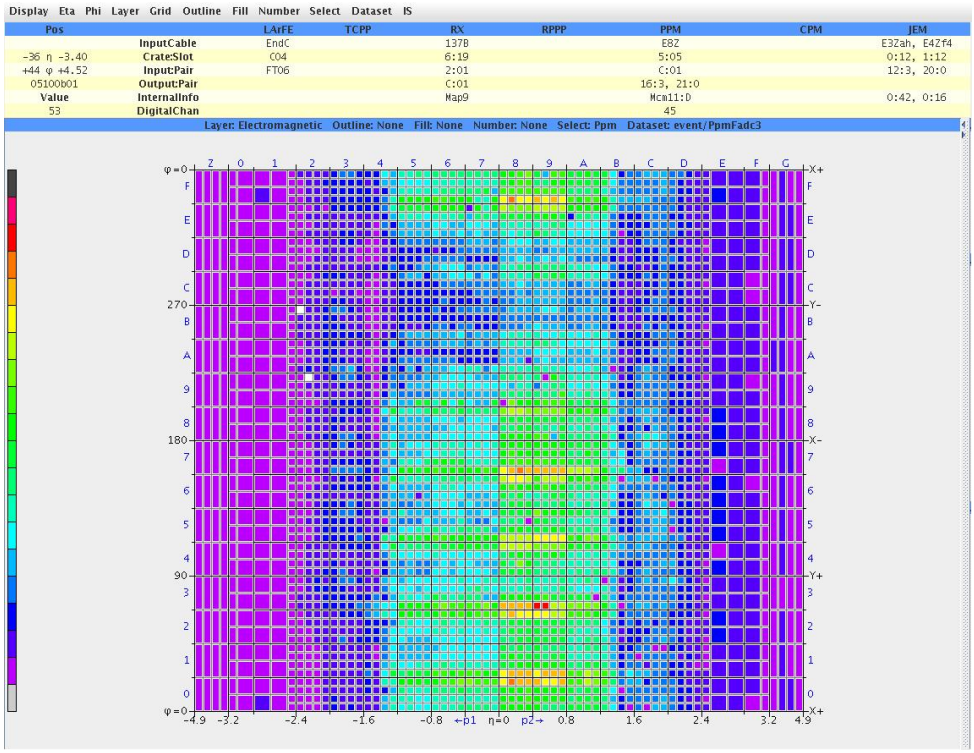


Fig. 24: LVL1 calorimeter trigger energy grid for a beam-splash event

At least in ATLAS, the first beam-spash events were recorded using triggers that had already been tested, with a free-runing 40 MHz clock, for cosmic-ray events. This approach was appropriate because of the importance of recording the first beam-related activity in the detector before the local beam instrumentation had been calibrated. However, it was crucial to move on as rapidly as possible to establish a precise and stable time reference.

Once beam-related activity had been seen in all of the LHC experiments, stopping the beam on the corresponding collimators, all of the collimators were removed and the beam was allowed to circulate. The first circulating beams passed around the LHC for only a short period of time, corresponding to a few turns initially, rising to a few tens of turns. For the 27 km LHC circumference, the orbit period is about 89 μs.

Upstream of the LHC detectors (and upstream of the collimators) are passive beam pick-ups that provide electrical signals induced by the passage of the proton beams. The photograph in the left-hand side of Fig. 25 shows the beam pick-up for one of the beams in an LHC experiment. Three of the four cables that carry the signals can be seen. The analog signals from electrodes above, below, to the left and to the right of the beam are combined (analog sum). The resulting signal is fed to an oscilloscope

directly and also via a discriminator (an electronic device that provides a logical output signal when the analog input signal exceeds a preset threshold, see Section 2).

On the right-hand side of Fig. 25 can be seen a plot, from CMS, of the relative timing of different signals. The upper three traces are 'orbit' signals provided by the LHC machine, whereas the bottom trace is the discriminated beam pick-up signal. As can be seen, the pick-up signal is present for only four turns and then disappears. The reason for this is that after a few turns the protons de-bunched and the analog signal from the pick-ups became too small to fire the discriminator. Similar instrumentation and timing calibration studies were used in all of the LHC experiments.

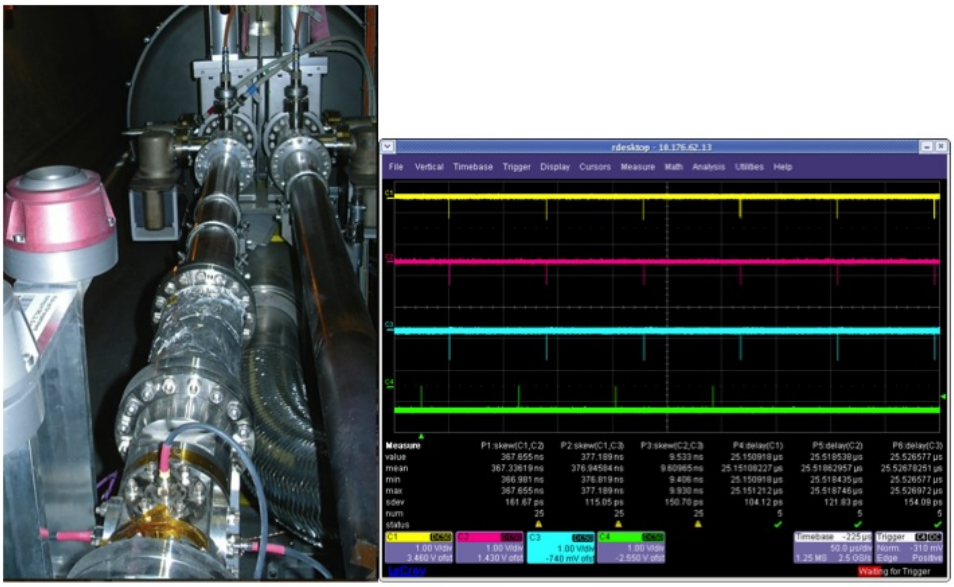


Fig. 25: Photograph of beam pick-up instrumentation (left) and display of timing signals recorded on a digital oscilloscope (right). The upper three traces are 'orbit' signals from the LHC machine, whereas the bottom one is the (inverted) discriminated signal from the beam pick-up.

A key feature of the beam pick-ups is that they provide a stable time reference with respect to which other signals can be aligned. The time of arrival of the beam pick-up signal, relative to the moment when the beam passes through the centre of the LHC detector, depends only on the proton time of flight from the beam pick-up position to the centre of the detector, propagation delays of the signal along the electrical cables, and the response time of the electronic circuits (which is very short).

Thanks to thorough preparations, the beam pick-up signals and their timing relative to the trigger could be measured as soon as beam was injected. Programmable delays could then be adjusted to align in time inputs to the trigger from the beam pick-ups and from other sources. For example, in ATLAS, the beam pick-up inputs were delayed so that they would have the same timing as other inputs that had already been adjusted using cosmic-rays.

Once the timing of the beam pick-up inputs to the trigger had been adjusted so as to initiate the detector readout for the appropriate bunch crossing (BC), i.e., to read out a time-frame that would contain the detector signals produced by beam-related activity, they could be used to provide the trigger for subsequent running.

It is worth noting that the steps described above to set up the timing of the trigger were completed

within just a few hours on the morning of 10 September 2008. From then onwards the beam pick-ups represented a stable time reference with respect to which other elements in the trigger and in the detector readout systems could be adjusted.

As already indicated, all of the beam operations in September 2008 were with just a single beam in the LHC. Operations were performed with beams circulating in both the clockwise and anti-clockwise directions. Beam activity in the detectors was produced by beam splash (beam stopped on collimators upstream of the detectors producing a massive number of secondary particles) or by beam-halo particles (produced when protons lost from the beam upstream of the detectors produced one or more high-momentum muons that traversed the detectors). In both cases one has to take into account the time of flight of the particles that reach one end of the detector before the other end. In contrast, beam–beam interactions have symmetric timing for the two ends of the detector.

The work on timing calibration performed over the days following the LHC start up can be illustrated by the case of ATLAS. Already on 10 September both sets of beam pick-ups had been commissioned (with beams circulating in the clockwise and anti-clockwise directions) giving a fixed time reference with respect to which the rest of the trigger, and indeed the rest of the experiment, could be aligned.

The situation on 10 September is summarized in the left-hand plot of Fig. 26. The beam pick-up signal, labelled 'BPTX' in the figure, is the reference. The relative time of arrival of other inputs to the trigger is shown in units of BC number (i.e. one unit corresponds to 25 ns which is the nominal bunch-crossing interval at LHC). Although there is a peak at the nominal timing (bunch-number zero) in the distributions based on different trigger inputs — the Minimum-Bias Trigger Scintillators (MBTS), the Thin-Gap Chamber (TGC) forward muon detectors, and the Tau5, J5 and EM3 items from the calorimeter trigger — the distribution is broad.

Prompt analysis and interpretation of the data allowed the timing to be understood and calibration corrections to be applied. Issues addressed included programming delay circuits to correct for time of flight of the particles according to the direction of the circulating beam and tuning the relative timing of triggers from different parts of the detector or from different detector channels.

The situation two days later on 12 September is summarized in the right-hand plot of Fig. 26. It is important to note that the scale is logarithmic — the vast majority of the triggers are aligned correctly in the nominal bunch crossing. Although shown in the plot, the input from the Resistive Plate Chambers (RPC) barrel muon detectors, which see very little beam-halo activity in single-beam operation, had not been timed-in.

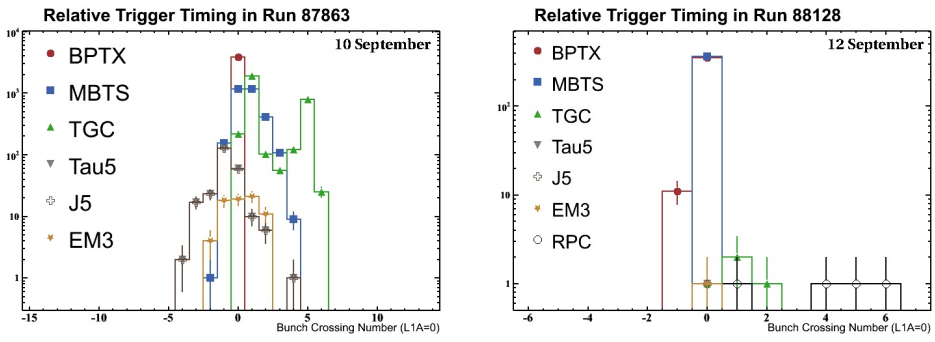


Fig. 26: Progress on timing-in ATLAS between 10 and 12 September 2008

As can be seen from the above, very significant progress was made on setting up the timing of the experiments within the first few days of single-beam operations at LHC. The experimental teams were eagerly awaiting further beam time and the first collisions that would have allowed them to continue the work. However, unfortunately, on 19 September there was a serious accident with the LHC machine that

required a prolonged shutdown for repairs and improvements. Nevertheless, when the LHC restarts one will be able to build on the work that was already done (complemented by many further studies that were done using cosmic rays during the machine shutdown).

A huge amount of work has been done using the beam-related data that were recorded in September 2008, as discussed in much more detail in the lectures of Andreas Hoecker at this School [27]. A very important feature of these data is that activity is seen in the same event in several detector subsystems which allows one to check the relative timing and spatial alignment. Indeed the fact that the same event is seen in the different subdetectors is reassuring — some previous experiments had teething problems where the readout of some of the subdetectors became desynchronized! A nice example of a beam-halo event recorded in CMS is shown in Fig. 27. Activity can be seen in the Cathode-Strip Chamber (CSC) muon detectors at both ends of the experiment and also in the hadronic calorimeter.

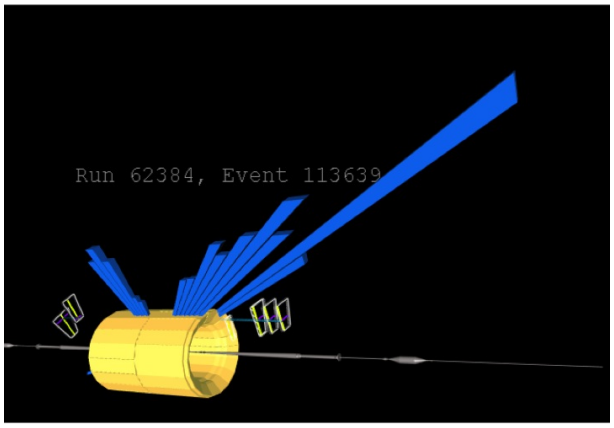


Fig. 27: A beam-halo event in CMS

The detectors and triggers that were used in September 2008 were sensitive to cosmic-ray muons as well as to beam-halo particles when a requirement of a signal from the beam pick-ups was not made. The presence of beam-halo and cosmic-ray signals in the data is illustrated in Fig. 28 which shows the angular distribution of muons reconstructed in CMS. The shape of the cosmic-ray distribution, which has a broad peak centred around 0.3–0.4 radians, is known from data collected without beam. The peak at low angles matches well with the distribution for simulated beam-halo particles.

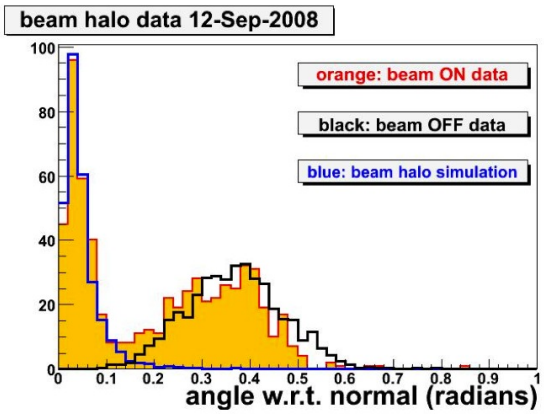


Fig. 28: Angular distribution of muons in CMS recorded with and without circulating beam. Also shown is the distribution for simulated beam-halo events.

Before concluding, the author would like to show another example of a study with single-beam data. Using a timing set-up in the end-cap muon trigger that would be appropriate for colliding-beam operations, in which the muons emerge from the centre of the apparatus, the distribution shown in the right-hand part of Fig. 29 was obtained. The two peaks separated by four bunch crossings, i.e., 4×25 ns, correspond to triggers seen in the two ends of the detector system. This is consistent within the resolution with the time of flight of the beam-halo particles that may trigger the experiment on the upstream or downstream sides of the detector. As indicated in the left-hand part of the figure, this is reminiscent of the very simple example that was introduced early on in the lectures, see Fig. 1.

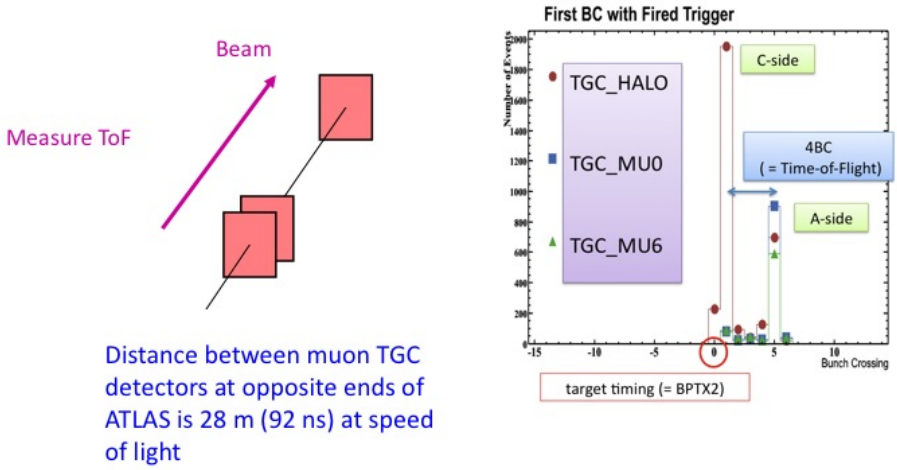


Fig. 29: Time of flight of beam-halo muons in ATLAS (one BC is 25 ns)

14 Concluding remarks

It is hoped that these lectures have succeeded in giving some insight into the challenges of building T/DAQ systems for HEP experiments. These include challenges connected with the physics (inventing algorithms that are fast, efficient for the physics of interest, and that give a large reduction in rate), and challenges in electronics and computing. It is also hoped that the lectures have demonstrated how the subject has evolved to meet the increasing demands, e.g., of LHC compared to LEP, by using new ideas based on new technologies.

Acknowledgements

The author would like to thank the local organizing committee for their wonderful hospitality during his stay in Colombia. In particular, he would like to thank Marta Losada and Enrico Nardi who, together, created such a wonderful atmosphere between all the participants, staff and students alike.

The author would like to thank the following people for their help and advice in preparing the lectures and the present notes: Bob Blair, Helfried Burckhart, Vincenzo Canale, Philippe Charpentier, Eric Eisenhandler, Markus Elsing, Philippe Farthouat, John Harvey, Andreas Hoecker, Jim Linnerman, Claudio Luci, Jordan Nash, Thilo Pauly, and Wesley Smith.

References

[1] L. Evans and P. Bryant (eds.), LHC machine, *JINST* **3** S08001 (2008).

[2] R.W. Assmann, M. Lamont, and S. Myers, A brief history of the LEP collider, *Nucl. Phys. B, Proc. Suppl.* **109** (2002) 17–31, http://cdsweb.cern.ch/record/549223.

[3] http://en.wikipedia.org/wiki/NIM

[4] http://en.wikipedia.org/wiki/Computer_Automated_Measurement_and_Control

[5] J. Schukraft, Heavy-ion physics at the LHC, in Proceedings of the 2003 CERN–CLAF School of High-Energy Physics, San Miguel Regla, Mexico, CERN-2006-001 (2006).
ALICE Collaboration, Trigger, Data Acquisition, High Level Trigger, Control System Technical Design Report, CERN-LHCC-2003-062 (2003).
The ALICE Collaboration, K. Aamodt et al., The ALICE experiment at the CERN LHC, *JINST* **3** S08002 (2008) and references therein.

[6] J. Nash, these proceedings.

[7] C. Amsler et al. (Particle Data Group), *Phys. Lett. B* **667** (2008) 1 also available online from http://pdg.lbl.gov/.

[8] W. von Rueden, The ALEPH data acquisition system, *IEEE Trans. Nucl. Sci*. **36** (1989) 1444–1448.
J. F. Renardy et al., Partitions and trigger supervision in ALEPH, *IEEE Trans. Nucl. Sci*. **36** (1989) 1464–1468.
A. Belk et al., DAQ software architecture for ALEPH, a large HEP experiment, *IEEE Trans. Nucl. Sci*. **36** (1989) 1534–1539.
P. Mato et al., The new slow control system for the ALEPH experiment at LEP, *Nucl. Instrum. Methods A* **352** (1994) 247–249.

[9] A. Augustinus et al., The DELPHI trigger system at LEP2 energies, *Nucl. Instrum. Methods A* **515** (2003) 782–799.
DELPHI Collaboration, Internal Notes DELPHI 1999-007 DAS 188 and DELPHI 2000-154 DAS 190 (unpublished).

[10] B. Adeva et al., The construction of the L3 experiment, *Nucl. Instrum. Methods A* **289** (1990) 35–102.
T. Angelov et al., Performances of the central L3 data acquisition system, *Nucl. Instrum. Methods A* **306** (1991) 536–539.
C. Dionisi et al., The third level trigger system of the L3 experiment at LEP, *Nucl. Instrum. Methods A* **336** (1993) 78–90 and references therein.

[11] J.T.M. Baines et al., The data acquisition system of the OPAL detector at LEP, *Nucl. Instrum. Methods A* **325** (1993) 271–293.

[12] http://en.wikipedia.org/wiki/VMEbus

[13] H1 Collaboration, The H1 detector, *Nucl. Instrum. Methods A* **386** (1997) 310.

[14] R. Carlin et al., The trigger of ZEUS, a flexible system for a high bunch crossing rate collider, *Nucl. Instrum. Methods A* **379** (1996) 542–544.
R. Carlin et al., Experience with the ZEUS trigger system, *Nucl. Phys. B*, *Proc. Suppl.* **44** (1995) 430–434.
W.H. Smith et al., The ZEUS trigger system, CERN-92-07, pp. 222–225.

[15] CDF IIb Collaboration, The CDF IIb Detector: Technical Design Report, FERMILAB-TM-2198 (2003).

[16] D0 Collaboration, RunIIb Upgrade Technical Design Report, FERMILAB-PUB-02-327-E (2002).

[17] R. Arcidiacono et al., The trigger supervisor of the NA48 experiment at CERN SPS, *Nucl. Instrum. Methods A* **443** (2000) 20–26 and references therein.

[18] T. Fuljahn et al., Concept of the first level trigger for HERA-B, *IEEE Trans. Nucl. Sci*. **45** (1998)

1782–1786.
M. Dam et al., Higher level trigger systems for the HERA-B experiment, *IEEE Trans. Nucl. Sci.* **45** (1998) 1787–1792.

[19] ATLAS Collaboration, First-Level Trigger Technical Design Report, CERN-LHCC-98-14 (1998).
ATLAS Collaboration, High-Level Triggers, Data Acquisition and Controls Technical Design Report, CERN-LHCC-2003-022 (2003).

[20] The ATLAS Collaboration, G. Aad et al., The ATLAS experiment at the CERN Large Hadron Collider, *JINST* **3** S08003 (2008) and references therein.

[21] CMS Collaboration, The Level-1 Trigger Technical Design Report, CERN-LHCC-2000-038 (2000).
CMS Collaboration, Data Acquisition and High-Level Trigger Technical Design Report, CERN-LHCC-2002-26 (2002).

[22] The CMS Collaboration, S. Chatrchyan et al., The CMS experiment at the CERN LHC, *JINST* **3** S08004 (2008) and references therein.

[23] See, for example, summary talks in Proc. Computing in High Energy and Nuclear Physics, CHEP 2003, `http://www.slac.stanford.edu/econf/C0303241/proceedings.html`

[24] LHCb Collaboration, Online System Technical Design Report, CERN-LHCC-2001-040 (2001).
LHCb Collaboration, Trigger System Technical Design Report, CERN-LHCC-2003-031 (2003).
The LHCb Collaboration, A. Augusto Alves Jr et al., The LHCb detector at the LHC, *JINST* **3** S08005 (2008) and references therein.

[25] R. Achenbach et al., The ATLAS level-1 calorimeter trigger, *JINST* **3** P03001 (2008).

[26] G. Navara et al., Electron trigger performance of the ATLAS detector, presented at *Signaling the Arrival of the LHC Era*, Trieste, Italy, 8–13 December 2008.

[27] A. Hoecker, these proceedings.

Commissioning and early physics analysis with the ATLAS and CMS experiments

A. Hoecker
CERN, Geneva, Switzerland

Abstract
These lecture notes for graduate students and young postdocs introduce the commissioning and early physics programme of the high-transverse-momentum experiments ATLAS and CMS, operating at the Large Hadron Collider (LHC) at CERN.

Preface *— This writeup of lectures given in March 2009 at the 5th Latin American School of High-Energy Physics, Recinto Quirama, Colombia, provides an overview of the various commissioning phases pursued by the ATLAS and CMS experiments to thoroughly prepare the detectors and data acquisition systems for physics. As an ATLAS member, the access to the relevant information from my own experiment was so invitingly easy that the document features an intolerable emphasis on ATLAS. I can only sincerely apologize to my CMS colleagues, and state that changing all figures shown into the corresponding ones from CMS would not alter the message the lectures seek to convey. In spite of their very different design, ATLAS and CMS have similar physics potential. Wherever significant performance differences exist, they are pointed out throughout these notes. Most of the analyses discussed here are taken from the vast ATLAS and CMS detector, performance, and physics reports [1–4]. No explicit reference is given when using results from these papers. While finalising these notes, the LHC restarted the commissioning programme in November 2009, after a year of repair and consolidation, achieving for the first time proton–proton collisions at 900 GeV centre-of-mass injection energy, and — for short periods — even the new world record energy of 2.36 TeV. Results from the analyses of collision data, which were not available at the time of the lectures, are not included in these notes.*

1 Motivation for a huge machine

The Large Hadron Collider (LHC) at the European Laboratory for Particle Physics Research (CERN) is the most powerful proton accelerator ever built. It collides two beams of protons accelerated to 7 TeV each and bent by dipole magnets with 8.3 T magnetic field strength within the 26.7 km circular collider, immersed in a ca. 100 m deep tunnel between Lake Geneva and the French Jura mountains. If the proton were an elementary particle, that is, if it were point-like, the 14 TeV centre-of-mass energy released by the collision could be fully transformed into mass. Dependent on quantum numbers and conservation laws (symmetries), for example one heavy particle of 14 TeV or a particle–antiparticle pair of 7 TeV each (masses of particles and antiparticles are identical) could be produced. Since the heaviest known particle is the top quark with mass of 173 GeV, any heavier particle found would be a discovery. These new heavy particles might decay to other new particles, still heavier than the top quark, and henceforth a full cascade of new particles could be discovered. The proton is, however, not an elementary particle, but is made out of a cloud of quarks and gluons (partons). The collision of two protons can thus be regarded as collisions between partons with momentum fractions that follow a density distribution with long tails towards one. Unlike for instance at e^+e^- colliders, increasing the number of recorded collisions increases the probability for the occurrence of very hard parton scattering involving large fractions of the proton–proton centre-of-mass energy. A high-luminosity 14 TeV proton–proton collider therefore allows the experiments to deeply explore the TeV scale.

What does TeV scale signify? Let us recall the relevant atomic, nuclear, and particle physics scales. The only known massless elementary particles are photons and gluons (bosons), which propagate the electromagnetic and strong forces, respectively. The lightest fermions are the neutrinos with masses probably lower than a few eV. This is below the atomic binding energy, which reaches tens of eV.

The next orders of magnitudes are represented by the electron mass (1 MeV), nuclear binding energy (up to 10 MeV), pion and muon masses (100 MeV), the heaviest known lepton as well as proton, neutron, and vector-meson masses (1 GeV), the $c\bar{c}$ and $b\bar{b}$ resonances and heavy-quark mesons (10 GeV), and finally the electroweak unification scale, represented by the masses of the Z and W weak-interaction bosons, the top quark, and (presumably) the Higgs boson (100 GeV) and the Higgs vacuum expectation value (246 GeV). No particles beyond that scale are known to date.

However, as we shall see later, the requirement of a stable Higgs sector suggests the existence of new phenomena at the TeV scale, which is precisely the area of sensitivity of the LHC. Little is known beyond that scale. Will new symmetries arise, the breaking of which generates new particles? The seesaw mechanism accommodating massive neutrinos predicts heavy right-handed Majorana neutrinos of mass up to 10^{14} GeV. Unification of the electroweak and strong interactions may occur at 10^{16} GeV. Finally, gravitation becomes strong at the particle level at the Planck scale of order 10^{18} GeV, requiring a quantum field theory that includes gravitation. The minimal Standard Model (assuming massless neutrinos) of unified electroweak and strong interactions includes 19 free parameters, among which are 3 coupling constants, 1 spin-1 and 1 spin-0 boson mass, 9 fermion masses, 3 weak quark mixing angles, 1 *CP*-violating weak phase, and 1 *CP*-violating strong phase, which is either tiny or zero. Including a massive neutrino sector increases the number of free parameters by at least 9, depending on the nature of the neutrinos.

The dynamical predictions of the Standard Model have been verified to extreme precision in the past thirty-five years at a large number of very different experiments. Let us recall a few eminent examples. The cross section of lepton pair production has been measured to order 1 TeV and found in agreement with the Z resonance being the highest particle decaying into two leptons, and Drell–Yan production being the dominant process beyond the Z (*cf.* topmost plot in Fig. 1 [5]). Electroweak unification has been tested by globally fitting the Standard Model prediction to precision measurements obtained at the high-energy e^+e^- colliders LEP (CERN) and SLC (SLAC), and at the $p\overline{p}$ collider Tevatron (FNAL). The second plot from the top in Fig. 1 shows the relation between measured and predicted W-boson mass versus the top-quark mass [6]. The universality of weak interactions has been verified at the 0.3% level by comparing the tau branching fractions to electron and muon plus neutrinos and to the tau-lepton lifetime (*cf.* bottom left plot in Fig. 1 [7]). The asymptotic freedom property of QCD has been verified at the 1% level by measuring the evolution ('running') of the strong coupling at

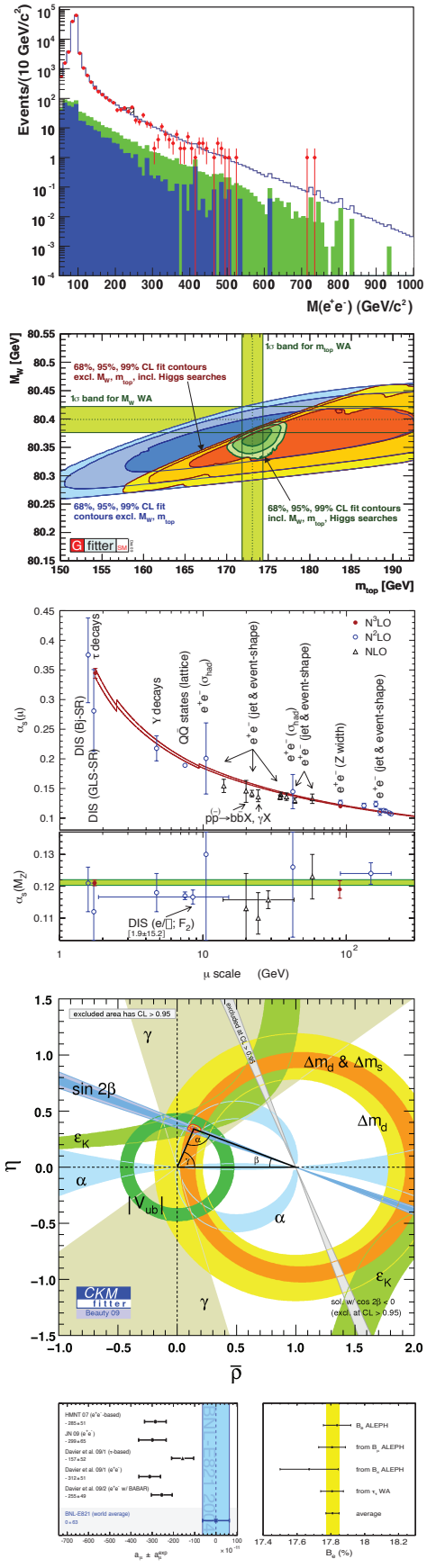


Fig. 1: Tests of the Standard Model.

various energy scales, the most precise of which being the ones at the τ and the Z mass scales (third plot in Fig. 1 [8]). The Standard Model predicts that all *CP*-violating phenomena involving weak charged currents originate from a single phase in the quark mixing matrix. This has been verified by relating different measurements of *CP* violation in the *B*-meson and kaon sectors to each other, all showing compatibility (fourth plot in Fig. 1 [9]). The *CP*-violating electric dipole moment of the electron has been found to be smaller than 10^{-27} ecm as predicted by the Standard Model. The anomalous magnetic moment of the muon has been measured to the parts-per-million level, verifying the predicted contributions from electromagnetic, weak, and hadronic loop corrections. A small deviation from the expectation is currently not at a sufficiently significant level to draw conclusions (*cf.* bottom right plot in Fig. 1 [10]). Many more examples all confirm the Standard Model. So, what's the problem?

As explained in much detail by John Ellis [11] and others at this school, the Standard Model — though describing so gloriously the experimental data — is, at best, incomplete. Firstly, the Higgs boson, the last elusive Standard Model ingredient, has not yet been discovered. Even if it were discovered, it would be the only elementary scalar particle in the Standard Model, which — for many physicists — is conceptually unsatisfactory. A popular question is the origin of the large mass hierarchy between fermions of different generations, amounting to more than 4 orders of magnitude between top and up quarks. Many astrophysical observations have established the presence of cold dark matter in the galaxies and galactic halos. Moreover, spurious repulsive 'dark energy' appears to accelerate the expansion of the universe. In particle physics, we can use the standard quantum field theory renormalisation groups to predict the energy-scale dependence of the electroweak and strong coupling constants. Evolving the three couplings to 10^{16} GeV, they *almost* converge towards a single unified coupling — almost, but not quite. While unification might be considered an aesthetic requirement, stability of the Higgs sector is not. Indeed, the virtual loop corrections, in particular from top-pair vacuum polarisation, diverge quadratically with their high-energy cut-off. Also, perturbativity of the Higgs quartic coupling and stability of the Higgs potential require the Higgs mass to lie within a small allowed window, if the Standard Model is to survive up to the (reduced) Planck scale $M_P \simeq 2 \cdot 10^{18}$ GeV. Moreover, how would the unification of the Standard Model and gravitation be established at that scale? A subtle, but no less intriguing problem is the apparent smallness of the strong-*CP* parameter, tightly bound from measurements of the neutron electric dipole moment, although no mechanism such as a symmetry in the Standard Model suggests such a small or even vanishing value. While the Standard Model features *CP* violation in the charged weak current, theoretical calculations show that the amount of *CP* violation is insufficient by many orders of magnitude to be at the origin of the matter–antimatter asymmetry currently observed in the visible part of the universe.[1]

The instability of the mass of the scalar Higgs boson against radiative corrections is denoted by the term 'gauge hierarchy problem', which also sets the scale at which new physics can be expected. It is — beyond the Higgs discovery and the strong Standard Model research programme — a primary motivation for the construction of the LHC. Indeed, if a Higgs boson with mass <1 TeV is discovered, the Standard Model is complete. However, when computing radiative corrections to the Higgs propagator, modifying the bare Higgs mass, such as $t\bar{t}$ vacuum polarisation diagrams, or boson self-energies including the Higgs self-coupling, the corresponding loop integrals diverge. To solve them, a cut-off parameter $\Lambda_{\text{cut-off}}$ is introduced to which the integrals are quadratically proportional. The cut-off parameter sets the scale where new particles and physical laws must come in, regularising the diverging integral.[2] However, above the electroweak scale we know of only two scales exhibiting new physics: grand unification of the

[1] We could thus ask ourselves what the role of the weak phase is in the evolution of the universe. Does it carry a hidden purpose? Or is weak *CP* violation a meaningless 'accident of Nature': because there are three generations and because all quark flavours have mass there is quark mixing with four parameters of which three are three Euler angles and one is a *CP*-violating phase. The phase is not constrained by a symmetry and thus of order one (68° [9]). Perhaps without major implications for Nature.

[2] In a renormalisable quantum field theory, divergences in single loop integrals frequently occur, but they are always cancelled to all perturbative orders by other diagrams contributing to the full matrix element of the scattering process under study.

electroweak and strong forces ($\approx 10^{16}$ GeV) and the Planck scale. A cut-off at such large energies would require an enormous amount of fine-tuning to keep the physical Higgs mass small and stable. What could be a 'natural' value for the scale $\Lambda_{\text{cut-off}}$? The following three diagrams give the largest contributions to the Higgs radiative corrections and hence to the physical Higgs mass: $t\bar{t}$ loop: $-(3/8\pi^2)\lambda_t^2\Lambda^2_{\text{cut-off}} \approx (2\,\text{TeV})^2$; gauge-boson loop: $(9/64\pi^2)g^2\Lambda^2_{\text{cut-off}} \approx (0.7\,\text{TeV})^2$; and Higgs loop: $-(1/16\pi^2)\lambda^2\Lambda^2_{\text{cut-off}} \approx (0.5\,\text{TeV})^2$, where we have used $\Lambda_{\text{cut-off}} = 10\,\text{TeV}$ everywhere, and where λ_t, g, λ are respectively CKM, weak, and quartic Higgs couplings. The total mass-squared of the Higgs is the sum of these contributions and the tree-level term. What would be the cut-off (= new physics) scales if only small ($\sim$10%) fine-tuning were allowed? We would find $\Lambda_{\text{top}} < 2\,\text{TeV}$, $\Lambda_{\text{gauge}} < 5\,\text{TeV}$, and $\Lambda_{\text{Higgs}} < 10\,\text{TeV}$. To naturally cancel these divergences, new physics at the TeV scale should couple to the Higgs and should be related to the particles in the loop (top, gauge, Higgs) by some symmetry.

The gauge hierarchy problem denotes this fine-tuning of parameters, and the strong dependence of physics at the weak scale on the physics at (presumably) much higher scale: if the Higgs radiative corrections are cut off at the scale of gravity, why is the scale of electroweak symmetry breaking so different from the scale of gravity? Why is $m_W \ll M_P$? Equivalently, why is gravity so weak? Possible solutions to the hierarchy problem include: (i) new physics appears not much above the electroweak scale and regularises the quadratic divergences, (ii) new physics modifies the running of the couplings, approaching grand unification to the electroweak scale, (iii) gravity is not as weak as we think, it is only diluted in our four-dimensional world but it is as strong as electroweak interactions in, e.g., five or more dimensions with Planck scale $M_P^{(5D)}$ $\mathcal{O}(\text{TeV})$, or (iv) the theory is fine-tuned and the explanation for the parameter values is statistical rather than dynamic (anthropic principle).

From the above discussion we retain that the Standard Model is in crisis. Most Standard Model extensions, developed with the goal to solve the hierarchy problem and/or to provide a dark matter candidate, introduce new particles at the TeV scale. To find these, we need a new, huge collider providing hard particle collisions with centre-of-mass energy well above 1 TeV.

2 The Large Hadron Collider

In principle, one could accelerate protons circulating in a magnetic ring almost illimitably to higher and higher energy by continuously passing them through a radio-frequency field. The energy loss through synchrotron radiation of a proton in the Large Hadron Collider (LHC) amounts to a few keV per turn (compared to a few GeV per turn for electrons in the e^+e^- collider LEP2), which is about one hundred times smaller than the acceleration the proton receives per turn. In practice however, the proton energy in the collider ring is limited by the superconducting dipole magnets that guide the circular beams: $E_{\text{proton}} \simeq 0.3 \cdot B \cdot r$. Because the radius of the LHC is fixed ($r = 4.3\,\text{km}$), one must use as strong fields as possible (8.3 T, compared to approximately 4 T at the HERA and Tevatron colliders), and fill all free LHC sections with dipole magnets ($\approx$2/3).[3] Because the effective centre-of-mass energy of the hard parton collision depends on the parton energy density distributions in the proton, with long tails towards a large energy fraction, accumulating larger statistics due to a high instantaneous luminosity effectively increases the available kinematic reach of the proton–proton collider. High luminosity (beyond $10^{33}\,\text{cm}^{-2}\text{s}^{-1}$), and good machine and data-taking efficiency (of the order of 10^7 seconds good-quality data taking per year), are also required to search for rare events, such as processes involving the Higgs boson, especially if the Higgs is light (Higgs production is an electroweak process with large momentum transfer, which has a cross section roughly a billion times smaller than inelastic QCD (so-called 'minimum bias') processes), and also for studies of the nature of new physics phenomena if discovered. To achieve high luminosity (L), strong currents are necessary, requiring dense proton bunches containing up to $N = 110$ billion protons each (for comparison: $1\,\text{cm}^3$ of hydrogen contains $\approx 10^{19}$ protons), and as many LHC bunches

[3] More precisely, the total number of dipole magnets in the LHC is 1232, each of which has a magnetic length of 14.3 m, giving a total length of 17 618 m. The effective 'bending radius' amounts thus to: $17\,618/(2\pi) = 2804\,\text{m}$, and hence $E_{\text{proton}} \simeq 0.3 \cdot B \cdot r \approx 7\,\text{TeV}$.

(k) as possible filled with protons (maximum of $k = 2808$ bunches out of a total of 3564 bunches). The bunches are spaced by 25 ns from each other, corresponding to a distance of 7.5 m. High luminosity also requires that the protons be transversely squeezed by magnetic lenses to a small spot to increase the probability that two protons collide. The typical transverse beam size, determined by the square-root of the product of an amplitude function characterising the beam optics (varying throughout the ring), and the constant phase space volume (emittance), amounts to $\sigma_x^\star = \sigma_y^\star = 16\,\mu$m at 7 TeV beam energy (for smaller beam energies, the beam emittance increases with $\varepsilon \propto 1/\gamma_p$, as does the beam spot size as $\propto \sqrt{\varepsilon}$).[4] The luminosity value is obtained from the formula

$$L = \frac{kN^2 f}{4\pi\sigma_x^\star\sigma_y^\star}\,, \tag{1}$$

where $f = 11.25$ kHz is the revolution frequency determined by the LHC circumference and the speed of light of the protons. We thus obtain $L = 3.5\cdot 10^{30}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$ per bunch, reaching $10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$ when all bunches are filled.

The LHC acceleration chain involves several steps (see Fig. 2 for a schematic view). The injector complex consists of the LINAC-2, preaccelerating the protons to 50 MeV, followed by the Proton Synchrotron Booster (PSB) consisting of four superimposed rings accelerating the protons to 1.4 GeV. Two large circular rings further accelerate the protons to 26 GeV (Proton Synchrotron – PS) and 450 GeV (Super Proton Synchrotron – SPS), which is the LHC injection energy. The beams are transferred from the SPS to the LHC via two newly built 3 km transfer lines. The PSB–PS–SPS complex required significant upgrades to be able to provide beams with the appropriate intensity, size, and bunch distance. The injection chain is particularly delicate because any increase of beam emittance during injection will be 'remembered' by the protons in the LHC and lead to a reduction of the available peak luminosity and/or beam lifetime (thus increasing beam-related backgrounds and reducing the integrated luminosity the LHC can deliver during a proton fill). We note that in each acceleration step, the energy increase lies between a factor of 10 and 20, which are reasonable ranges for the dipole magnets. The injector also has the task of creating the proton bunches and (fixed) bunch pattern for the LHC. The chain is as follows: 6 booster bunches are injected into the PS; each of these is split into 12 smaller bunches giving a total of 72 bunches at extraction; between 2 and 4 batches of 72 bunches are injected into the SPS giving from 144 up to 288 bunches; finally, a sequence of 12 extractions of (up to) 288 SPS bunches is injected into the LHC, giving a maximum of 2808 bunches (39 groups of 72 bunches). The filling scheme (difference between the 3564 possible and 2808 actually filled bunches) foresees a number of short gaps for, e.g., kicker magnet rise times in the injection chain, and one long gap of 119 empty bunches (3 μs) for the rise time of the LHC beam dump kicker magnet. Once injected into the LHC, the protons are accelerated from 450 GeV to 7 TeV in a 20-minute acceleration process, during which the protons receive an average energy gain of 0.5 MeV per turn when passing the electrical radio-frequency (RF) fields created in 8 superconducting cavities per beam with a peak accelerating voltage of 16 MV.

The LHC consists of eight 2.45-km-long arcs with bending dipole magnets (see Fig. 3 for a schematic drawing of a dipole section),[5] and eight 545-m-long straight sections. Four particle detectors have been constructed and are housed in huge underground caverns located at four of the straight sections. They record the objects left by collision debris by interacting with them. The detectors are: ATLAS (A Toroidal LHC ApparatuS), CMS (the Compact Muon Solenoid), ALICE (A Large Ion Collider

[4] The free 'volume' occupied by each proton in the interaction point is of the order of $10^{-4}\,\mu\mathrm{m}^3$, which is huge compared to the size of an atom, so that strong-interaction collisions between protons are still rare. The probability of two protons colliding can be estimated to be approximately $4\cdot 10^{-21}$, so that with $1.1\cdot 10^{10}$ protons per bunch one finds ≈ 50 interactions per bunch crossing, of which, however, only one-half are inelastic.

[5] The LHC magnet systems consists of a total of 1232 superconducting dipoles (cooled with 120 tons of superfluid helium down to 1.9 K), in which currents of 12 kA create the required 8.33 T magnetic field; 392 focusing quadrupoles; and 3700 multipole corrector magnets.

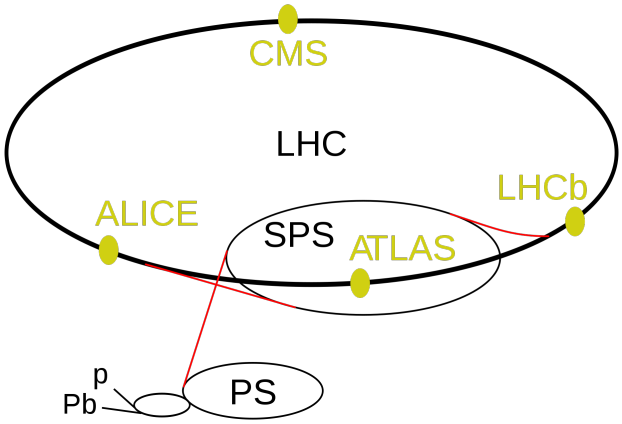


Fig. 2: Schematic view of the main elements of the LHC accelerator complex (see text) and the location of the four largest LHC experiments ALICE, ATLAS, CMS and LHCb.

Experiment), and LHCb (study of physics in B-meson decays at the LHC).[6] The remaining four straight sections are used by the RF cavities, the beam dump, and by two beam-cleaning systems using chains of collimators to absorb off-beam protons that would provoke magnet quenches and create so-called beam-halo backgrounds in the experiments. Although the energy of a single 7 TeV proton corresponds to only that of a flying mosquito (1 μJ), the total stored energy of 2808 bunches each filled with 10^{11} 7 TeV protons amounts to 360 MJ.[7] It is a huge challenge to control this energy and avoid damage to accelerator and experiments.

3 The high-p_T general-purpose detectors ATLAS and CMS

The broad range of physics opportunities and the demanding experimental environment at high-luminosity 14 TeV proton–proton collisions impose unprecedented performance requirements and technological constraints upon the LHC particle detectors. ATLAS and CMS are general-purpose detectors, capable of adequately covering the entire physics programme reachable with high-luminosity 14 TeV proton–proton collisions: from charm and beauty physics at lowest transverse momenta ($\sim$3 GeV), to new physics searches up to the highest reachable scales ($\sim$4 TeV). The cross sections of the dominant QCD processes and those representing the primary physics channels for research differ by many orders of magnitude. For example, while at 14 TeV centre-of-mass energy, the total inelastic pp cross section amounts to approximately 70 mb (giving a 1 GHz event rate at $L = 10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$),[8] hard quark and gluon scattering into pairs of jets (or more) occurs roughly a thousand times less frequently; inclusive b-hadron production has a cross section of approximately 0.5 mb; inclusive $W \to \ell\nu$ and $Z \to \ell\ell$ boson production and decay have cross sections times branching fractions of approximately 20 nb and 2 nb,[9] respectively (compared to roughly a factor of 8 smaller at the Tevatron); top and antitop production has a cross section of almost 1 nb (rate of 10 Hz), two orders of magnitude higher than at the Tevatron; inclusive Higgs-boson production, dominated by gluon-gluon-to-Higgs fusion via a triangular top-quark loop, has a Higgs-

[6] In addition, there are two smaller experiments: TOTEM (Total Cross Section, Elastic Scattering and Diffraction Dissociation at the LHC) and LHCf (Large Hadron Collider forward) for very low-p_T physics.

[7] The stored energy is sufficient to heat up and melt 12 tonnes of copper. It is equivalent to an Airbus A380 flying at 700 km/h speed, to 90 kg of TNT, 8 litres of gasoline, or 15 kg of chocolate.

[8] Recall that $1\ \mathrm{mb}^{-1} = 10^{27}\,\mathrm{cm}^{-2}$.

[9] Because of the proton quark structure, producing more $u\bar{d}$ than $\bar{u}d$ quarks in scattering reactions, roughly a quarter more W^+ than W^- are produced at the LHC [12] (while equal amounts of both charges are produced at the CP symmetric Tevatron collider).

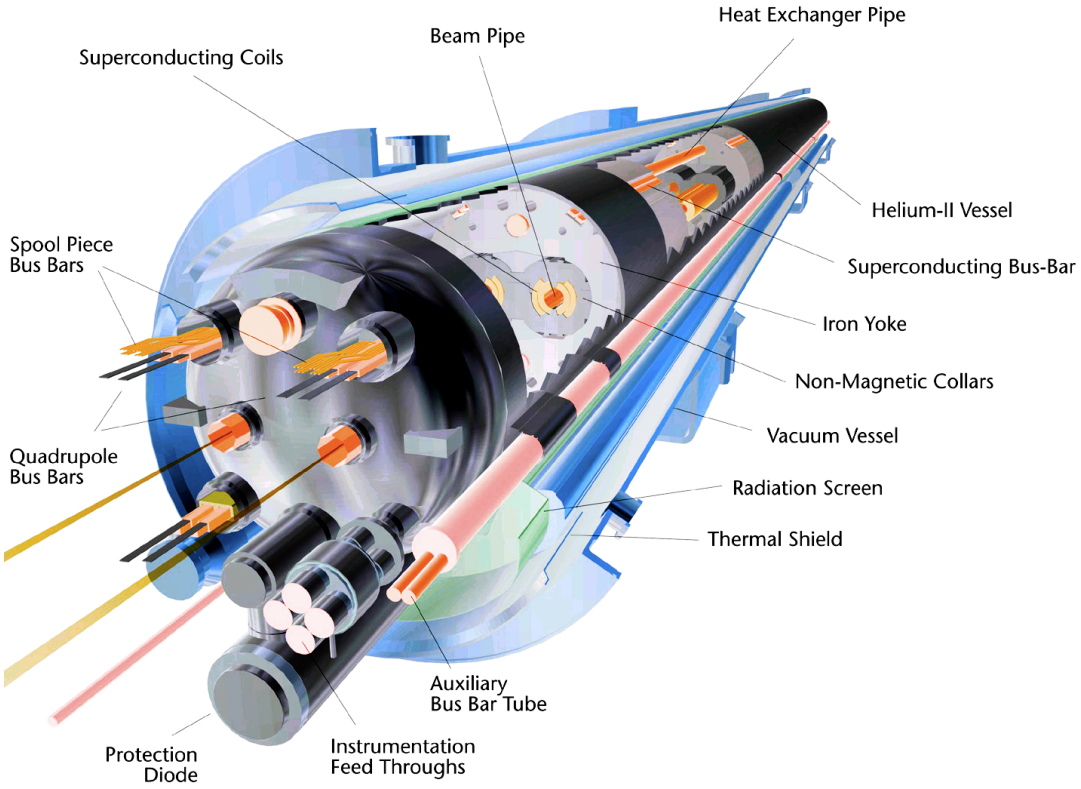


Fig. 3: Section view of a superconducting LHC dipole magnetic. The two beam pipes are wrapped into two oppositely poled superconducting coils.

mass dependent cross section between 45 pb ($m_H = 120\,\text{GeV}$) and 20 pb (180 GeV); and the production via gluon–gluon scattering of 1 TeV supersymmetric squarks and gluinos has a cross section of a few pb. These vast disparities, rendering physics analysis at the LHC like searching for needles in a giant haystack, drive the detector design.

Let us list some of the most outstanding LHC conditions and derive from these the corresponding design challenges.

- The 40 MHz bunch crossing rate[10] requires a fast trigger decision, precise timing and 'pipeline' electronics, locally storing readout data until the Level-1 (hardware) trigger response signal has been derived. For a pipeline memory depth of 100 bunch crossings, the Level-1 trigger latency must not exceed 2.5 μs.
- The interaction rate of up to 1 GHz at maximum peak luminosity of $10^{34}\,\text{cm}^{-2}\text{s}^{-1}$ (LEP and Tevatron: $L_{\text{max}} = 10^{32}\,\text{cm}^{-2}\text{s}^{-1}$ and $3.5\cdot 10^{32}\,\text{cm}^{-2}\text{s}^{-1}$, respectively), corresponding to approximately 25 inelastic interactions piling up in a single collision event, requires efficient pattern recognition to reduce the event rate from 1 GHz to 75 kHz (Level-1 output, high-level trigger input) to approximately 200 Hz (HLT output rate, events written to disk).
- The approximate data size of 1.5 MB per event together with the 200 Hz accepted trigger rate provides an average raw data throughput of 300 MB per second. Storage, worldwide distribution, prompt reconstruction and reprocessing of these data require adequate storage media, and powerful network and computing resources. The paradigm of distributed computing chosen by the LHC

[10] For comparison, the bunch crossing rates at LEP and the Tevatron are 45 kHz and 2.5 MHz, respectively, while the *B* factory PEP-II, an e^+e^- collider, has achieved 240 MHz, and the CLIC design foresees 2 GHz.

experiments requires the availability of several (order 10) large-scale computing centres (Tier-1s, demarcating 'computing clouds'), with resources similar to those at CERN, and located representatively for the collaborations' geographical extensions. These clouds embrace smaller computing centres for user analysis and simulation production.

– The irradiation rate after 10 years of successful LHC operation is expected to reach $5 \cdot 10^{14}$ neutron equivalents per cm^2 (300 kGy), requiring radiation-hard inner tracker (pixel detector with large signal-to-noise ratio and small silicon volume close to the interaction point) and forward calorimeter technology.
– The high charged multiplicity of up to 1000 tracks per event ($4 \cdot 10^{10}$ tracks per second) requires the use of high-granular pixel/silicon or fine-grained straw tracker technologies. Three-dimensional pixel technology, replacing traditional silicon strip detectors close to the beam pipe, is mandatory to provide sufficient pattern recognition capability.
– Large background rates from beam-gas interactions, beam-halo muons, thermal neutrons and photons ('cavern background', bathing the detector during event pileup and afterwards due to activation of materials in the detector, its support structure, and the cavern), require precise muon timing, redundant pattern recognition, and radiation hardness.

Similarly, the detector design reflects the challenges posed by the physics programme.

– The search for rare $B_{s(d)} \to \mu\mu$ decays, which have Standard Model branching fractions of $3.3 \cdot 10^{-9}$ and $1.1 \cdot 10^{-12}$, respectively, and the measurement of time-dependent *CP* violation and the unitarity triangle angle β_s using (among others) flavour-tagged $B_s \to J/\psi\phi$ decays, require good trigger efficiency and purity for muon tracks with transverse momenta as low as 3 GeV. To achieve sufficient purity, the HLT tracking algorithm must reconstruct charges as well as the B vertex and mass.
– Measuring the W mass to a precision better than the current world average [13] of (80.399 ± 0.023) GeV, requires excellent alignment of the tracking detectors, good track reconstruction efficiency, calorimeter uniformity, and missing transverse energy resolution.
– A precision measurement of the top mass needs — apart from a better theoretical understanding of the nature of the measured top mass — excellent jet energy calibration, resolution and uniformity, as well as excellent b-tagging purity and efficiency.
– A sensitive search for the Higgs boson in the most promising final states $2e(\mu)2\nu$, $4e(\mu)$, $2e2\mu$, $\gamma\gamma$, $\tau\tau$ (via weak boson fusion accompanied by forward jets) requires very pure and efficient particle identification, excellent electromagnetic and hadronic calorimeter resolution and uniformity, efficient high-luminosity tracking, and efficient reconstruction of forward jets.
– Searching for the multifaceted signatures from supersymmetry requires excellent jet and missing transverse energy resolution, low calorimeter noise, excellent τ identification and reconstruction, as well as maximum detector acceptance.
– The search for heavy resonances of masses beyond 1 TeV, as they are predicted in models with excited weak bosons or extra spatial dimensions, requires good tracking (including charge reconstruction) and calorimeter resolution, and a large dynamic range (small calorimeter saturation) up to the highest reachable energies.

3.1 Detector design

The high-p_T detectors, ATLAS and CMS, are designed as a result of careful optimisation processes to respond as well as possible to these unprecedented and sometimes conflicting requirements, while respecting budget limitations (approximately 550 million Swiss francs per detector). Both detectors have fast, multi-level trigger systems allowing one to select complex signatures, fast data acquisition based on broadband network switches, excellent inner tracking devices allowing efficient high-p_T tracking

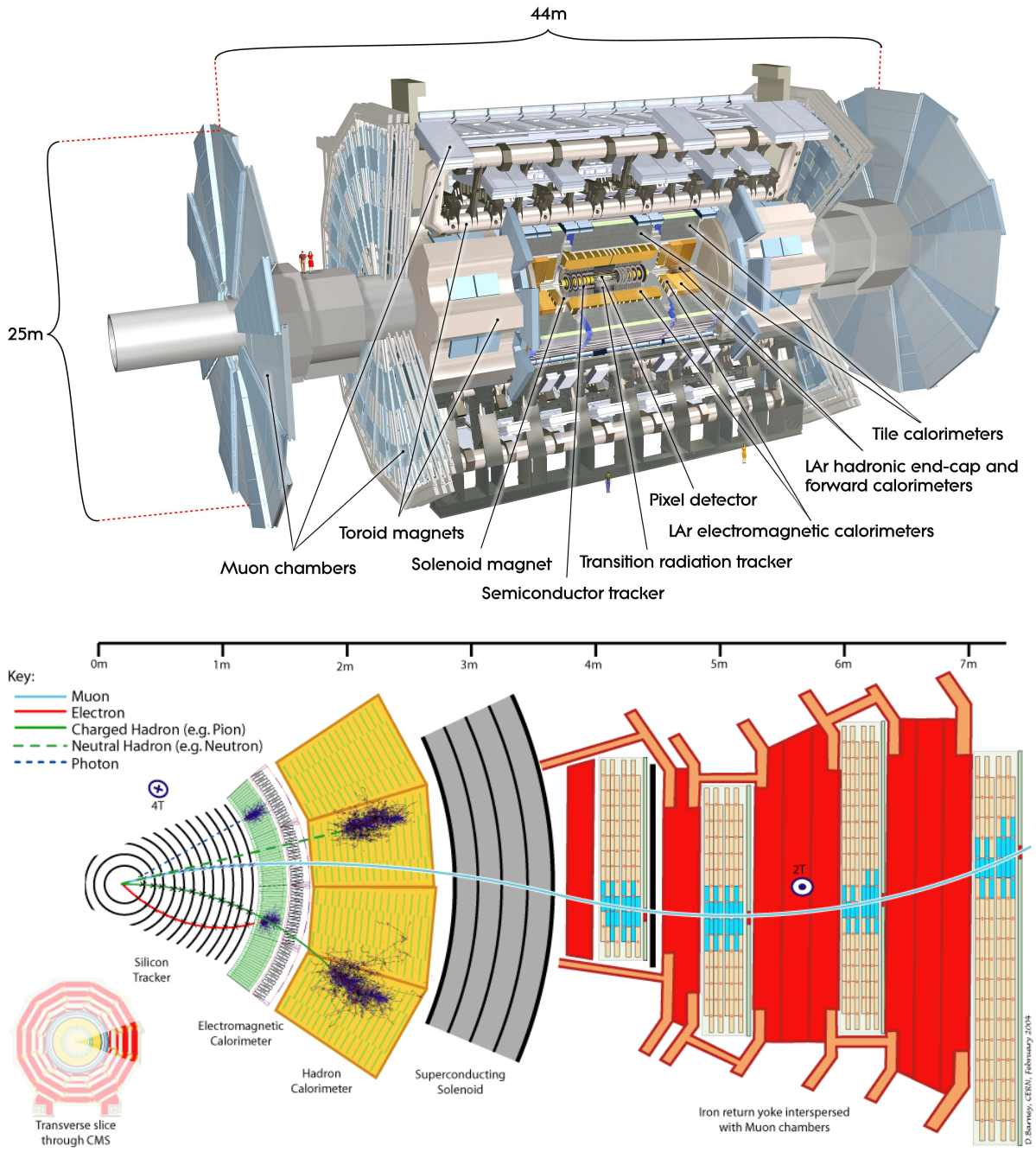


Fig. 4: Schematic drawings of the ATLAS detector (upper) and a slice of the CMS detector (lower), showing the trajectories of charged and neutral particles interacting with the various detector layers.

and secondary (b) vertex reconstruction in a high-luminosity environment; fine-grained, high-resolution electromagnetic calorimeters for excellent electron and photon reconstruction, complemented by full coverage hadronic calorimetry for accurate jet and missing transverse energy measurements, and an efficient identification of semileptonic τ lepton decays; as well as high-precision muon systems with standalone tracking capability [1, 2, 14, 15]. Schematic drawings of the ATLAS and CMS detectors are shown in Fig. 4.

The most striking difference between ATLAS and CMS, strongly determining the entire detector design, is the **magnet structures**. CMS has a single, albeit huge solenoid (inner diameter 5.9 m, thick-

ness 60 cm, axial length 12.9 m), fully immersing the inner tracking systems and electromagnetic and hadronic calorimeters in a 3.8 T axial magnetic field[11] (18.2 kA current), and providing muon momentum measurement via the ∼2 T field in the flux return yoke made out of 10 000 tonnes of steel. ATLAS has three different magnet systems: a thin solenoid (inner diameter 2.46 m, thickness 5 cm, axial length 5.8 m, axial magnetic field 2 T at the centre of the tracking volume, 7.7 kA current) around the inner tracking system, and 8 barrel and 2×8 endcap air-core toroid magnets (magnetic fields between 0.5 T and 4 T, strongly varying with the radial distance from the toroids, 20.5 kA currents), arranged radially around the hadron calorimeters such that the Lorentz force bends charged tracks along their z coordinates. The toroid magnets do not affect the central solenoid field. All magnet systems are superconducting.

The **inner tracking systems** are made out of semiconducting silicon pixel and silicon strip detectors for the inner and outer layers (disks in the endcaps), respectively, comprising approximately 80 million channels. Pixel systems close to the collision impact point are mandatory to cope with the large track density. The innermost barrel pixel layer, of a total of 3 layers, is as close as 5.0 cm (ATLAS) and 4.4 cm (CMS) to the beam line. The design $R\phi$ position resolution of the pixel system is 10 μm. In CMS silicon strip technology is used to cover the entire inner detector between pixel and electromagnetic calorimeter (radius of the outermost layer: 107–110 cm), providing a total of 14 measurement points. The ATLAS silicon strip detector, being shorter in radius, provides 8 measurement points. A transition radiation tracker made of 350 000 Kapton straw tubes of 4 mm diameter, providing on average 35 measurement points for pseudorapidity[12] lower than 1.8 (resolution of 130 μm per straw), and between 18 and 35 $R\phi$ measurement points (no η measurement) between $1.8 \leq |\eta| \leq 2.5$, is inserted between the silicon strip tracker and solenoid. Transition radiation with 8 keV photons on average, emitted when charged ultrarelativistic particles traverse the boundary of two different dielectric media (foil and air), increases the signal size so that dual readout with low and high thresholds allows the identification of $\beta = 1$ particles (electrons).

Owing mainly to the stronger solenoid magnetic field, CMS has better momentum resolution with $\sigma(p_T) \simeq 1.5\%$ compared to 3.8% (ATLAS) for 100 GeV tracks at $\eta = 0$. At low momentum, multiple scattering that occurs due to the significant material in the tracking systems of both detectors (varying between $0.3X_0$ at $\eta \simeq 0$ and $1.4X_0$ at $\eta \simeq 1.5$) reduces this difference.

The **electromagnetic calorimeters** consist of a lead and liquid-argon sampling technique, radially shaped as an accordion to minimise inhomogeneities and cracks, chosen by ATLAS, versus high-granular lead tungstate ($PbWO_4$) scintillating crystals in CMS (61,200 crystals in the barrel and 7,324 in each endcap). Both calorimeters have a geometry pointing towards the collision point, which simplifies the energy reconstruction of the incident particles. The lead absorber in the ATLAS calorimeter reduces the available light yield for energy measurement, thus limiting the stochastic resolution to $\sigma(E) \simeq 10\text{–}12/\sqrt{E}$ with a constant term of 0.2–0.35%, compared to $\sigma(E) \simeq 3\text{–}5.5/\sqrt{E}$ and a constant term of 0.5% for the CMS crystals. The influence of the constant term, originating from non-uniformities in the calorimeter response due to inhomogeneities and non-linearities, is small for ATLAS, while it becomes a limiting factor at energies beyond 40 GeV for CMS (hence, for example, affecting the measurement of $H \to \gamma\gamma$). While CMS has only a single electromagnetic layer, the ATLAS calorimeter is longitudinally segmented in four layers (including the presampler, which corrects the measured energy for early electromagnetic showers in solenoid and cryostat), permitting one to measure the shower development and so distinguish electromagnetic from hadronic showers. It also allows one to reconstruct the direction of the incoming

[11] The solenoid is designed to deliver a 4 T field. Longevity considerations have however led to the decision to decrease the current from 19.5 kA to 18.2 kA, reducing the field to 3.8 T.

[12] The pseudorapidity is defined by

$$\eta \equiv -\ln\left(\tan\frac{\theta}{2}\right) = \frac{1}{2}\ln\left(\frac{|\mathbf{p}| + p_L}{|\mathbf{p}| - p_L}\right), \tag{2}$$

where θ is the polar angle between the particle momentum $\mathbf{p}$ and the beam axis (z), and p_L is the longitudinal component of $\mathbf{p}$. In hadron collider physics, the pseudorapidity is preferred over the use of the polar angle because particle production is constant as a function of the pseudorapidity.

particle. The cell granularity for the ATLAS main sampling layer is $\Delta\eta \times \Delta\phi = 0.025^2$ rad, improved in η by fine strips with $\Delta\eta = 0.003$ (barrel number) in front of the main sampling layer to help identifying π^0. CMS has a crystal granularity of $\Delta\eta \times \Delta\phi = 0.017^2$ rad in the barrel, and 0.018×0.003 to 0.088×0.015 in the endcaps. Saturation of the energy reconstruction occurs for energy depositions beyond 3 TeV (ATLAS) and 1.7 TeV (CMS). Biases due to saturation are corrected but lead to a decrease in the energy resolution.

The **hadronic calorimeters** use similar sampling techniques, based on iron (ATLAS) and brass absorbers (CMS) and scintillating tiles read out via wavelength shifting optical fibres guiding the light to photomultiplier tubes. The main difference in performance originates from the strong constraint imposed by the maximum achievable size of the CMS solenoid, resulting in a barrel hadronic calorimeter with insufficient absorption (radiation length of 7.2λ at $\eta = 0$ for all calorimeter layers including the crystals, compared to 9.7λ for ATLAS) before the coil. A tail catcher had to be added around the CMS coil to complete the hadronic shower reconstruction and provide better protection against punch-through to the muon system, faking muons. The reduced sampling fraction of CMS versus ATLAS leads to an approximately twice worse jet resolution of $100\%/\sqrt{E}$ for CMS, and a worse constant term of up to 8% in the barrel. It similarly affects the missing transverse energy resolution. Energy flow algorithms, attempting to replace charged hadrons in the shower by the corresponding measurement in the inner tracker, improve the energy resolution for hadrons and jets, in particular at low energies.

Hermeticity of the detectors for an excellent missing transverse energy measurement, but also to tag forward jets occurring, for example, in weak boson fusion processes, requires calorimeter coverage up to the very forward direction. The **forward calorimeters** of ATLAS and CMS extend the energy measurement to pseudorapidities of 5 (polar angle of 0.77 degrees). They are located in different parts of the detector. The ATLAS forward calorimeter, made of copper and tungsten absorbers with gaps filled with liquid argon, is fully integrated into the cryostat that houses the end-cap calorimeters, which reduces the neutron fluence in the muon system and, with careful design, has minimal impact on the neutron fluence in the inner tracker. The CMS forward calorimeter, made out of steel and quartz fibres and operating with Cherenkov light, is situated 11 m from the interaction point, thereby minimising the amount of radiation and charge density during operation.

Driven by the design of the magnets, the **muon systems** strongly differ between ATLAS and CMS. While CMS measures muons within the instrumented flux return, requiring the extrapolation of the track into the inner tracker, ATLAS has standalone muon tracking inside the large area spanned by the air-core toroids. Both experiments use drift tubes and cathode strip chambers (forward direction) for the precision muon measurements, and fast resistive plate chambers (thin gap chambers in the ATLAS endcaps) for fast muon Level-1 trigger signals. The pseudorapidity coverage amounts to $|\eta| < 2.7$ (2.4) for ATLAS (CMS) for muon measurements, lowering by 0.3 units for triggering. The combined momentum resolution for a 100 GeV (1 TeV) track at $\eta = 0$, reconstructed in the inner tracker and muon systems, is $\sigma(p_T) \simeq 2.6\%$ (10.4%) (ATLAS) and $\sigma(p_T) \simeq 1.2\%$ (4.5%) (CMS). The resolution significantly deteriorates in CMS for forward muons due to the reduced bending power of the solenoid (6 T.m at $|\eta| = 2.5$ compared to 16 T.m at $\eta = 0$).

Apart from these main detector systems, both ATLAS and CMS have dedicated luminosity detectors in their forward regions.

In summary, we may recall that ATLAS has put emphasis on excellent jet and missing transverse energy resolution, particle identification, and standalone muon measurement, while CMS has prioritised excellent electron, photon and tracking (muon) resolution. Both detectors have good hermeticity (very few 'cracks').

References [1, 2] present the essential performance parameters of the ATLAS and CMS experiments, sub-divided into track reconstruction, muon, electron and photon identification and reconstruction, jet and hadronic tau reconstruction, b-flavour tagging and the trigger selection (see below). Many of the results given are supported by existing test beam and cosmic ray measurements (also discussed in

these lecture notes), in particular for the single-particle response of the detector elements to electrons, photons, pions and muons at various benchmark energies. Others rely on the simulation of the detector response and the underlying physics processes. They are affected by numerous uncertainties also due to hard-to-quantify soft-QCD and machine background effects.

3.2 Trigger and data acquisition

In former times, when particle physics experiments used bubble and cloud chamber techniques, data acquisition (DAQ) was made by means of stereo photographs. There was effectively no trigger. Instead, each bubble expansion was photographed based on the constant (and known) accelerator cycle. The high-level trigger was *human*, realised by scanning teams operating worldwide with varying trigger efficiencies (rumours claim that physicists had the worst scanning efficiency). The slow operation rate of this setup allowed one to measure only the most common processes. Later, electronic signals were used to trigger the camera to photograph an event (a single trigger level). The dead time occurred while the film advanced after a trigger.

The trigger [16] is a function of the fast detector response to a collision event providing a binary accept or reject signal. Its task is to look at (almost) all bunch crossings and select the most interesting ones. Data acquisition (DAQ) collects all detector information and stores it for offline analysis. Requirements for a DAQ system are the provision of online services, such as a state machine ('Run Control'), governing the run sequences, and data quality monitoring. It must keep records of the detector configuration and run conditions, avoid corruption or loss of data (and hence verify the data sanity), be robust against imperfections in the detector and associated electronics and readout systems, and minimise dead time.[13] Because the trigger latency even for the fastest level is longer than the 25 ns bunch crossing period, the electronics signals need to be saved locally in so-called pipelines until the trigger signal arrives.

A problem for any trigger at the LHC is that one cannot (and does not want to) save all events. 'Old' (known) physics occurs more often than 'new' physics, i.e., the new physics is buried under huge amounts of old physics. We have seen that the interesting physics occurs at rates of 10 Hz (for top antitop production) and below at highest peak luminosity. The remit is thus to keep *all* of those events, while rejecting most of the others. One exception to this is low-mass flavour physics, which — although being 'old' — has still important potential for discoveries. We hence must aim at fitting the best possible physics cocktail into the available bandwidth. Efficient selection and background rejection requires one to include the response of the entire detector in the trigger decision. This can only be achieved by splitting the trigger decision into several levels with increasing complexity. The first level has short latency and high efficiency and must only aim at the rejection of the 'obviously' uninteresting events (once rejected, events are rejected forever!). Later levels, which can be slower thanks to the rejection in the previous level, perform fine-grained selection and rejection.

The trigger systems of ATLAS and CMS are separated into a first-level ultra-fast hardware trigger, based on information from the calorimeters and dedicated muon systems only. The detector data are transferred to large buffer memories after a Level-1 accept. The data rates to DAQ and the next level triggers are massive: with approximately 1 MB event size at 100 kHz event rate one has a rate of 100 GB/s (i.e., 800 Gbit/s). The subsequent high-level trigger (HLT) uses partial event data readout or powerful network switches to feed reconstruction and software selection algorithms running on farms

[13] Dead time is the fraction of time where valid interactions could not be recorded for various reasons. Typical system-imminent dead time is of the order of up to 10%. It originates from the readout and trigger system, from operational dead time (e.g., the time to start and stop a run or to configure the detector systems), trigger or DAQ down-time (e.g., following computer failure), or detector down-time (e.g., following a high-voltage trip). For a multi-level trigger, the total dead time is the sum of the dead times of all levels. The trigger dead time for a given level is computed from the product of the trigger rate of the previous level and the latency for this level. The readout dead time is given by the product of the final (highest-level) trigger rate and the local readout time. Note that trigger dead-time logic is *required* to prevent triggering another event before the detector has been fully read out. Given the investment in the accelerators and the detectors for a modern HEP experiment, it is clearly important to keep dead time to a minimum.

with several thousand central processing units. In ATLAS the HLT is separated into two independent steps. A fast Level-2 trigger using only detector information from so-called 'regions of interest', which are sections along azimuthal and pseudorapidity cuts around triggered Level-1 objects, and including only the detector systems required by the Level-2 algorithm. The Level-2 decision must come within a few milliseconds and reduce the outgoing Level-1 rate from 75 kHz to 2 kHz, which is the input rate to the event builder requiring to read out the full detector. A subsequent Level-3 trigger ('Event Filter') then further reduces the event rate to approximately 200 Hz, which is written to disk. These events are promptly reconstructed at CERN and, in parallel, distributed to 10 worldwide computing centres. In CMS, the large HLT input rate is tamed by factorising the event building into a number of slices each of which sees only a fraction of the rate. This requires a large and expensive total network bandwidth, but avoids the need for a very large single network switch.

An important requirement for the event building is a proper timing-in of the various detectors. Indeed, within the 40 MHz bunch crossing rate, particles can only travel 7.5 m through the detectors, which are significantly larger than that (ATLAS has a height of 2×11 m and a length of 2×23 m). In addition, the collection of the detector signals, notably in the large muon drift tubes, can take up to 40 bunch crossings (1 μs). To properly collect the signals belonging to the same bunch crossing (i.e., 'event') and to keep the exposure time per event as small as possible, trigger-decision and detector response collection delays must be aligned to a few nanoseconds. Timing-in is one of the first commissioning tasks for all detector systems.

4 Detector commissioning — Overview

All detector systems, as well as the performance and physics groups developed detailed commissioning strategies for initial running with colliding beams. Even before beams collide in the LHC, as more and more systems are being installed, extensive stand-alone and combined studies with comic ray events and detector calibrations are performed. These studies as well as dress rehearsals using simulated data exercise the full data acquisition chains, including the online and offline data quality assessment tools, and the streaming of the events into several physics and calibration streams based on the trigger decision.

The cosmic ray data provide important information to align the detectors relative to each other (but not relative to the beam axes). They set an initial reference geometry for most of the barrel muon detector, and will be used to correct the alignment based on precise survey data and optical sensors. Muons from beam halo data taken during single-beam LHC commissioning runs will be used as an initial validation of the end-cap muon detector alignment. For example, in ATLAS the magnetic field strengths of the toroids, determining the muon energy scale, are known to better than 0.5% versus ϕ from survey data of the measured coil positions. Later the precision can be improved to 0.1–0.2%, using a system of Hall probes. The field of the solenoid immersing the inner detector has been mapped to a precision of a few Gauss, which approaches the design goal.

Charge injection or pulsed calibrations of the electronic boards and pedestal runs provide initial settings for channel thresholds, ramp and delay values, pedestals, etc. for the various systems, and are used to map noisy and to some extent dead channels. Hadronic calorimeters also perform calibration with laser-light and radioactive caesium sources. These tasks together with test beam measurements contribute to achieving a sufficient quality of the first collision data.

As an example, the ATLAS operational status as of autumn 2009 is given in Table 1. The experiment's start-up and ultimate design goals in terms of the tracking and calorimeter performance are summarised in Table 2.

With the start-up of the LHC,[14] and after timing-in the detector systems with the colliding LHC bunches and the trigger signal, minimum bias triggers from scintillator counters will provide Level-1

[14] All event numbers given in this overview section refer to 14 TeV LHC centre-of-mass energy. The impact from lower centre-of-mass energies (10 TeV and 7 TeV) is briefly discussed in Section 11.

Table 1: Number of channels and operational status as of autumn 2009 of the ATLAS subdetectors.

ATLAS subdetector	Number of channels	Operational fraction (%)
Pixel Tracker	80 million	97.9
Silicon Strip Tracker	6.3 million	99.3
Transition Radiation Tracker	350 000	98.2
Liquid-Argon Electromagnetic Calorimeter	170 000	98.8
Tile Hadronic (Extended) Barrel Calorimeter	9800	99.2
Hadronic Endcap Liquid-Argon Calorimeter	5600	99.9
Forward Liquid-Argon Calorimeter	3500	100
Muon Drift Tubes	350 000	99.7
Muon Cathode Strip Chambers	31 000	98.4
Barrel Muon Trigger	370 000	98.5
Endcap Muon Trigger	320 000	99.4
Level-1 Calorimeter trigger	7160	99.8

accepts for initial physics studies at a luminosity less than or equal to 10^{31} cm^{-2}s^{-1}. These events can be used to provide first occupancy tests of the inner tracking systems, and to refine the dead channel maps. Copious isolated tracks from minimum bias events will allow the experiments to refine the inner detector alignment using the distributions of residuals between measured hits and fitted tracks, and the comparison of E/p for pions of opposite charge. Alignment monitoring information will also be derived from K^0_S and Λ invariant mass and azimuthal decay vertex distributions. The K^0_S invariant mass together with the known, ideally uniform decay-angle distribution can be used for a data-driven determination of the tracking efficiency. In ATLAS, high and low threshold transition radiation hits from isolated pion tracks will be compared to the expectation from simulation. Minimum bias events will help both experiments to monitor the uniformity of the calorimeter response, which can be done azimuthally and by comparing positive and negative pseudo-rapidity regions. In this initial phase it will also be possible to some extent to validate the calorimeter simulation by comparing shower shapes for isolated hadronic tracks and low energetic jets. The statistics corresponding to a few days of low-luminosity data taking without toroid fields will allow the collection of enough straight muon tracks to calibrate the ATLAS muon optical alignment system to better than 100 μm. It will be improved to up to 30 μm at higher luminosity, which is required to take full advantage of the spatial resolution of 40 μm per muon chamber, providing a 10% measurement of 1 TeV muon tracks.

While the trigger system is being commissioned, simple inclusive Level-1 calorimeter and muon triggers will be included first, followed by more complex Level-1 triggers, involving, for example, isolation and missing transverse energy. At the same time, the HLT systems will begin to operate, initially in pass-through mode, allowing the experiments to test the algorithms, and later using the full power of the HLT, while continuing to run pre-scaled triggers in pass-through mode. Combinations of pre-scaled multi-threshold triggers will be used to determine efficiency curves for the three trigger levels (so-called 'bootstrapping method'). The data collected with the complete low-luminosity trigger menu will contain copious quantities of low-energy leptons from heavy quark decays and also from direct J/ψ and Υ production. Approximately 5000 $W \to \mu\nu$ and 500 $Z \to \mu\mu$ decays should be reconstructed per 1 pb^{-1} of integrated luminosity (the expected rates are somewhat lower for electrons). The low-luminosity trigger menu will also provide abundant samples of high-p_T jets, prompt photons mainly from γ-jet events, and semileptonic τ decays.

Table 2: Expected calibration and alignment accuracies at the LHC start-up and the ultimate design goals for the ATLAS experiment. Examples for physics channels or measurements driving the requirements are given in the last column.

	Start-up of LHC	**Ultimate goal**	**Physics goals**
EM energy uniformity	1–2%	0.7%	$H \to \gamma\gamma$
Electron energy scale	∼2%	0.02%	W mass
Hadronic energy uniformity	2–3%	$< 1\%$	Missing E_T
Jet energy scale	$< 10\%$	1%	Top-quark mass
Inner detector alignment	50–100 μm	$<10\,\mu$m	b tagging
Muon spectrometer alignment	$<200_{\text{barrel}}\,\mu$m	30 μm	$Z' \to \mu\mu$
Muon momentum scale	∼1%	0.02%	W mass

All these events will be crucial for the initial validation of the detector performance. More specifically, the inner detector material can be mapped with photon-to-e^+e^- conversions to order 1% with the statistics available after a few months of data taking. This procedure can be validated by studying the momentum dependence of the reconstructed invariant masses of low-mass resonances. Inclusive electrons can be used to test bremsstrahlung recovery in the inner detector. The inner detector alignment is expected to converge to the relative design accuracy of approximately 10 μm soon after the full detector commissioning has started (the alignment with cosmic ray events will be insufficient in the endcaps), allowing the constant term in the tracking resolution to be below 20% of the full resolution. Local inner detector misalignment can be studied with the use of resonances with known masses and lifetimes decaying to lepton pairs, and with high-p_T muons in combined track fits with the muon spectrometer.

Preliminary electromagnetic inter-calibration can be obtained at low luminosity using the azimuthal and $\pm\eta$ symmetry of inclusive isolated electrons from various sources. It is, however, not clear whether this procedure improves the intrinsic electromagnetic calorimeter inter-calibration determined in test beams at the higher energy scales of interest for most of the physics analyses (it will be useful for CMS where only 9 out of 36 supermodules of the electromagnetic calorimeter could be calibrated in the H4 test beam, see Section 5). The ultimate high-energy electromagnetic inter-calibration will use $Z \to ee$ events, requiring about 100 pb^{-1} recorded integrated luminosity to significantly improve the expected initial uniformity of 1–2% to a statistical precision of ∼0.7% (ATLAS) with high granularity, provided the inner detector material is well enough understood. These events will also serve to calibrate the global electromagnetic energy scale.

Hadronic track and jet inter-calibration will employ E/p measurements (assuming an aligned inner tracker) and E_T balancing in di-jet, γ-jet and also Z-jet events, versus ϕ. The latter two channels also determine the global jet energy scale with an expected precision better than 5% after a few months of data taking. Di-jet events will also be used to validate the forward E_T scale and resolution. The expected number of ∼500 fully reconstructed $t\bar{t}$ events for 100 pb^{-1} with one W decaying hadronically and the other one leptonically (electron or muon) allows a first calibration of narrow jets with invariant mass fits to $W \to qq'$ decays. It will also be important to study the stability of the electromagnetic and hadronic cluster reconstruction with respect to varying calorimeter noise (significant event pileup is expected to occur only above peak luminosities of $\mathcal{O}(10^{33}\,\text{cm}^{-2}\text{s}^{-1})$ for the nominal LHC bunch pattern scheme).

The performance of heavy-flavour jet tagging crucially relies on locally aligned silicon detectors. Flavour tagging will be calibrated using $t\bar{t}$ events, but initially also using orthogonal information from tagging algorithms based on track fits and soft-muon reconstruction in di-jet events.

One of the most difficult detector observable to measure accurately is missing transverse energy

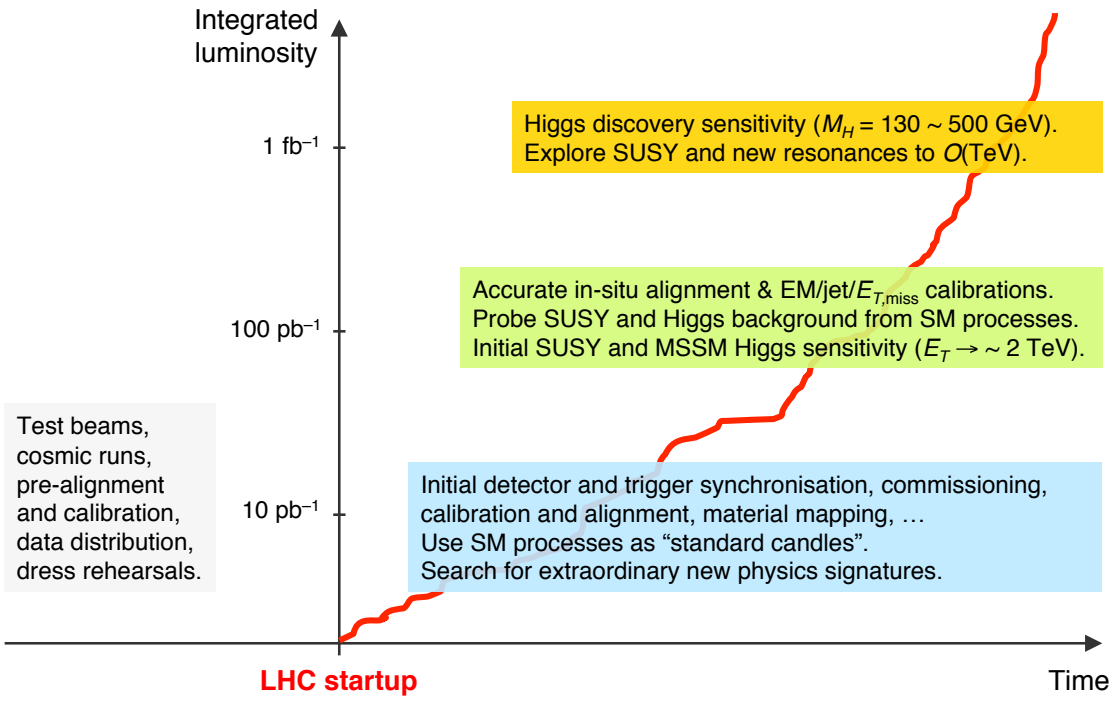


Fig. 5: Sketch for a commissioning and early physics roadmap at the LHC.

($\not\!\!E_T$). Because it is sensitive to many new physics signatures, the tails of its distribution, dominated by resolution and instrumental effects, must be precisely calibrated with data before they can be used for discrimination and reconstruction purposes. The computation of $\not\!\!E_T$ requires the cleaning of the event from beam halo muons, beam gas collisions, cavern background, and cosmic rays. Moreover, the calorimeter cells must be calibrated (for both electromagnetic and hadronic showers), and deficient calorimeter cells (including noise) must be mapped and corrected. Initial data-driven $\not\!\!E_T$ studies will use minimum bias events, analysing the $\not\!\!E_T$ resolution as a function of the E_T sum and comparing it with the expectation from simulated data, the transverse W mass in $W \rightarrow e(\mu)\nu$ events, $Z \rightarrow ee(\mu\mu)$ events, and, with rising statistics, mass-constrained $t\bar{t}$ and $Z \rightarrow \tau\tau$ events decaying to charged leptons and hadrons (approximately 7000 of the latter events with $p_T(\mu) > 15\,\mathrm{GeV}$ are expected in 100 pb^{-1}, allowing one to calibrate the absolute $\not\!\!E_T$ scale to about 5%).

For muon tracks, the correlation of muon spectrometer and inner detector provides powerful reconstruction cross-checks for both systems. The muon reconstruction efficiency for stand-alone (muon spectrometer or inner detector only) and combined tracks can be determined with $Z \rightarrow \mu\mu$ events by reconstructing one muon and probing the reconstruction of the other one ('tag-and-probe method'). The muon fake rate, expected to be negligible at low luminosity, will become significant above $10^{33}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$, due to the neutron and photon background in the cavern. The fake rate concentrates, however, at very low p_T, and remains small enough so that the impact on most physics analyses should be negligible. The overall muon energy scale will be calibrated with $Z \rightarrow \mu\mu$ events, where a statistical precision of 0.8% and reasonable geometrical granularity can be reached with 100 pb^{-1} integrated luminosity. With more data available, local misalignment problems in towers of chamber triplets could also be resolved with Z-mass constraints. A sketch for the commissioning and early physics roadmap at the LHC is displayed in Fig. 5.

Initial physics measurements will primarily focus on Standard Model processes with high cross-sections. Among these are the multiplicity and pseudo-rapidity distribution of minimum bias events and cross sections of events with jets. Low-p_T physics mainly dedicated to the study of B_s decays will

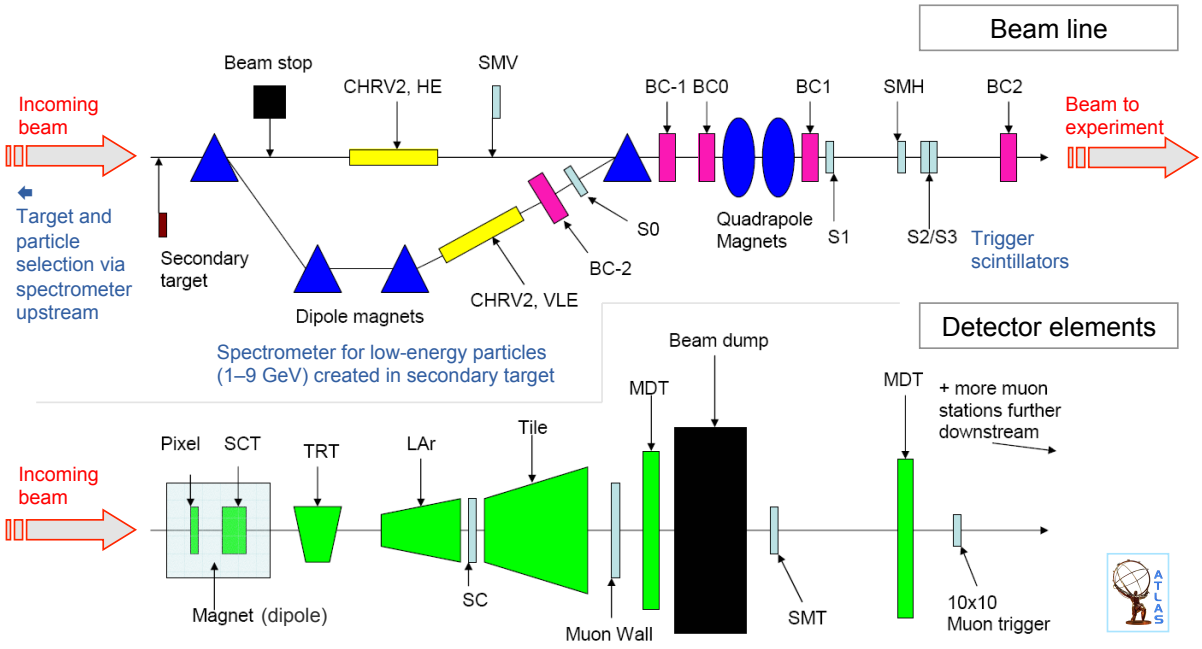


Fig. 6: H8 beam line of the ATLAS combined test beam 2004. Protons from the SPS, accelerated to 450 GeV energy, hit a target producing hadrons, electrons and muons with energies in the range of 1 to 350 GeV, which are selected upstream by a mass spectrometer. The composition of the incoming monochromatic particle beam is measured with Cherenkov counters (upper picture). The beam is focused and passes trigger scintillators before entering a complete ATLAS barrel slice (lower picture) with realistic geometry composed of Pixel and silicon strip detector (SCT) layers, immersed in a 1.4 T magnetic dipole field parallel to the beam, a transition radiation tracker (TRT) module outside the magnetic field, liquid-argon electromagnetic and tile hadronic calorimeter layers, interleaved with a scintillator to measure the energy lost in the liquid argon cryostat, and a series of muon drift tube and resistive plate chambers before and after a beam dump block.

begin by measuring J/ψ to Υ cross section ratios, which involves the validation of vertexing tools, and cross sections and lifetimes of B, B_s and Λ_b mesons using decays to J/ψ. Statistically competitive lifetime measurements for these mesons can be expected with $\sim$100 pb^{-1} integrated luminosity. The cross section of $t\bar{t}$ production using semileptonic decays can be measured to a precision better than 20% with 100 pb^{-1} integrated luminosity, without requiring b tagging. Moreover, a significant single-top signal is expected to be seen in this data sample. Analyses aiming at searches for new phenomena will initially concentrate on the understanding of the detector performance and Standard Model processes, using calibration channels and studying phase space areas where new physics contamination is expected to be small.

The subsequent sections describe in some detail several of the commissioning and early physics studies mentioned above.

5 Commissioning with test beams

Both ATLAS and CMS have performed series of measurements with test beams of known energies and particle types. Electrons, photons, muons, pions, protons with energies between 1 and 350 GeV and varying magnetic field configurations were collected to test the tracking efficiency, alignment and particle identification, (inter-)calibrate the electromagnetic and hadronic calorimeters, test the muon trigger efficiency, tune Monte Carlo simulation, etc.

Figure 6 shows a sketch of the 'H8' beam line used for the ATLAS 2004 combined test beam.

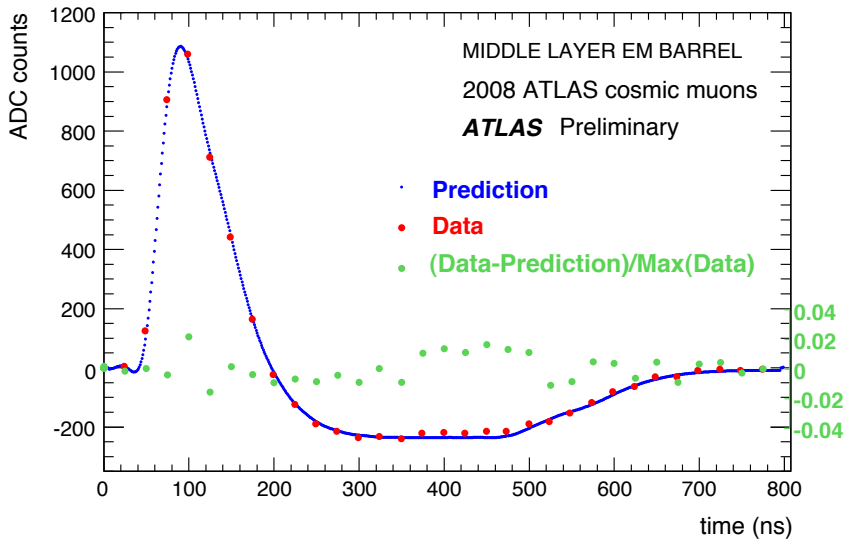


Fig. 7: Digitised bipolar ionisation pulse shape of a 15 GeV cosmic-ray signal measured in the middle layer of the ATLAS electromagnetic calorimeter. The signal is shaped and sampled with 40 MHz frequency, corresponding to a sample period of 25 ns, and a total sampling window of 800 ns (during normal data-taking only 5 samples (125 ns) are read out). The study of the pulses measured with 32-sample readout allows one to determine the drift time in the liquid argon gaps related to the undershoot of the pulse, and the electrode position related to the rise at the end of the pulse. The curve shows the expectation agreeing to better than 2% with the measurement.

A full barrel slice, from the innermost tracking detectors and magnetic field to the outermost muon spectrometer, was exposed to the particle beams. The experimental setup was kept as close as technically possible to the ATLAS geometry. The distance between subdetectors, the pointing geometry, and the magnetic field orientation were preserved where permitted. The most important goals of this test beam campaign were: (i) test the detector performance with final or close to final electronics equipment, data acquisition and trigger infrastructure and reconstruction software, (ii) validate the description of the data by Monte Carlo simulations down to energies of 1 GeV to prepare the simulation of the ATLAS data, and (iii) perform combined studies in a setup very close to that of ATLAS (e.g., combined electromagnetic and hadronic calorimetry, and combined tracking and calorimetry).

5.1 Energy reconstruction in the ATLAS liquid-argon electromagnetic calorimeter

The ionisation signal generated in the ATLAS electromagnetic calorimeter is collected from the readout electrodes and brought via cables to the front-end electronics where it is amplified, shaped and sampled at a 40 MHz frequency. The samples (usually five) are stored in an analog pipeline until the arrival of a trigger accept decision. The samples belonging to the accepted event are then digitised and transmitted by the calorimeter backend electronics to readout driver modules, where the signal amplitude is reconstructed and converted to MeV. Figure 7 shows a fully digitised pulse shape with 32 samples from a cosmic-ray event with an unusually large energy deposit. The full cell-energy reconstruction from the digitised pulse samples is encoded in the following conversion formula

$$E_{\mathrm{cell}} = F_{\mu\mathrm{A}\to\mathrm{MeV}} \cdot F_{\mathrm{DAC}\to\mu\mathrm{A}} \cdot \left(\frac{M_{\mathrm{phys}}}{M_{\mathrm{calib}}}\right)^{-1} \cdot R\left(\sum_{i=1}^{N_{\mathrm{samples}}} a_i\,(s_i - p)\right), \tag{3}$$

where the subscripts specify the conversion type. The sum over the digitised samples on the right-hand side is computed from the measured ADC counts, corrected for an overall pedestal (p), obtained in regular calibration runs — together with noise and autocorrelation terms — from random triggers in physics

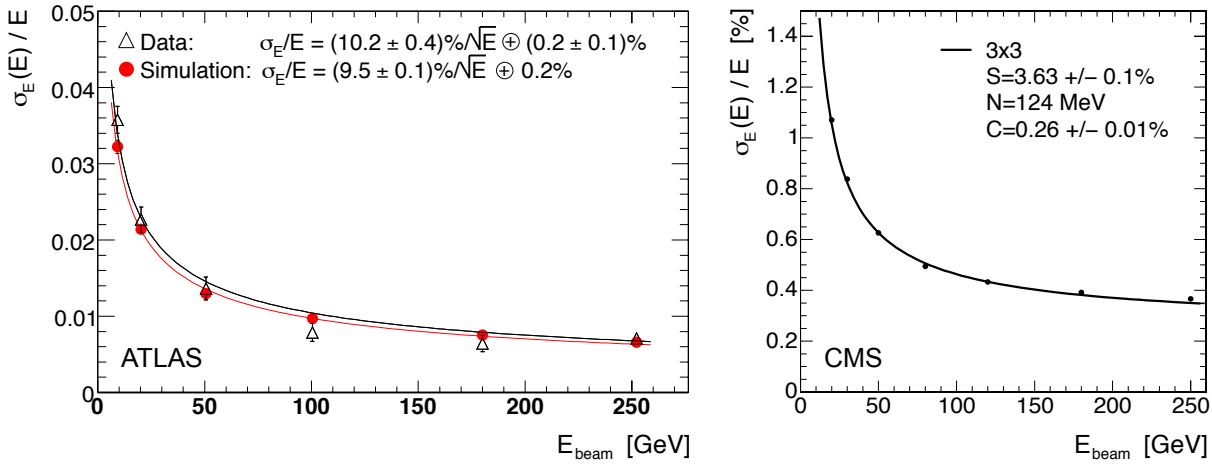


Fig. 8: Fractional electromagnetic energy resolution versus the incident energy obtained from electron test beams. Left: result for the ATLAS liquid-argon calorimeter obtained, behind $1.6X_0$ material thickness, from the H8 combined test beam in 2004. Right: result for the CMS crystal calorimeter obtained without upstream material for 9 out of 36 tested supermodules at the H4 test beam in 2006. The energy was measured in an array of 3×3 crystals with electrons impacting the central crystal.

events, and multiplied by the sample-specific 'optimal filtering coefficients' a_i, obtained from so-called 'delay' runs where calibration signals are injected to measure the pulse shape. The sum is taken as an argument to the ADC-to-DAC ramp function R, obtained from dedicated electronics calibration runs, where known charges are injected and the corresponding ADC output is measured and fit to a linear function. Differences between the calibration and physics pulse shapes are corrected via the M ratio. The DAC values are then converted to μA, which is related to the calibration injection resistance and computed taking into account cable and other attenuation effects. Finally, the μA signal is converted to MeV by applying the corresponding current-to-energy conversion factor, and by correcting the energy lost in the absorber material (sampling fraction). Once the cell energies are reconstructed, cells are summed to form a cluster over all three longitudinal compartments and the presampler of the electromagnetic calorimeter.

This procedure provides the electromagnetic energy scale. Physics events such as $Z \to ee$ will be used to achieve absolute energy calibration. For hadrons and jets, one needs to account for hadronic shower corrections, that is, one must pass from the electromagnetic to the hadronic energy scale.

5.2 Electromagnetic energy resolution

The resolution of an electromagnetic calorimeter is driven by the amount of active material in which the electromagnetic shower develops, and by the shower containment. Containment requires a calorimeter thickness of many radiation lengths $X/X_0 > 20$, where the radiation length X_0 is a material characteristic related to the energy loss of high-energy particles interacting electromagnetically with the material.[15] Test beams with known particle content and energy allow the experiments to measure resolution, linearity and uniformity of the calorimeter energy response. The resolution results obtained by ATLAS and CMS for electron beams with different energies are shown in Fig. 8 (the measurements were obtained under different experimental conditions, see figure caption). Calorimeter resolution is conveniently expressed as a function of the incident electron/photon energy, E, by the expression

[15]The radiation length is both the mean distance over which a high-energy electron loses all but $1/e$ of its energy by bremsstrahlung, and 7/9 of the mean free path for pair production by a high-energy photon.

$$\frac{\sigma(E)}{E} = \frac{S}{\sqrt{E\ (\mathrm{GeV})}} \oplus C \oplus \frac{N}{E\ (\mathrm{GeV})}\,, \tag{4}$$

where the first term on the right-hand side determines the *stochastic* resolution resulting from statistical fluctuations in the number of shower particles[16] and in the shower containment, the second *constant* term is due to non-uniformities in the calorimeter response introduced by inhomogeneities and non-linearities, and the third *noise* term quantifies electronics noise and in-time physics pile-up. The '$\oplus$' indicates that the different resolution terms are added in quadrature. Some numbers obtained for these terms from fits to electron test beam data are quoted on the plots in Fig. 8. Taking into account the full detectors and materials, one expects for ATLAS (CMS) the following benchmark resolution parameters: $S = 10$–12% (3–5.5%), $C = 0.2$–0.35% (0.5%), $N = 250$ MeV (200–600 MeV), where the better (worse) numbers refer to the barrel (endcaps).[17] With the 9 out of 36 super-modules calibrated in the 2006 test beam, CMS also found excellent energy-response uniformity of 0.27%.

5.3 Hadronic energy resolution

During the ATLAS H8 combined test beam campaign, pion beams with 6 discrete energies ranging from 10 GeV to 350 GeV were used to study the hadronic energy reconstruction in the calorimeters. Hadron showers originate from interactions of hadrons with nuclei. The density of hadron calorimeters is therefore appropriately expressed in terms of the nuclear interaction length λ, which quantifies the mean free path of hadrons in material between strong collisions. For example, silicon has $\lambda = 45.5$ cm, iron 16.8 cm, lead 17.1 cm, and water 83.6 cm, to be compared to $X_0 = 0.56$ cm for lead and 1.76 cm for iron. Hence $\lambda \gg X_0$ and one can separate electromagnetic showers, which are short-ranged, from far-ranged hadronic showers, which also clarifies why calorimeters are called and arranged as they are: electromagnetic calorimeters fully absorb electromagnetic showers, but only parts of the showers initiated by hadrons; the following calorimeter layers (usually sampling calorimeters) entirely absorb the hadronic showers.

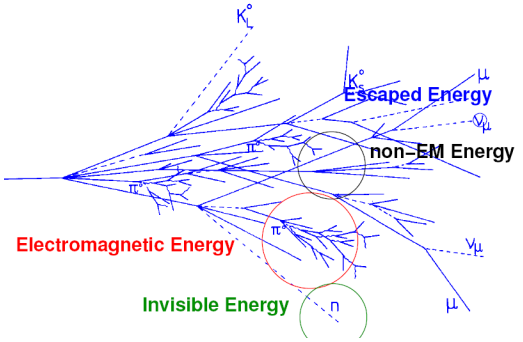


Fig. 9: Simulated hadron shower consisting of electromagnetic and non-electromagnetic, invisible and escaped energy.

Hadronic showers (Fig. 9) consist of approximately 50% electromagnetic energy (e.g., $\pi^0 \to \gamma\gamma$), 25% non-electromagnetic energy (such as dE/dx from $\pi^\pm$, $\mu^\pm$, $K^\pm$), another 25% invisible energy (nuclear fission and excitation, neutrons), and 2% escaped energy (e.g. neutrinos). Invisible and escaped

[16] The number of particles produced in the shower is proportional to the energy of the incident particle: $N_{\mathrm{part}} \propto E$. The error in the energy measurement is due to statistical fluctuations in N_{part}, i.e., $\sigma(E) \propto \sqrt{N_{\mathrm{part}}}$. One thus finds for the stochastic contribution to the energy resolution $\sigma(E)/E \propto 1/\sqrt{E}$. Because in sampling calorimeters the absorber material does not contribute to the energy measurement, the electromagnetic energy resolution is worse than for crystal calorimeters, provided that the crystals have sufficiently large X/X_0 so that the full shower can be contained. This is the case for the $PbWO_4$ scintillating crystals used by CMS, which have very high density so that the total calorimeter has $28X_0$ (for comparison, the ATLAS calorimeter has $22X_0$). The sampling fractions in the ATLAS electromagnetic calorimeter are $f_{\mathrm{sampl}} = 0.17$ (0.20) for $|\eta| \leq 0.8$ ($|\eta| > 0.8$). The measured energy must thus be corrected for the dead material $E_{\mathrm{true}} = f_{\mathrm{sampl}}^{-1} E_{\mathrm{meas}}$, so that the stochastic resolution becomes $\sigma(E)/E \propto \sqrt{d_{\mathrm{sampl}}/f_{\mathrm{sampl}}}/\sqrt{E} \approx 3/\sqrt{E}$, where d_{sampl} is the thickness of the sampling layers (finer sampling provides better resolution). Hence the approximately three times worse intrinsic electromagnetic energy resolution in ATLAS compared to CMS.

[17] With these parameters, a back-of-the-envelope calculation for $H \to \gamma\gamma$ gives for the di-photon mass resolution as a function of the photon energy: $\sigma(M_{\gamma\gamma})|_{E_\gamma} \propto M_H \sigma(E_\gamma)/(\sqrt{2}E_\gamma) \approx 1.2\,\mathrm{GeV}$ (0.7 GeV), for ATLAS (CMS) and where $M_H = 120$ GeV has been assumed. To obtain a realistic estimate of the resolution one must also include the error on the opening angle (photon directions), as well as $\gamma \to e^+e^-$ conversions (20–60% of all photons from $H \to \gamma\gamma$ decays, strongly increasing for large $|\eta|$). Both effects reduce the effective resolution difference between the experiments.

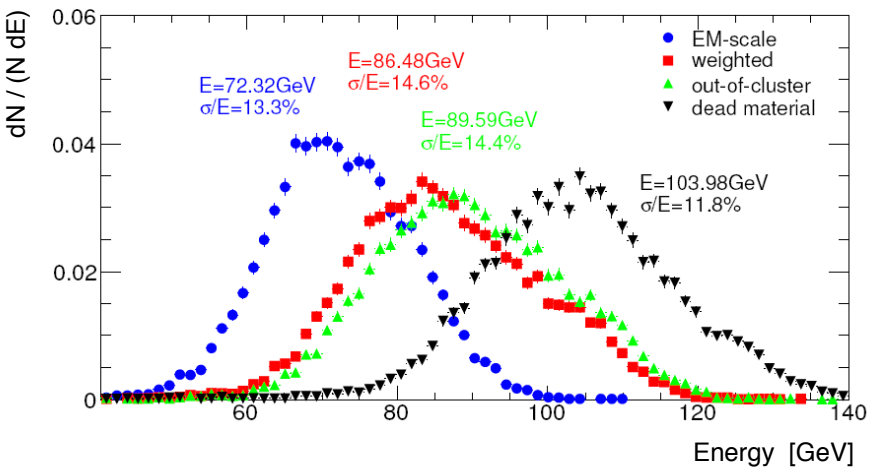


Fig. 10: Reconstructed energy for 100 GeV test beam pions in a slice of the ATLAS barrel electromagnetic and hadronic calorimeters. Shown are: raw measured energy (circles), after reweighting from the electromagnetic to the hadronic scale (squares), after applying out-of-cluster corrections from shower leakage (top-oriented triangles), and after dead-material corrections (bottom-oriented triangles).

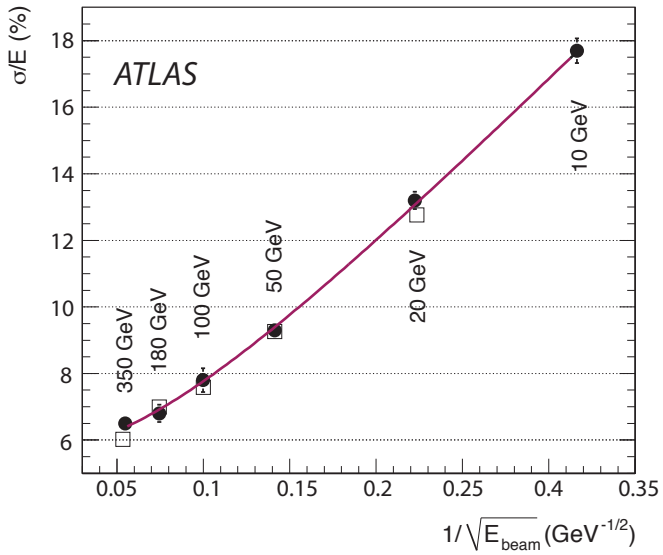


Fig. 11: Fractional energy resolution for pions at 0.35 pseudorapidity (equivalent calorimeter depth 7.9λ), versus the incident energy from test beam data in the ATLAS hadronic calorimeter (full circles), and compared to Monte Carlo simulation (open squares).

energy causes worse resolution for hadronic showers than for electromagnetic ones. When uncorrected it also causes an underestimate in the measured energy with respect to the true hadron energy. Figure 10 shows the reconstructed energy in the ATLAS barrel calorimeter slice for 100 GeV pions from test beams. The raw measured energy at the electromagnetic scale undershoots by 28% with the largest contributions to the bias coming from invisible and escaped energy, and from dead material. While the various corrections recover the overall energy scale, they cannot improve the resolution (unless event-by-event corrections as a function of the longitudinal and transverse shower shapes are applied).

The final energy resolution obtained from pion test beam data for the ATLAS calorimeter is shown in Fig. 11, and compared to the expectation from Monte Carlo simulation (Geant-4). One finds benchmark values for single hadrons of 53%, 3%, and 0.5 GeV, for the stochastic, constant and

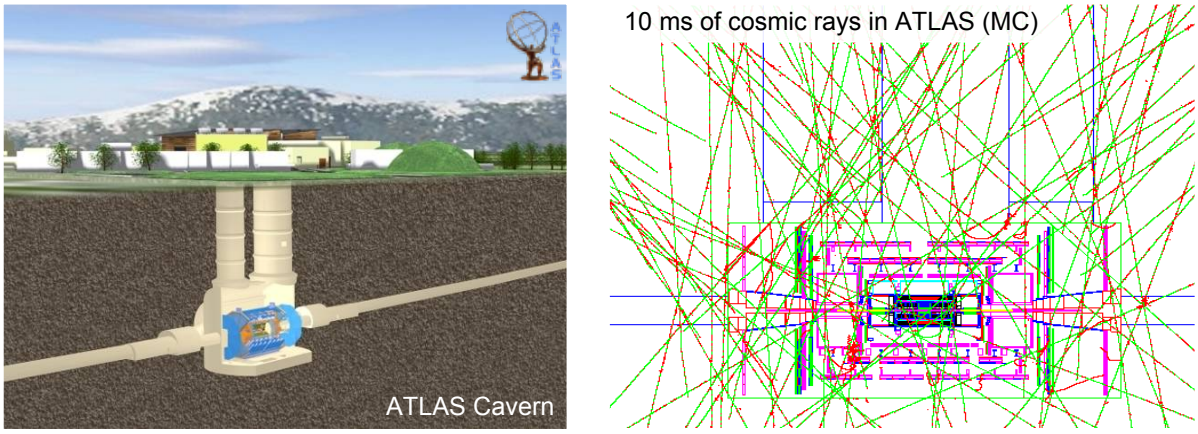


Fig. 12: Schematic drawings of the ATLAS underground cavern with supply shafts (left — two lateral elevator shafts are not drawn), and simulated cosmic rays through ATLAS within 10 ms exposure time (right).

noise terms, respectively (*cf.* Eq. 4). For comparison, for central jets Monte Carlo simulation predicts 60%, 3%, and 0.5 GeV for the resolution parameters, and a missing transverse energy resolution of $\sigma(E_T^{\text{miss}})/\sum E_T \approx 55\%$. These values are somewhat worse in CMS due to the reasons mentioned in Section 3.

6 Commissioning with cosmic rays

ATLAS and CMS have performed extensive campaigns of cosmic ray data-taking, initially with the individual systems, later including more and more detector systems with the completion of the installation in the pits. The goals of these studies are — along with exercising the detector operation, and the full data taking, reconstruction and analysis chain — tracking alignment (with and without magnetic field), deriving dead channel maps, measuring the muon trigger and tracking efficiencies, analysing calorimeter pulse shapes, improving the detector timing, tuning Monte Carlo simulation, etc.

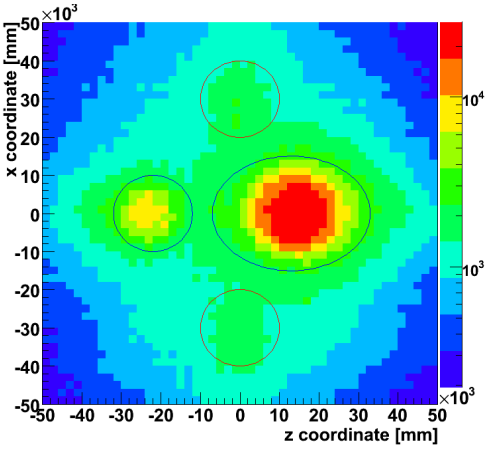


Fig. 13: Reconstructed cosmic tracks (6.6 million) in the ATLAS resistive plate chambers, extrapolated to the surface. The ellipses indicate the supply and elevator shafts.

Cosmic rays stem from cosmic nuclei (90% protons, i.e., hydrogen nuclei) that interact strongly with the Earth's atmosphere, creating hadrons — mainly pions and kaons with relative intensity 1:0.054 [17], which decay to minimum ionising relativistic muons that reach sea level on Earth,[18] or which undergo nuclear interactions with nuclei in air. The muon flux at the surface is approximately 130 Hz per m^2 for $E_\mu > 1$ GeV, and the average muon energy is about 4 GeV. The ATLAS detector, being separated from the surface by 100 m of earth and stone, receives a muon flux of approximately 4 kHz in the fiducial volume of the muon spectrometer, and 15 Hz in the TRT barrel (numbers from Monte Carlo simulation). The supply and elevator shafts (see left-hand plot of Fig. 12) provide reduced shielding, which translates into an increased occupancy of the detector elements underneath the shafts or close by

[18] Cosmic rays have been, and are still, sources of major discoveries in particle physics. For example, in 1932, Anderson (Cal Tech, USA) discovered the antielectron (positron) in cosmic rays. Later in 1946, Rochester and Butler (Manchester, England) observed two tracks 'out of nothing' in cosmic rays, which were pions from the decay of a neutral ('strange') kaon, thereby initiating particle physics. Today, very high energy cosmic rays are extensively studied.

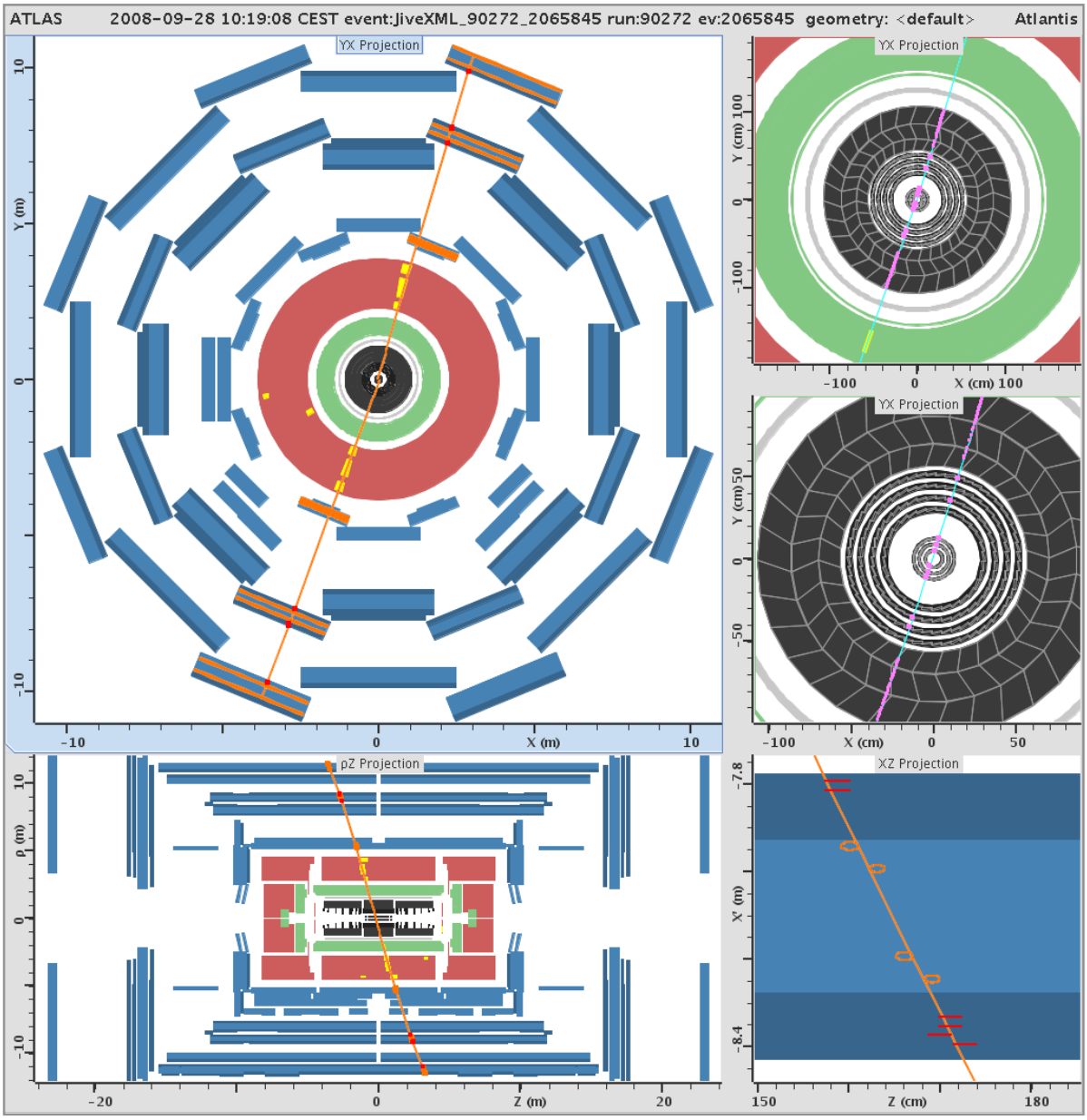


Fig. 14: A cosmic ray muon measured by ATLAS. Seen are hits in the muon spectrometer and the inner tracking systems, as well as energy deposits in the hadronic tile calorimeter. All magnets were switched off in this run.

(Fig. 13). The right-hand plot of Fig. 12 shows a simulated 10 ms snapshot of the ATLAS detector bombarded by cosmic rays. High-energy cosmic rays sometimes also produce so-called 'air showers' (and *extensive* air showers), where an avalanche of secondary scattering particles is created. Such air showers have been observed by the experiments, giving rise to events with large numbers of muons (order 10 to 100), jets, and large deposited energy (events with 6 jets, all exceeding 20 GeV transverse energy, have been seen).

Figures 14–17 show event displays of cosmic rays in ATLAS and CMS, measured with the full detectors. ATLAS accumulated 580 million combined cosmic ray events between September 13 and October 29, 2008, and in June/July and October/November 2009. CMS recorded 370 million combined events between October 13 and November 11, 2008 during the CRAFT exercise (many more cosmic ray data have been recorded by CMS during other campaigns). All events have been promptly reconstructed

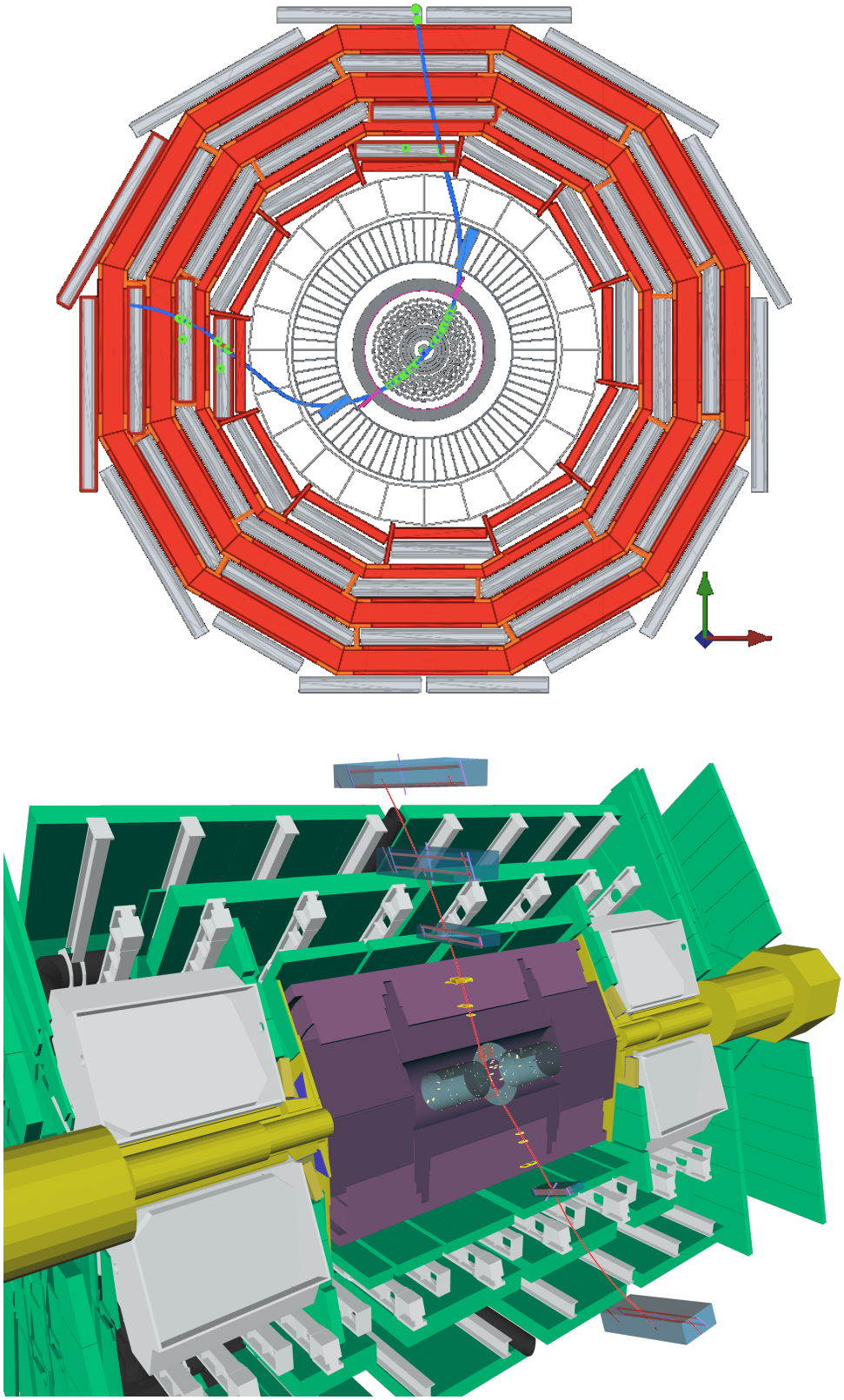


Fig. 15: Top: a cosmic ray muon measured by CMS, strongly bent in the transverse plane by the 3.8 T solenoid field. Bottom: three-dimensional view of a cosmic ray muon in ATLAS.

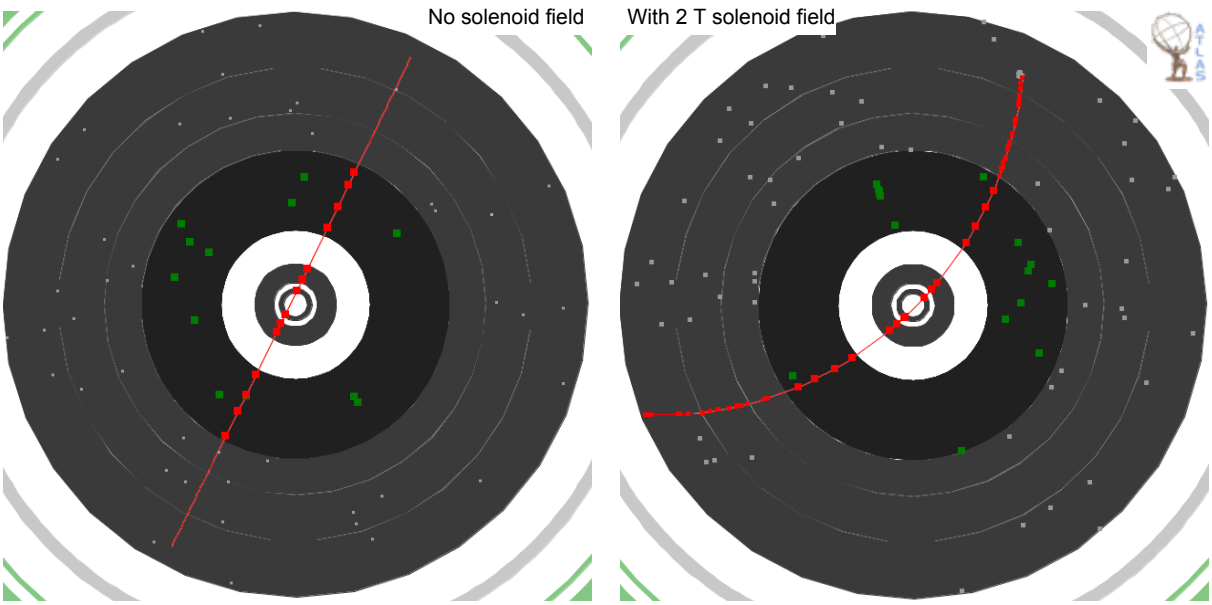


Fig. 16: Transverse views of cosmic ray tracks measured in the ATLAS pixel (the three innermost hits depicted by the dots) and silicon strip detectors (four double hits at about half radius in the event displays). The left (right) drawing shows a straight track measured with the solenoid field off (on). The right plot shows also transition radiation tracker hits.

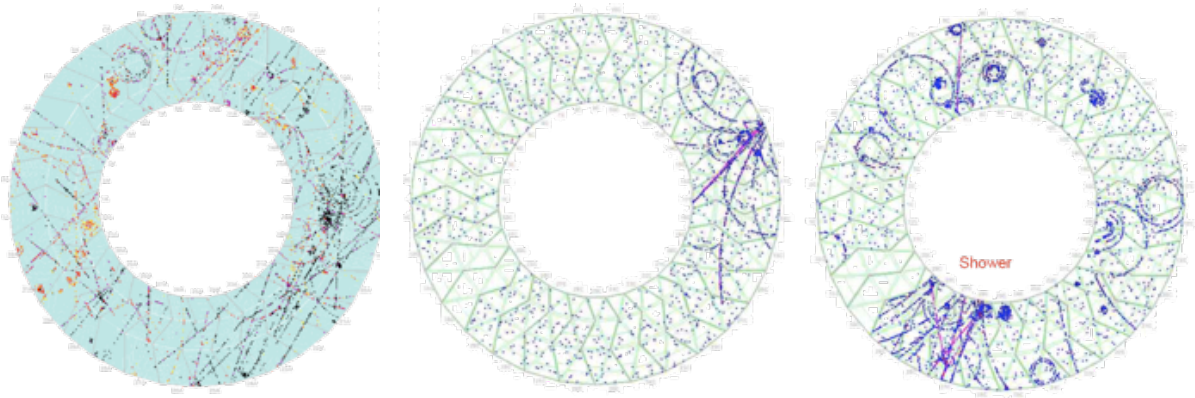


Fig. 17: Cosmic ray shower tracks seen in the ATLAS transition radiation tracker.

at the CERN Tier-0 centre, reprocessed after software and conditions upgrades at the Tier-1 worldwide computing centres, and distributed for analysis on the LHC Computing Grid.

6.1 Cosmic ray spectra in the inner tracker

Tracks bent in a magnetic field are characterised by five parameters. The parameters are defined with respect to a reference point, the perigee, which is the point of closest approach to the beam axis (along z). The impact parameters d_0 and z_0 are the signed distances to the z-axis and the z-coordinate of the perigee, respectively. Accordingly, the angles ϕ_0 and θ_0 are defined in the transverse plane and with respect to the z-axis at the perigee, respectively. The fifth parameter, q/p, is the charge of the cosmic muon divided by its momentum, defining curvature and orientation of the track helix.

Figure 18 shows the angular and impact parameter distributions of cosmic muon tracks measured in the ATLAS inner tracker. The asymmetries reflect the top-down nature of cosmic tracks, and the shaft

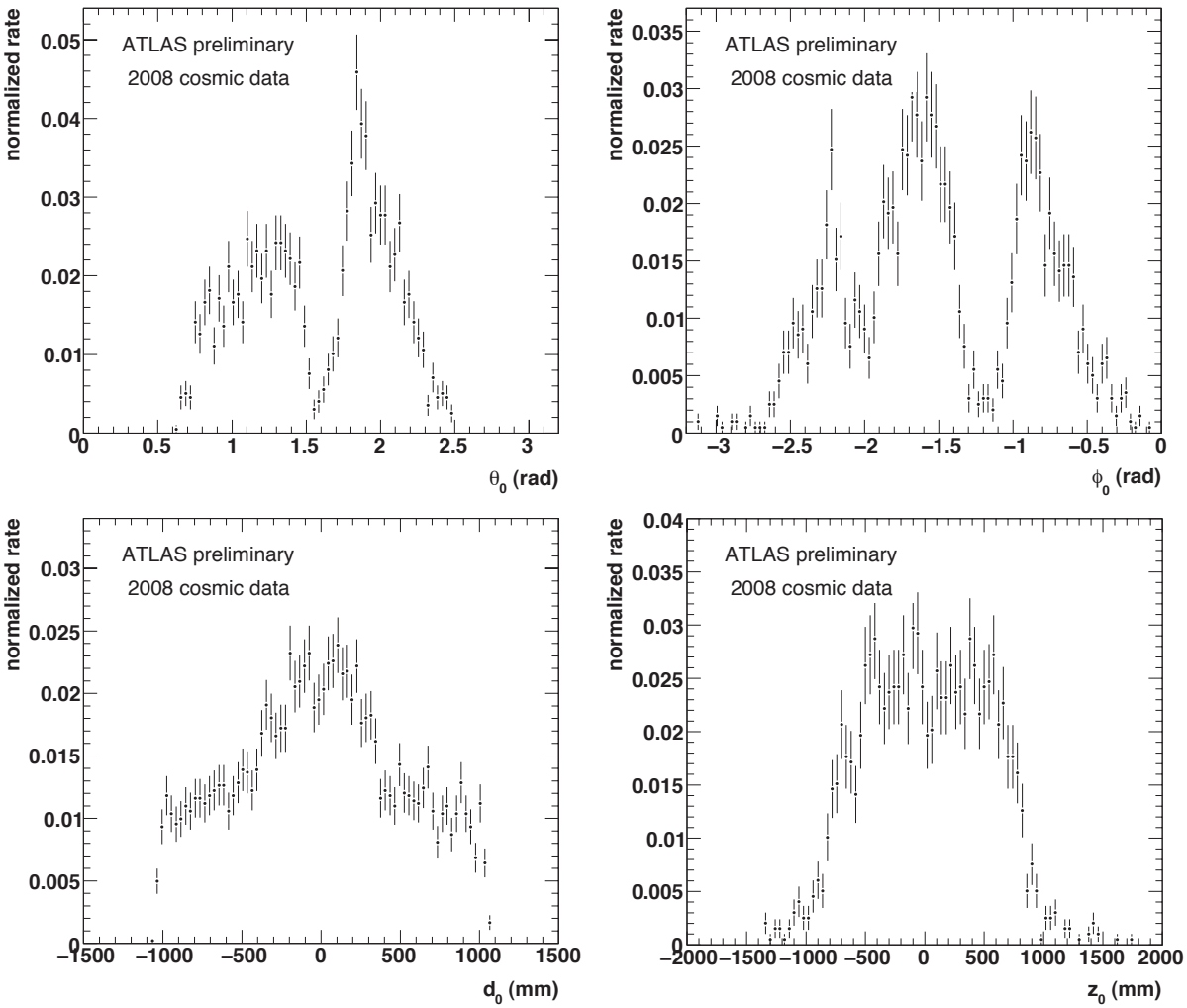


Fig. 18: Track parameter distributions of cosmic muon tracks measured in the ATLAS inner tracker. Shown are the polar and azimuthal angles (upper plots) and transverse and longitudinal impact parameters (lower plots). The asymmetries reflect the top-down nature of cosmic tracks, and the shaft architecture of the ATLAS cavern.

architecture of the ATLAS cavern (Fig. 13). For the θ_0 and z_0 distributions, the tracks are required to have hits in the silicon detectors, because these parameters are not measured by the transition radiation tracker (barrel).

6.2 Inner tracker alignment

The high-precision tracking detectors of ATLAS and CMS, and the huge muons systems (especially in ATLAS) challenge the accuracy with which the positions of the active detector elements must be known. And although the detectors have been built and installed with the greatest care, it does not meet the requirements imposed by the detector performance and by physics. Therefore the detectors have to be empirically *aligned*. Alignment signifies measuring the real detector positions and orientations from data, and correcting the reconstruction software accordingly. (It does not mean moving detector parts!). Several methods of varying complexity to solve alignment problems exist, and it is convenient to separate the alignment procedure into alignment levels, such as system, layer, and module, requiring increasing statistics due to an increasing number of degrees of freedom.

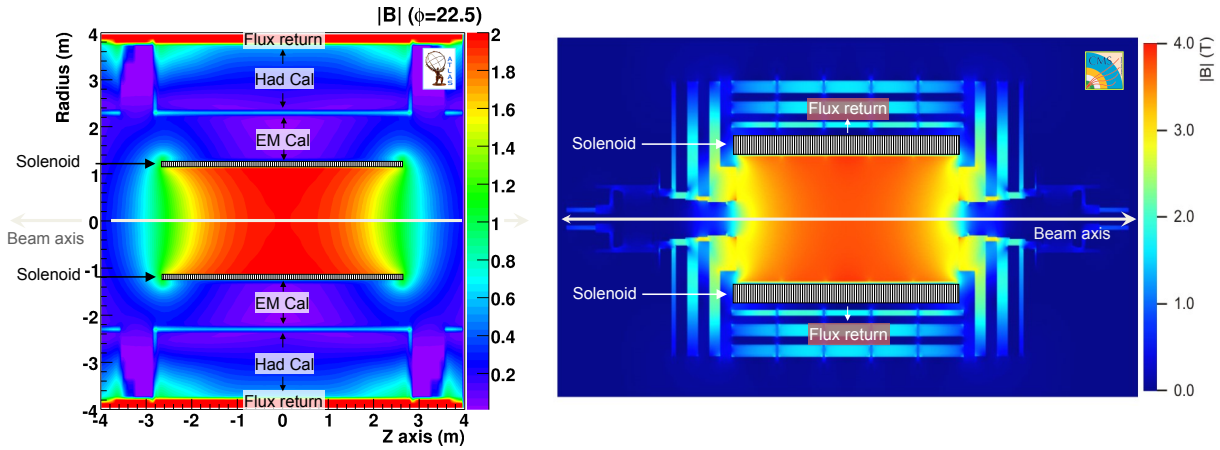


Fig. 19: Solenoid fieldmaps for ATLAS (left) and CMS (right). The colour scales are indicated on the vertical axes. Because the CMS solenoid is much longer (axial length of 12.9 m compared to 5.3 m in ATLAS), the inner tracking detectors, with total active lengths of 5.6 m (ATLAS) and 5.4 m (CMS), see a more homogeneous field in CMS than in ATLAS, where the inhomogeneities in the endcaps can reach up to 50% (which are however accurately mapped with magnetic field surveys and properly included in the reconstruction).

Alignment of the inner tracking systems

The inner tracking systems of ATLAS and CMS (*cf.* Section 3) provide excellent position resolution, with (ATLAS-barrel numbers) 10 μm ($r\phi$), 115 μm (z) for the Pixel device (total of 1744 modules), 17 μm ($r\phi$), 580 μm (z) for the silicon strip detector (4088 modules), and 130 μm ($r\phi$) per straw for the transition radiation tracker (2688 modules). A reasonable challenge is to align all parts of the detectors so that the track degradation due to misalignment not exceed 20% of the intrinsic resolution. The sources of information used for alignment are fourfold: (*i*) assembly knowledge: construction precision and survey data, for the initial alignment precision, and for corrections and uncertainties; (*ii*) online monitoring and alignment: lasers and optical cameras, before and during a run; (*iii*) offline track-based alignment: using physics and track residual information; (*iv*) offline monitoring: using physics observables, tracks and particle identification parameters.

Before coming to the alignment based on track residuals, let us briefly recall how a track momentum is measured. Charged particles are deflected in the homogeneous[19] axial field (i.e., the field is oriented parallel to the z coordinate along the beam line) of the solenoid magnet. Since the Lorentz force is perpendicular to the magnetic (B) field and to the particle's flight vector, the particle trajectory projected onto the plane perpendicular to the B field describes a circle with radius $r\,[\mathrm{m}] = p_T\,[\mathrm{GeV}]/(0.3 \cdot B\,[\mathrm{T}])$. Thus, for transverse momenta between 10 GeV and 1000 GeV, one finds radii between 17 m (9 m) and 1700 m (895 m), for ATLAS (CMS), which are to be compared with the radius of $\sim$1 m of the ATLAS and CMS inner tracking systems. Tracks with transverse momenta smaller than 0.3 GeV (ATLAS) or 0.6 GeV (CMS) become so-called 'loopers', which travel a full circle in the inner tracker and do not reach the barrel electromagnetic calorimeter. The r and p_T values of a track are derived from the measurement of the track's sagitta (s) by $r \approx L/(8s)$ (if $s \ll L$), where L is half the length of the transverse distance vector between the two extreme measurement points of the arc in the tracking system, and the sagitta determines the maximum distance between the intersection of the transverse distance vector with the radius vector, and the arc (the sagitta measures the deviation of the arc from a straight line, L, *cf.* Fig. 26). The smaller the sagitta s the larger the radius and therefore the momentum of the track and, for constant precision on s, the larger the relative error on the sagitta determination and hence on p_T: $p_T \propto s^{-1}$ and $\sigma(p_T)/p_T \propto p_T$.

[19] Not quite, as seen below.

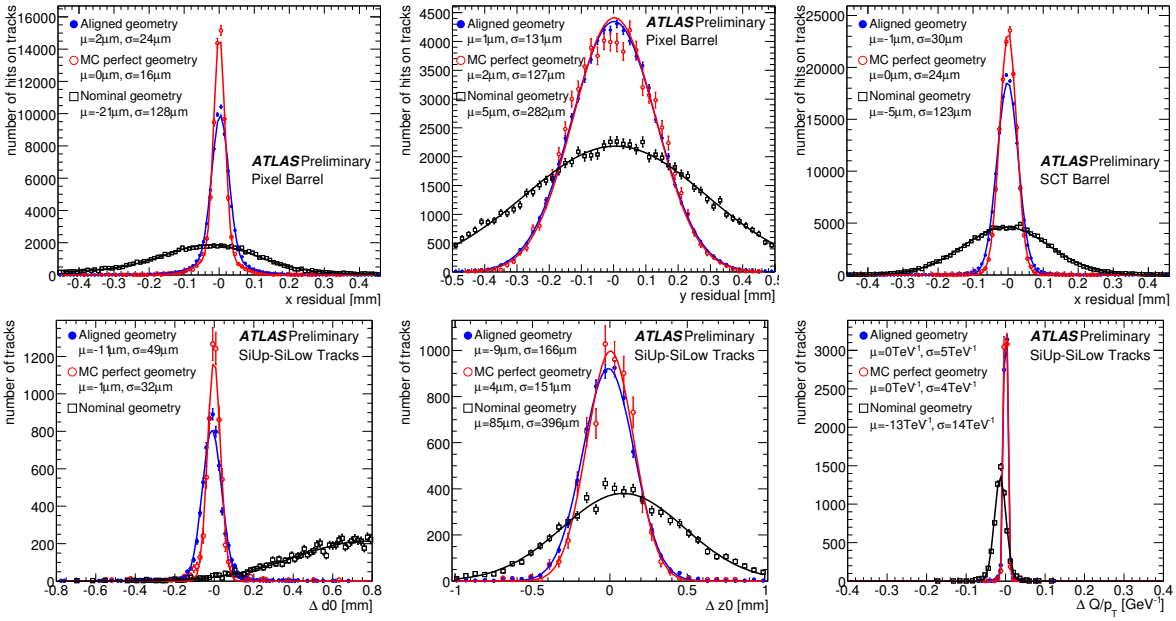


Fig. 21: Hit residuals (upper plots) for the ATLAS pixel and silicon strip detectors before alignment (open squares), after alignment with cosmic ray tracks (full circles), and for ideal conditions from Monte Carlo simulation (open circles). The lower plots give the impact parameter resolution for the same three data samples. The resolution is obtained with the track-splitting technique (see text).

Track fitting in the LHC environment is very challenging. It must deal with ambiguities, hit overlaps, multiple scattering, bremsstrahlung, multiple vertices, etc. Track fitters take Gaussian noise (e.g., Kalman filter) and non-Gaussian noise (e.g., Gaussian sum filter) into account. Owing to the large number of tracks per event and because tracks are used for selection in the high-level trigger, the fits must be very fast.

Figure 19 shows the superconducting solenoid field maps for ATLAS and CMS. Inhomogeneities in the magnetic field strengths occur towards the end of the solenoids, which are strongly influenced by the magnetic structure of the nearby detector elements. The $\sim$2 T flux return yoke in CMS is used for muon momentum measurement. (The ATLAS return yoke, integrated into the tile hadronic calorimeter and its support structure, also produces a $\sim$2 T.m azimuthal track deviation, which is, however, not measured precisely in the muon spectrometer and hence not used for momentum measurement.)

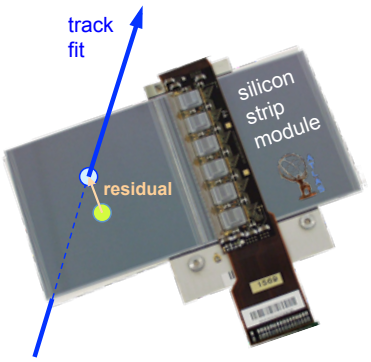


Fig. 20: Sketch of a track model through an ATLAS silicon strip tracker module, and a measured close-by hit defining the hit residual.

The alignment algorithm minimises the track residuals by fitting detector positions (layers and modules) to measured tracks (Fig. 20). The fit minimises a global estimator, which could be written by $\chi^2 = \sum_{i\in\text{hits}}(m(\vec{\alpha}) - h_i)^2/\sigma_i^2$, where the function m corresponds to the model prediction (track) at module of hit i, $\vec{\alpha}$ is the vector of track parameters, and h_i and σ_i are the measured hits and their errors. The full global χ^2 function must, however, also account for correlations so that it becomes: $\chi^2 = \sum_{\text{tracks}}(r^T V^{-1} r)$, where the residuals r are functions of the track parameters, the alignment parameters and the hit measurements along a track. The χ^2 function is simultaneously minimised with respect to the track and the alignment parameters.

The smallest movable object in the alignment procedure is a module, which has 6 degrees of

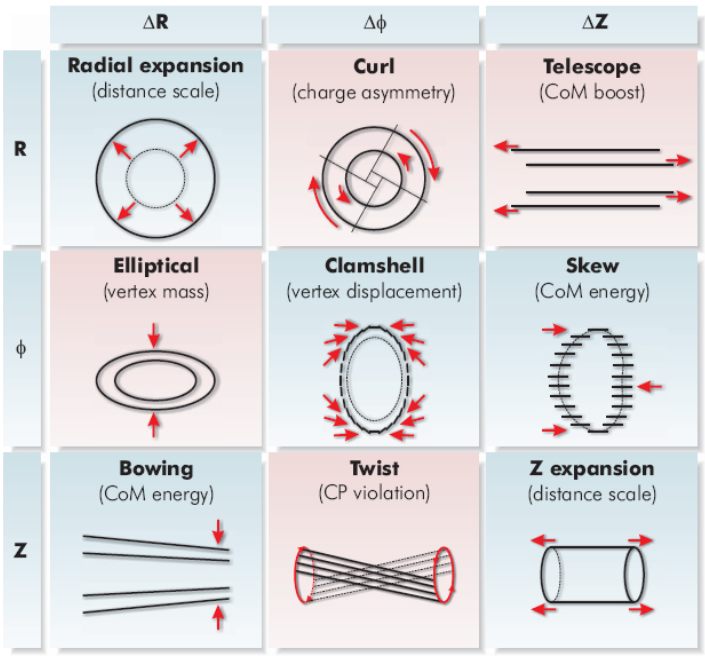


Fig. 22: Different types of misalignment according to transverse distortions in R, ϕ, and deformations along the beam axis (z). The pink types leave the χ^2 estimator approximately invariant ('weak modes').

freedom: 3 translation coordinates and 3 rotation angles. Taking into account the total number of modules of ca. 8500 (ATLAS number), one obtains 51 000 degrees of freedom that need to be determined by the fit. Depending on the alignment level (whole barrel/endcap, layers/disks, modules) different techniques can be used, where for either of these the correlations between fit parameters are important ingredients to help the fit converge rapidly. Neglecting correlations may not lead to a wrong fit result, after full convergence, but it is less efficient.

Figure 21 shows residual distributions for the ATLAS pixel and silicon strip detectors, as well as impact parameter and Q/p_T distributions, before and after alignment with cosmic ray tracks. The widths of these distributions are convolutions of the intrinsic hit and tracking resolution (seen under ideal conditions), and misalignment effects. The impact parameter and transverse momentum resolutions are obtained by splitting a cosmic ray muon track traversing the full detector into two tracks that are re-fit independently and compared.[20] A total of 4.9 (2.7) million tracks with solenoid field on (off) have been used by ATLAS (similar numbers of tracks are used by CMS for alignment), of which 1.2 million (230 thousand) have silicon strip (pixel) track components so that they can be used to align these detectors. Alignment results close to ideal have been obtained.

Weak modes

Unfortunately, the minimisation of hit residuals does not guarantee that indeed the true positions of the detector elements have been determined. This is because the residuals, and hence the χ^2 estimator, are insensitive against some types of misalignment, which may nevertheless impact the physics performance. Examples for such 'weak modes' are elliptical skews, i.e., distortions of the type $\delta\phi = \lambda + \beta/R$ or $\delta z \approx R$. Figure 22 summarises the various types of misalignment. The pink-coloured types represent weak modes in the global residual-based χ^2 estimator. Weak modes contribute to the lowest part of the eigenspectrum. Their deformations bias physics measurements and lead to systematic effects. The understanding of these effects is thus of utmost importance. Weak modes can be constrained by adding

[20] The resolution is the RMS of the difference divided by $\sqrt{2}$.

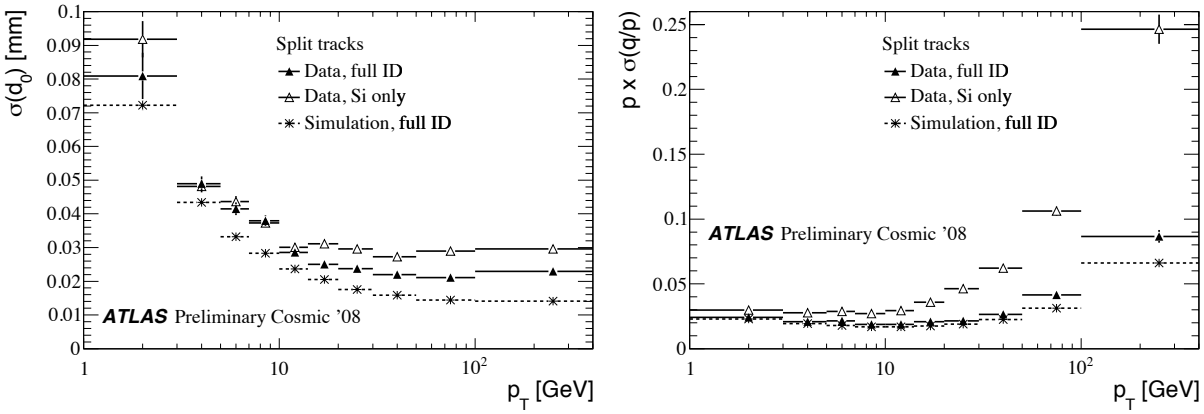


Fig. 23: Transverse impact parameter resolution (left) and relative momentum resolution (right) versus the transverse momentum for the ATLAS inner tracker. The full (open) triangles give the results for all inner tracker detectors combined (only silicon pixel and strip detectors), and the asterisk is the expectation from Monte Carlo simulation with ideal alignment conditions.

more information to the fit, such as: (*i*) cosmic ray and beam halo tracks (off-beam axis) in addition to beam collision data; (*ii*) vertex and beam-spot constraints; (*iii*) resonance masses (Z, J/ψ, Υ, K_S, ...); (*iv*) E/p measurements for electrons; and (*v*) survey data and mechanical constraints.

6.3 Inner tracker resolution

The tracking resolution for cosmic ray muons in the inner tracker is studied by comparing track parameters at the perigees using the track-splitting technique. Because both tracks emerging from the splitting have errors, the quoted resolution is the RMS of the residual distribution of a track parameter divided by $\sqrt{2}$. Well reconstructed tracks are selected for these studies. ATLAS requires a minimum number of hits in Pixel, silicon strip detector and transition radiation tracker of 2, 6 and 25, respectively, and $|d_0| < 40\,\text{mm}$ and $p_T > 1\,\text{GeV}$, and good timing properties. The left-hand plot of Fig. 23 shows the transverse impact parameter resolution versus the transverse momentum for the ATLAS inner tracker. In the low p_T region, the resolution is dominated by multiple scattering. At higher momenta, the resolution becomes independent of the momentum as is expected for almost straight tracks. Including the transition radiation tracker information improves the resolution due to the extended lever arm. The difference between data and the Monte Carlo prediction is a measure of the remaining misalignment. The right-hand plot of shows the relative momentum resolution versus p_T. At intermediate momentum, reduced multiple scattering counterbalances the p_T-dependent rise of the error due to a decreasing relative accuracy of the sagitta measurement. This latter effect dominates at higher momentum. Again, the difference with respect to the Monte Carlo expectation stems from residual misalignment.

6.4 Muon spectrometer alignment

The huge active volumes of the ATLAS and CMS muon spectrometers require a detailed understanding of the inhomogeneous magnetic fields (especially for ATLAS and the CMS endcaps) and the chamber positions to achieve design performance. To derive quantitative requirements, let us briefly recall how the muon precision measurements are obtained. Both experiments use drift tubes, which are standalone coaxial cylindrical drift chambers functioning similarly to proportional tubes, in the barrel (ATLAS also in the outer endcaps for $|\eta| < 2.0$), and cathode strip chambers in the forward direction.

The drift tubes in ATLAS (denoted 'monitored drift tubes' — MDT) are made of thin aluminium tubes with 3 cm diameter (4 cm in CMS, 4 mm for the ATLAS transition radiation tracker), filled with a 93% argon and 7% CO_2 gas mixture at 3 bar pressure (Fig. 24). A 50 μm gold-plated tungsten wire in

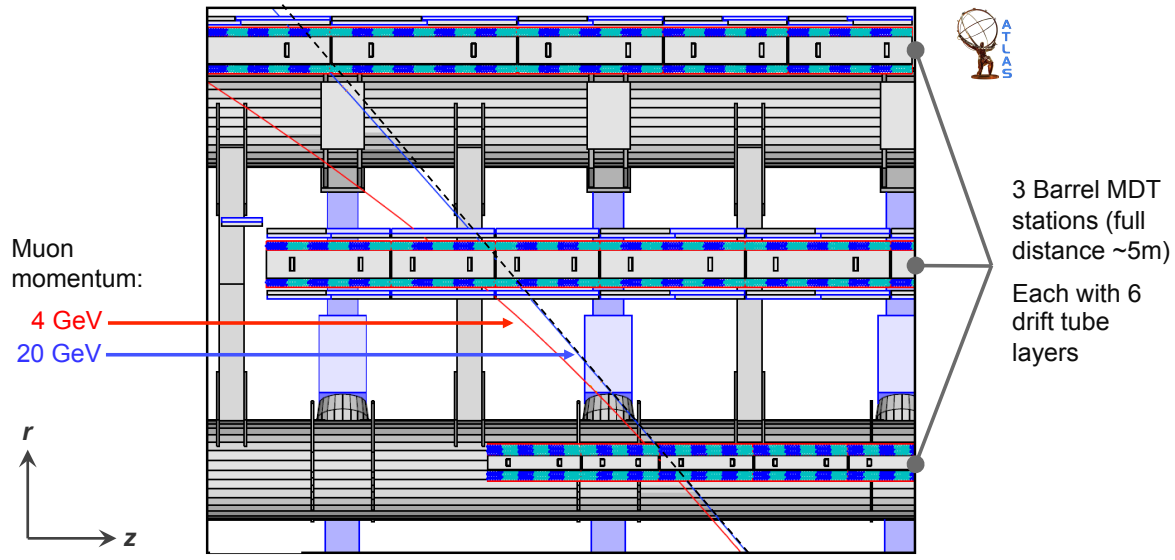


Fig. 25: Drawing of ATLAS barrel monitored drift tube stations. There are three of these spanning a full radial distance of ∼5 m. Shown by the curved lines are simulated muon tracks with 4 GeV (red) and 20 GeV (blue), bent in the z direction by the toroidal magnetic fields. The curvature is hardly visible for the latter track (see straight dashed line for comparison). The MDT system is designed to measure 1 TeV tracks with 10% relative accuracy, requiring a position alignment of better than 40 μm.

the centre of each tube serves as anode with an applied potential of 3080 V. A charged track traversing the tube ionises the gas and the ionised electrons drift in the electrical field to the wire, while the ions drift to the cathode (cylinder). From the measured hit time of the induced electrical pulse, and the known drift velocity ('space-drift time (r-t) relation'), it is possible to determine a *drift circle* around the anode wire, tangential to which the track has passed.

The measurement of several adjacent layers of tubes provides the redundant information required for a full track fit. The measured drift time in a tube reaches up to 800 ns corresponding to a drift velocity of approximately 18 km/s. The average position resolution is 80 μm per tube (250 μm in CMS), but varies strongly along the drift radius: tracks very far from the anode wire are measured with better precision than close tracks, due to the smaller dispersion in the drift time of the incoming electrons.

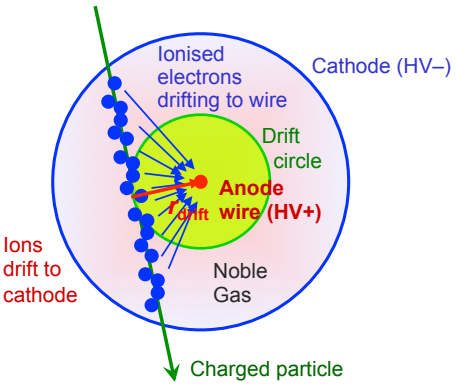


Fig. 24: Principle of a drift tube used for precision measurement in the ATLAS and CMS muon systems, and also in the ATLAS transition radiation tracker.

The ATLAS drift tubes are arranged in large-sized MDT chambers with six tube layers oriented along ϕ to allow for a precise measurement of the z coordinate, the direction of which the charged particles are bent in the toroidal magnetic fields. Three almost equally spaced stations of MDT chambers (inner, middle and outer) are installed in the barrel with about 2.5 m radial distance from each other (Fig. 25). A 1 TeV track has a sagitta of about $s = 500\,\mu$m at $\eta = 0$ (*cf.* sketch in Fig. 26). A measurement of that sagitta with 10% accuracy requires the error induced by misalignment to be significantly smaller than 50 μm. With $\sigma(s) \approx \sqrt{3/2} \cdot \sigma(z)$, one finds $\sigma_{\text{misallign}}(z) \ll 40\,\mu$m, which represents a tremendous alignment challenge given the size of the system.

Figure 27 shows an example of a misaligned MDT chamber in ATLAS (from simulation). In the

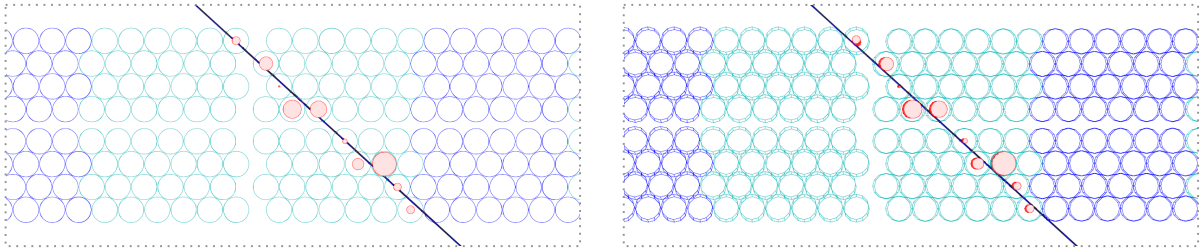


Fig. 27: Example of a misaligned drift tube chamber in the ATLAS barrel muon spectrometer (simulation). In the left-hand picture, without alignment corrections, it is not possible to draw a straight line track through the drift circles. After alignment (right-hand picture) the chambers have been slightly tilted so that a good track fit can be obtained.

left drawing, where no alignment corrections have been applied, the track is not tangential to all drift circles. The χ^2 of the track fit is bad. In the right drawing the chambers have been aligned leading to a good track fit.

Optical muon chamber alignment in ATLAS

ATLAS implements a twofold alignment strategy for the muon system: fits to measured tracks from cosmic rays and collision events, in particular using straight tracks without the toroid fields, provide the absolute MDT chamber positions.[21] Relative chamber movements due to temperature-dependent 'breathing' and when switching on the toroid magnets, are monitored by means of an optical alignment system, designed to detect slow chamber displacements, occurring at a timescale of hours or more. The system is based on optical and temperature sensors, and on alignment bars, which are up to 9.6 m long instrumented aluminium tubes used as precision reference rulers. The information from the optical system together with the track-based alignment is used in the offline track reconstruction to correct for the MDT chamber misalignment. Similar to ATLAS, CMS is instrumented with a precise and complex opto-mechanical alignment system that provides a common reference frame between tracker and muon detection systems by means of a net of laser beams. We discuss in the following the ATLAS system.

To first order, only the relative alignment of triplets of chambers traversed by the same muon track is important for a precise sagitta measurement. The barrel optical alignment system thus uses 3-point straightness monitors, which are installed on the inner, middle and outer chambers to form projective lines pointing to the interaction region.[22] The straightness monitor creates a highly redundant image of a coded mask (for example a chess-like pattern) through a lens onto a charged-coupled device (CCD) acting as screen. The mask is lit by infrared LEDs passed through a diffuser to minimise effects of imperfections in the light source. The relative position in transverse direction to the projective lines is measured along the line mask, the

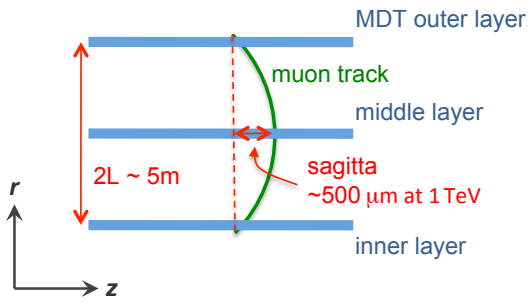


Fig. 26: Sketch for the muon sagitta measurement in ATLAS. For a 1 TeV track the sagitta measures about 500 μm.

[21] Full alignment not only requires a proper positioning of the chambers and tubes in the chambers, but one must also correct for the wire sag in the drift tubes, which has been measured from survey data for a fraction of the tubes, and must be derived from track fits for the remaining ones. The wire-sag induced error in the position measurement amounts to 20–30 μm, depending on the size of the MDT chamber.

[22] In the endcaps, projective lines cannot be installed because the cryostats of the endcap toroid magnets block the way to the interaction region. The optical alignment system thus relies on high-precision reference rulers and alignment bars forming an alignment grid.

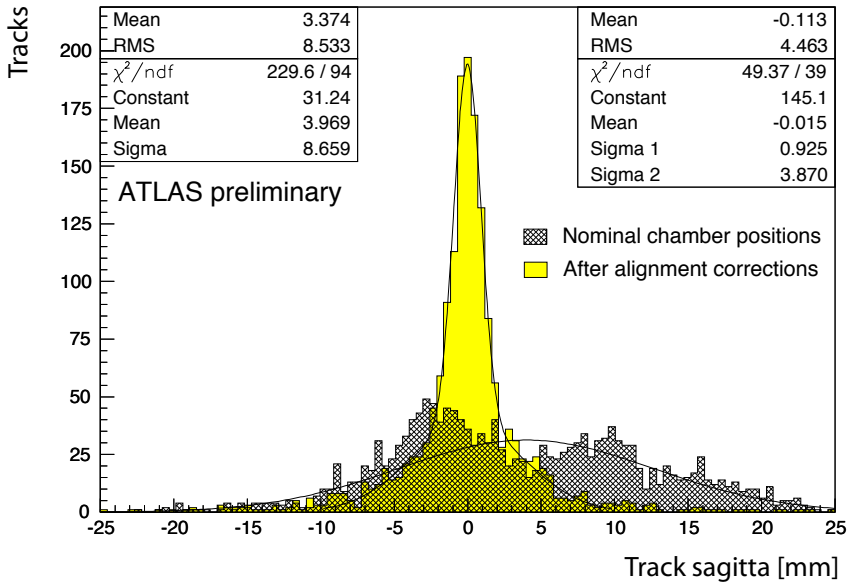


Fig. 28: Track 'sagitta' for straight cosmic ray muon tracks (toroid fields off) in the ATLAS endcaps before (dark shaded) and after (light shaded/yellow) applying the optical alignment. The sagitta is calculated from the distance in the precision coordinate of the middle chamber segment from the line joining the inner and outer endcap segments. After alignment, the resolution (width) is dominated by multiple scattering effects.

optical centre of the lens, and the CCD camera. It is also possible to measure the (relative) rotation of the mask or the sensor, and the relative rotation around any axis of the mask with respect to the CCD camera. Finally, by computing the actual image size and comparing it with the known mask size (magnification), the position of the lens along the longitudinal axis can be obtained. A total of 6000 (7000) optical lines have been installed in the ATLAS barrel (endcap). Not all of these are projective. In the barrel, praxial lines align adjacent chambers in each layer. In the endcaps there are bars, polar and proximity lines.

The absolute resolution of the optical alignment system is of the order of 300–500 μm, which is insufficient for precision measurements. Hence the necessity to rely on track measurements for absolute chamber positions. The relative optical alignment accuracy has been evaluated with simulated muon shifts of the H8 test beam arrangement and found to correct misalignment within 14 μm error (RMS) on the sagitta, which is well within the specified requirement [18]. Figure 28 shows the distribution of sagitta values for straight cosmic muon tracks (the toroid magnets were turned off so the expected sagitta is zero) in the ATLAS endcaps before and after applying the optical alignment. The sagitta is computed from the distance in the precision coordinate of the middle chamber segment from the line joining the inner and outer chamber segments. The resolution found is compatible with the expectation. The tails in the sagitta distribution after alignment originate from *multiple scattering*.

Digression. Multiple scattering denotes the deflection by (or convolution of) successive small-angle scatters of a charged particle traversing a medium. The multiple scattering cross section, dominated by Coulomb scattering from nuclei, is proportional to $\sqrt{\text{pathlength}/X_0} \cdot p^{-1}$, i.e., it is enhanced for soft particles and dense matter. The angular distribution is approximately Gaussian at small angles (owing to the central limit theorem), but also large-angle Rutherford scattering occurs with a differential cross section $\propto \sin^4(\theta/2)$. Multiple scattering is analogous to diffusion. Figure 29 shows the effect of light diffusion on a wet windscreen. The more matter in terms of radiation lengths a particle traverses in the tracking volume, the more the detector 'sees' the particle as we see other cars at night in rainy weather with a broken wiper. Multiple scattering complicates the track fitting and limits the resolution of the momentum measurement.

Figure 30 (left) gives the contributions to the standalone muon momentum resolution versus the incident momentum of the ATLAS barrel muon spectrometer. Multiple scattering (black line) determines the resolution for momenta below ~200 GeV. At very low momentum (below 20 GeV) the fluctuations in

Fig. 29: Multiple scattering (diffusion) of light passing through a wetter and wetter windscreen (left to right).

the energy loss of the muon traversing the calorimeters becomes the dominant effect (cyan coloured line — the blue line indicates the resolution with respect to the entrance at the muon spectrometer). However, below 100 GeV the momentum measurement is in any case dominated by the inner tracking system. For high-momentum muons the contribution from the intrinsic drift tube resolution and *r-t* calibration is of similar magnitude as the expected systematic error in the mechanical alignment, hence the challenge for the alignment system. The right-hand plot in Fig. 30 shows the fractional standalone momentum resolution measured by comparing top and bottom muon spectrometer tracks in cosmic ray data (track splitting method). The measured resolution is compatible with the expected one from Monte Carlo simulation at transverse momenta below 100 GeV, and is degraded at higher momenta. The degradation is caused by imperfect alignment of the muon chambers and by limited timing accuracy because cosmic muons are not synchronous with the artificial LHC clock used in drift time measurements (no fixed time reference).

6.5 Muon charge asymmetry in cosmic rays

The charge ratio of positive to negative muons in cosmic rays, with momenta in the range 10–300 GeV, has been measured to be 1.27 at sea level [19], and is expected to increase somewhat with the muon momentum due to a growing influence from kaon decays (the charge ratio of pion decays is expected to be approximately 1.25, while it is 2 for kaons [20]).

In 2006, during the 'Magnet Test and Cosmic Challenge (MTCC)', CMS performed a measurement of the muon charge asymmetry on the surface, using a 30° slice of the detector including the muon drift tubes in presence of a 4 T solenoid field [21]. Owing to the high muon rate at the surface, 337 000 high quality tracks with hits in at least 3 (of 4) barrel stations and transverse momentum larger than 3 GeV could be selected. The most important systematic effect on the charge measurement stems from the charge-dependent alignment uncertainty, in particular for high muon momenta. The resolution-induced charge misidentification probability is estimated from Monte Carlo simulation and also contributes significantly to the systematic error above 100 GeV (no inner tracking used). The total systematic error varies between 2% below 10 GeV, ∼8% at 100 GeV, and up to and beyond 20% above 100 GeV. It exceeds the statistical errors at all muon momenta. To compare the raw charge ratio measurement with other measurements, the result is expressed in terms of the muon momentum before entering CMS using Monte Carlo simulation. The resulting momentum correction is about +7 GeV and almost independent of the muon momentum. Figure 31 (right plot) shows the charge-ratio measurements versus the corrected muon momentum, together with results from other sources (see references in Ref. [21]). Within their uncertainties, the CMS results can be regarded as independent of the muon momentum, giving the

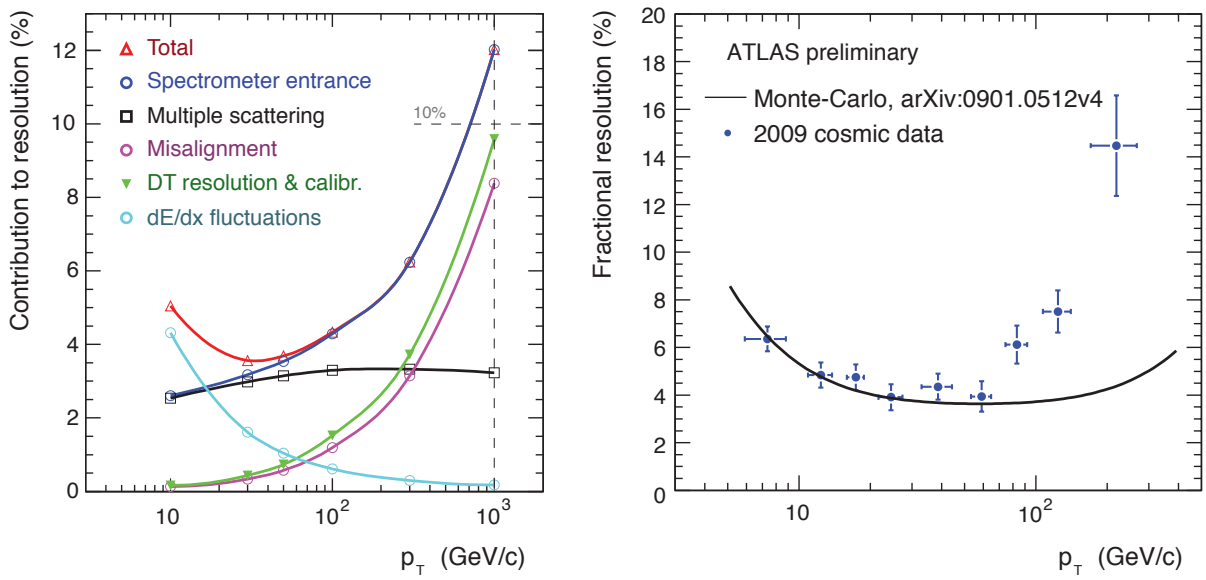


Fig. 30: Left: expected contributions to the standalone muon momentum resolution of the ATLAS barrel muon spectrometer (Monte Carlo simulation). See text for a discussion of the various terms. Right: fractional standalone momentum resolution measured by comparing split top and bottom muon spectrometer tracks in cosmic ray data. The degradation of the measured resolution with respect to the expected one is mainly due to imperfect alignment, but also due to *r-t* relation inaccuracies due to the missing LHC clock reference.

average $R_{\mu^+/\mu^-} = 1.282 \pm 0.004 \pm 0.007$, where the first error is statistical and the second systematic. The left plot in Fig. 31 gives a compilation of previous muon charge-ratio data between 0.1 and 7 TeV taken from a MINOS publication [20]. Superimposed is the model expectation.

6.6 Combining muon and inner tracker reconstruction

The comparison of cosmic muon track measurements in the muon system and in the inner tracker allows one to study the momentum scale and the energy loss in the calorimeters, and to tune the Monte Carlo simulation. Figure 32 shows a comparison between standalone track fits to cosmic ray muons in the ATLAS spectrometer and the inner tracker. Shown are the polar and azimuthal angle correlation, the azimuthal angle and impact parameter differences, and the momentum scale difference. A satisfactory agreement is observed between the detectors, and between data (dots) and the Monte Carlo prediction (histograms), showing that the relative alignment and the momentum scales are understood within the available statistics (a single run was used for these plots).

The difference in the momentum scale of 3 GeV on average corresponds to the energy loss of the muons between spectrometer and inner tracker, mainly when traversing the calorimeters.[23] It is well described by the simulation.

6.7 Cosmic ray muons in the inner tracker

One of the first measurements performed with cosmic ray muons is the verification of the hit reconstruction efficiency in the silicon trackers, which is expected to be very high ($> 99\%$). The method is as

[23] One can attempt a back-of-the envelope calculation of the expected energy loss to understand the magnitude of the effect. The barrel ATLAS hadronic calorimeter uses iron absorber and plastic scintillator tiles. Inserting the corresponding densities and dE/dx expectations for cosmic ray muons one finds: $\langle \Delta E(\text{Had cal}) \rangle \simeq 200\,\text{cm} \cdot (0.4 \cdot dE/dx|_{\text{Fe}} \cdot 11.8\,\text{g/cm}^3 + 0.6 \cdot dE/dx|_{\text{C}} \cdot 2\,\text{g/cm}^3) \approx 2.1\,\text{GeV}$. Similarly one finds for the electromagnetic liquid-argon accordion calorimeter: $\langle \Delta E(\text{EM cal}) \rangle \simeq 100\,\text{cm} \cdot (0.4 \cdot dE/dx|_{\text{Pb}} \cdot 16.9\,\text{g/cm}^3) \approx 1.0\,\text{GeV}$, and for the contribution from the thin solenoid magnet: $\langle \Delta E(\text{solenoid}) \rangle \simeq 5\,\text{cm} \cdot (0.4 \cdot dE/dx|_{\text{Cu}} \cdot 8.9\,\text{g/cm}^3) \approx 0.1\,\text{GeV}$. The sum of all contributions gives roughly 3.2 GeV expected energy loss.

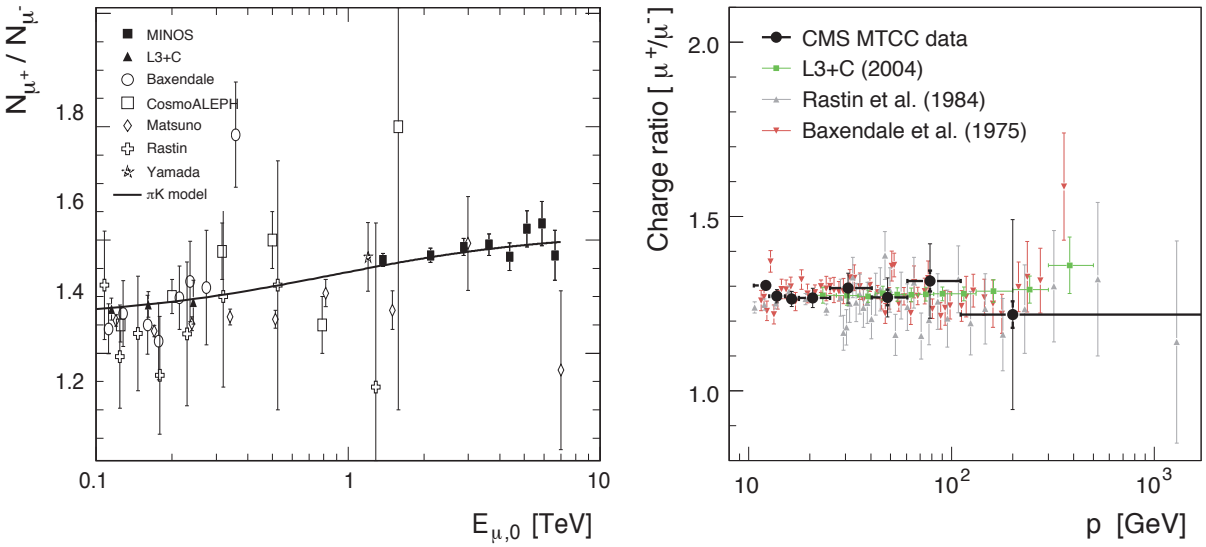


Fig. 31: Left: muon charge-ratio measurements compiled by the MINOS experiment [20]. Right: muon charge ratio measured by CMS (black dots) with statistical (bold bars) and systematic errors (thin bars), together with results from other experiments (see Ref. [21] for references).

follows.

1. Selection of good quality tracks by requiring a large number of silicon hits, satisfying goodness-of-fit and a small incident angle.
2. To measure the efficiency of the i-th layer, the hits from this layer (if there are any) are excluded, and the track is refitted without the i-th layer.
3. The hit efficiency is computed by searching for hits in the i-th layer within a narrow road around the refitted track.

The hit reconstruction efficiencies obtained with this method for the ATLAS barrel silicon strip tracker are shown in Fig. 33. Here the tracks were required to have at least 10 hits in the silicon tracker, 30 hits in the transition radiation tracker, and an average χ^2 per degree of freedom smaller than 2. Furthermore their intersection with the modules had to be within 40 degrees of normal incidence, and there had to be a hit of some kind on the track before and after the module being studied. Finally a guard region around the edge of the active silicon was excluded. The silicon efficiency was then found to be 99.75% on average. Very similar results have been found for the ATLAS pixel detector using the same measurement technique, and also for the CMS silicon pixel and strip detectors.

The hit reconstruction efficiency per straw for the ATLAS transition radiation tracker depends on the distance of the track to the anode wire (maximum distance 2 mm). There is a plateau region below 1 mm where the efficiency reaches 97.2%, decreasing to ∼90% (80%) at 1.5 mm (1.8 mm) and steeply dropping beyond that distance.

6.8 Measurement of the Lorentz angle

The solenoid field applies a Lorentz force on moving charges that deflects the track-induced ionisation electrons and holes, travelling through the depleted substrate of the silicon junction along the high-voltage potential (Hall effect). The deflection angle is denoted *Lorentz angle*. The value of the Lorentz angle depends on the mobility of the charge carriers as well as the external magnetic field. For silicon immersed in a magnetic field B the Lorentz angle α_L is given by $\tan\alpha_L = \mu_H B = \gamma\mu_d B$, where μ_H is the Hall mobility, γ represents the Hall factor which is of order unity, and μ_d is the drift mobility, which is a

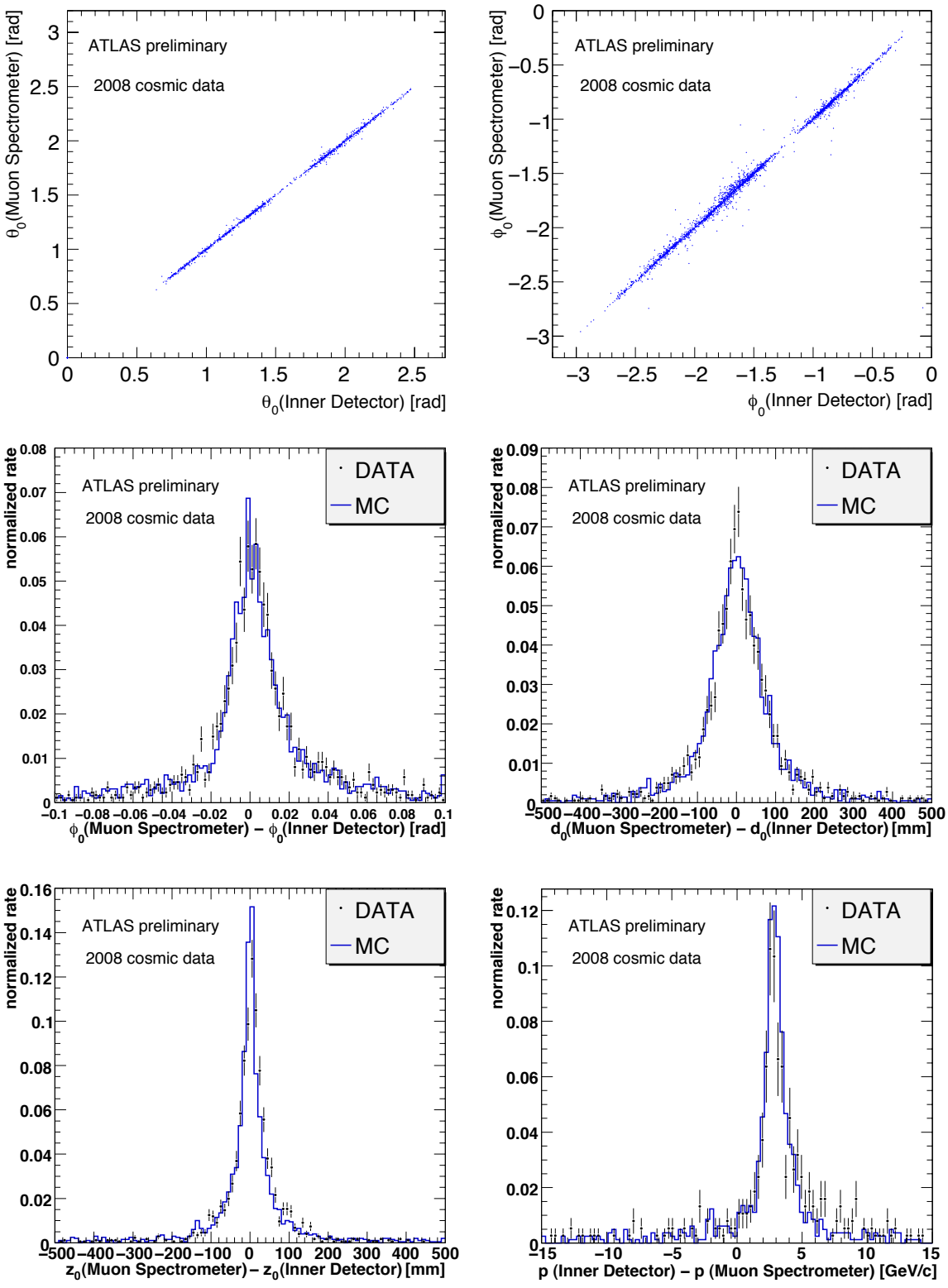


Fig. 32: Comparison between standalone track fits to cosmic ray muons in the ATLAS muon spectrometer and the inner tracker. Shown are the polar and azimuthal angle correlation (upper plots), azimuthal angle and impact parameter differences (middle and lower left plot), and momentum scale difference (lower right plot, sensitive to the energy loss of the muons when traversing the calorimeters). The dots are data and the histograms correspond to the Monte Carlo prediction.

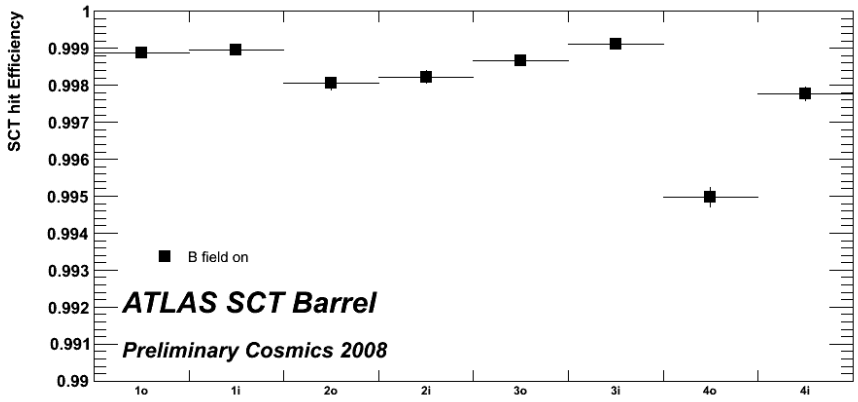


Fig. 33: Hit efficiencies for the ATLAS barrel silicon strip tracker as measured with cosmic muon tracks (see text for details of track requirements and procedure).

function of the ratio of drift velocity to the electric field induced by the bias voltage. The drift velocity for both electrons and holes saturates at high electric field. This leads to a drop in the mobility thus decreasing the Lorentz angle.[24]

Figure 34 sketches the Lorentz deflection effect. Owing to the opposite charge of electrons and holes, both carriers are deflected into the same transverse direction along the Lorentz force. The deflection generates a bias in the position measurement (cluster barycentre) of the track incident in the silicon strip or pixel. The bias could be reduced by tilting the modules in the direction of the Lorentz angle, and indeed the modules in the ATLAS and CMS silicon detectors are tilted (*shingled*). The values for the tilts chosen are, however, due to technical reasons to allow overlaps between adjacent modules.[25] Instead of a mechanical solution, the position bias due to the Lorentz deflection is corrected by software. The correction must be recalibrated at regular intervals because the size of the depletion region in the semiconductor reduces with rising irradiation and constant bias voltage, thus reducing the position bias.

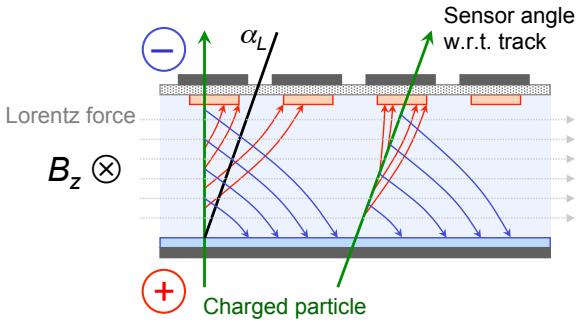


Fig. 34: Sketch illustrating the deflection of moving ionisation charges in the solenoid field, leading to a bias in the position measurement. Tilting the modules by the amount of the Lorentz angle α_L would correct for the bias.

The Lorentz angle is determined empirically by minimising the measured cluster width of hits on tracks. Figure 35 shows the cluster width versus the cosmic muon track incident angle with respect to the module normal for the ATLAS barrel silicon strip tracker. Measurements with and without magnetic field are shown. The value of the Lorentz angle, extracted at the minimum cluster size, is found to be $\alpha_L = (3.93 \pm 0.03 \pm 0.10)^\circ$, where the first error is statistical and the second systematic (for comparison, the Lorentz angle for the ATLAS pixel device is 12.3°).

[24] The electron and hole mobility and hence the Lorentz angle also depend on the temperature: increasing temperature reduces the mobility and thus the Lorentz angle.

[25] In ATLAS the chosen tilts with respect to the pointing axis are 11 degrees (−20 degrees) for the silicon strip tracker (pixel tracker), whereas the Lorentz angle for non-irradiated modules is 4 degrees (13 degrees).

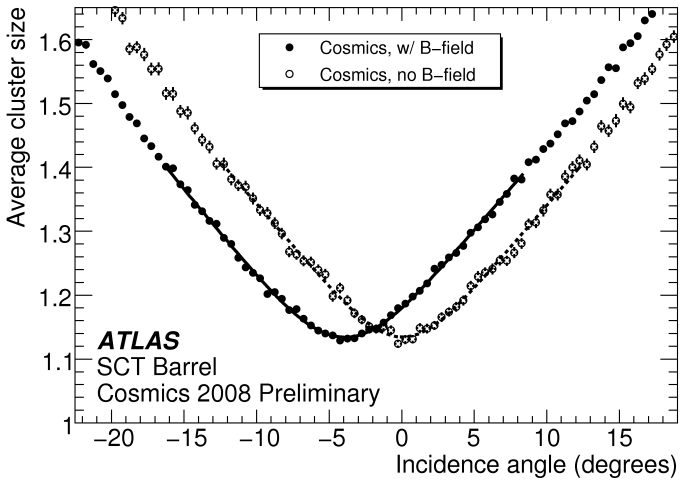


Fig. 35: Measurement of the mean cluster size versus the incidence angle with respect to the module normal in the ATLAS barrel silicon strip tracker, using cosmic ray muon tracks. Measurements with and without magnetic field are shown (the Lorentz angle vanishes without external field). The value of the Lorentz angle is extracted from the position of the minimum cluster size.

6.9 Particle identification with transition radiation

Hits from ultrarelativistic particles, generating transition radiation photons in the keV range that contribute to the gas ionisation in the ATLAS transition radiation tracker (TRT), are identified via dedicated high-threshold readout. It turns on at a gamma factor above $\simeq$1000 (with $p = \beta\gamma m \simeq \gamma m$, the threshold momenta for $\gamma = 1000$ are 0.5 GeV, 105 GeV and 139 GeV for electrons, muons and pions, respectively), and thus essentially only for electrons in the typical energy range, so that it can be used for electron identification.

The principle of the creation of transition radiation via an electric dipole is sketched in Fig. 36. Figure 37 shows the high-threshold hit probability obtained for the ATLAS barrel TRT from 2004 combined test beam data (*cf.* Section 5) for different particle species (left plot), and for cosmic ray muons (right plot). The turn-on curves are found to be in good agreement.

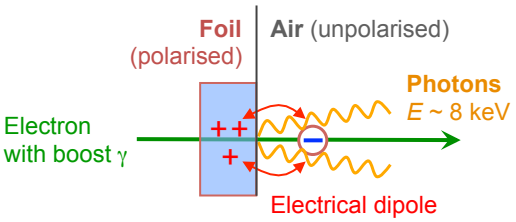


Fig. 36: Transition radiation is produced when charged ultrarelativistic particles traverse the boundary of two different dielectric media (e.g., polymer fibres/foil and air). The radiation is intense enough to be measured for $\gamma > 1000$ and more than 100 boundaries.

6.10 Calorimeter performance with cosmic ray muons

Cosmic ray muons have also been exploited by the calorimeter groups of ATLAS and CMS to study pulse shapes, and occupancy distributions, detect bad channels, understand the muon energy loss in the calorimeters and tails in energy distributions, and for energy inter-calibration purposes.

The total energy sum of all cells along a muon track in the ATLAS hadronic calorimeter is shown in the left-hand plot of Fig. 38. The peak of the minimum-ionising particles (i.e. a particle whose mean energy loss rate through matter is close to the minimum), is well distinguished from the correspond-

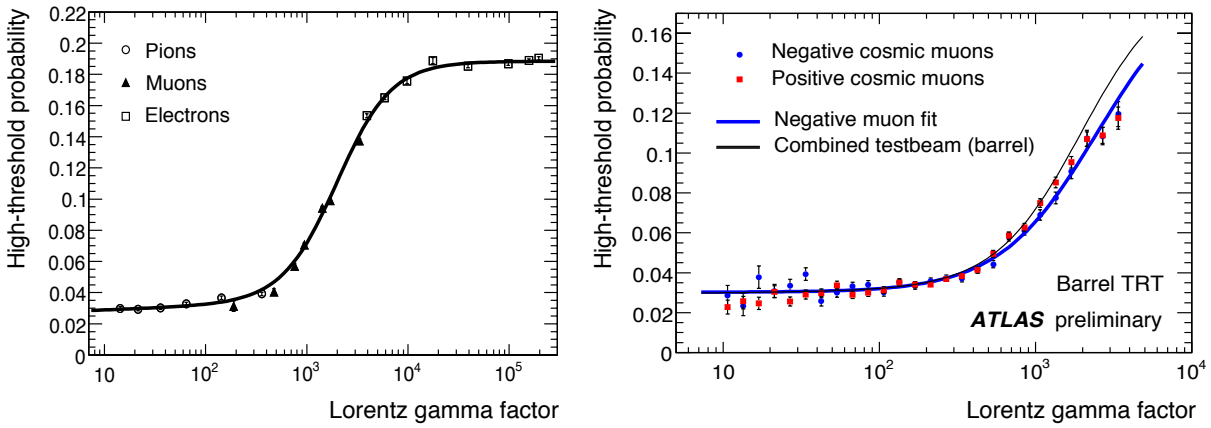


Fig. 37: Left: average probability of a high-threshold hit in the ATLAS barrel transition radiation tracker (TRT) as a function of the Lorentz γ factor for electrons (open squares), muons (full triangles) and pions (open circles) in the energy range 2–350 GeV, as measured in the 2004 combined test beam. Right: transition radiation turn-on versus γ in the ATLAS barrel TRT for cosmic muon tracks. The data points are shown for both muon charges (positive: red dots, negative: blue dots) and are compared with test beam results (black line). The blue line gives a fit to the results obtained with the cosmic data.

ing noise distribution obtained from randomly triggered events. The energy of the cosmic ray muons deposited in the active parts of the hadronic calorimeters of ATLAS and CMS exceeds the one in the electromagnetic calorimeters by approximately a factor of 10. The ionisation energy loss of the muons when traversing the electromagnetic calorimeters is measured by comparing the momenta between the muon system and the inner tracker (*cf.* Fig. 32). It can be correlated on an event-by-event basis to the measured calorimeter energy deposits. This has been done by CMS in the right-hand plot of Fig. 38, where the average electromagnetic calorimeter deposits are drawn versus the muon momentum. Overlaid is the expected energy loss, which is found to be in good agreement with the measurement. The results indicate the correctness of the tracker momentum scale and of the calorimeter energy scale calibrated with electrons at test beams.

The energy deposition can also be directly compared to Monte Carlo simulation, as done by ATLAS in the upper plot of Fig. 39 (see Ref. [22]), showing the energy reconstructed in the first and second layers for data and Monte Carlo cosmic ray events. Good agreement is observed up to the tails both for the shape and the absolute scale. This result can be used to measure the uniformity in the energy response of the calorimeter versus the pseudorapidity by integrating over the response in the azimuth angle (the statistics is insufficient to make a full $\eta \times \phi$ uniformity map). The estimation of the muon energy is done with a fit of the cluster energy distribution using a Landau function, which accounts for fluctuations of the energy deposition in the ionisation process, and a Gaussian describing essentially electronic noise (and also cluster non-containment). The response uniformity is computed from the RMS of the normalised difference between the data and Monte Carlo most probable values (MPV) of the Landau distribution in each η bin. The resulting distribution for the second (and main) liquid-argon calorimeter layer in ATLAS is shown in the lower plot of Fig. 39. The observed dispersion is in agreement with statistical fluctuations, i.e., no significant non-uniformity is seen at the per cent level. Similar results have been obtained by CMS where an intercalibration with cosmic muons (aligned to the crystal axis and with a reference energy of 250 MeV (MPV)) achieved an intercalibration of better than 1.5% in the barrel and better than 2.2% in the forward region. All 36 CMS crystal supermodules could thus be intercalibrated with cosmic muons, which was an important achievement because only 9 supermodules (25%) had been calibrated with electron test beam data prior to the calorimeter installation.

The reconstruction of jets and missing transverse energy requires the electromagnetic and hadronic

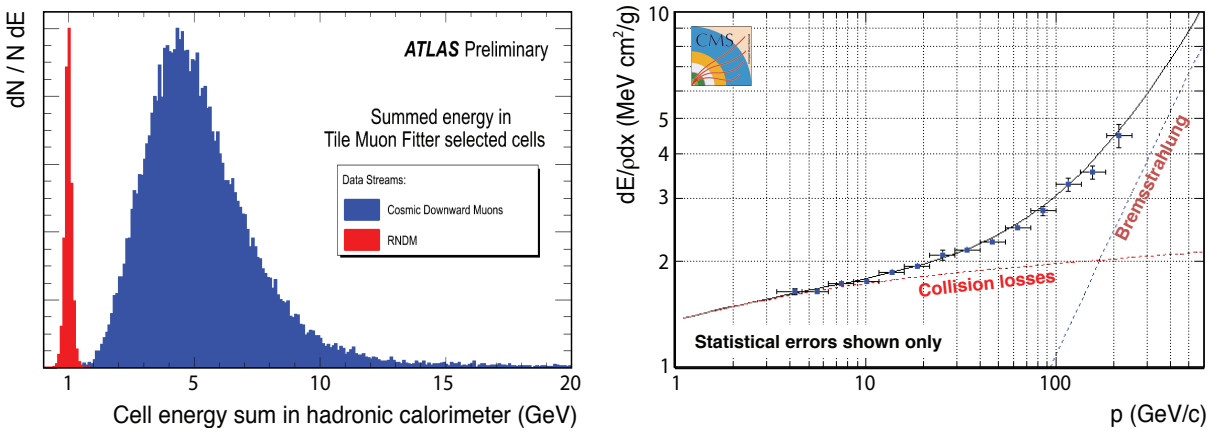


Fig. 38: Left: total energy sum of all cells along a muon track in the ATLAS hadronic calorimeter (blue) and the corresponding noise distribution obtained from randomly triggered events (red). The minimum ionising muon signal is well separated. Right: average energy deposits in the CMS electromagnetic calorimeter versus the muon momentum measured in the tracking devices. Overlaid is the expected energy loss for the lead-tungsten calorimeter. Indicated by the dotted lines are the contributions to the energy loss from collisions (red) and bremsstrahlung (blue).

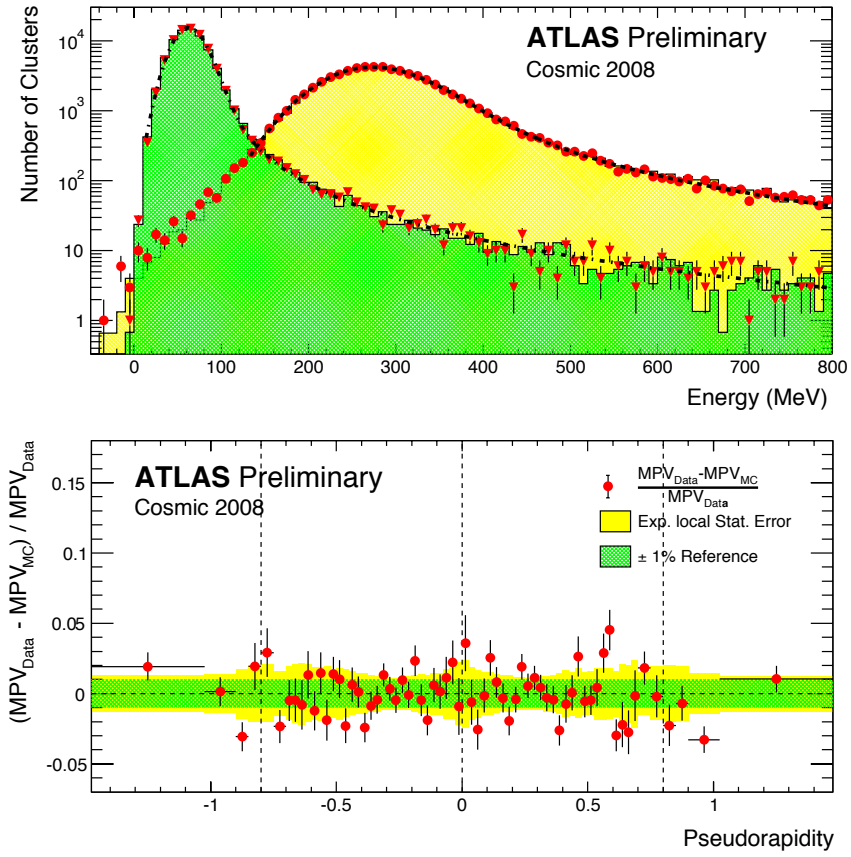


Fig. 39: Top: reconstructed cosmic muon energy in a 2×1 cluster in the first layer (dark shaded/green histogram for Monte Carlo and triangles for data) and in a 1×3 cluster in the second layer (light shaded/yellow histogram and dots for data) of the ATLAS electromagnetic calorimeter. Bottom: electromagnetic calorimeter energy response dispersion between data and Monte Carlo simulation versus the pseudorapidity, as measured with cosmic muons for the second (main) layer of the ATLAS electromagnetic calorimeter. The dark shaded (green) band indicates the $\pm1\%$ region for reference, and the light shaded (yellow) band indicates the expected statistical accuracy (1σ error band) of the measurement.

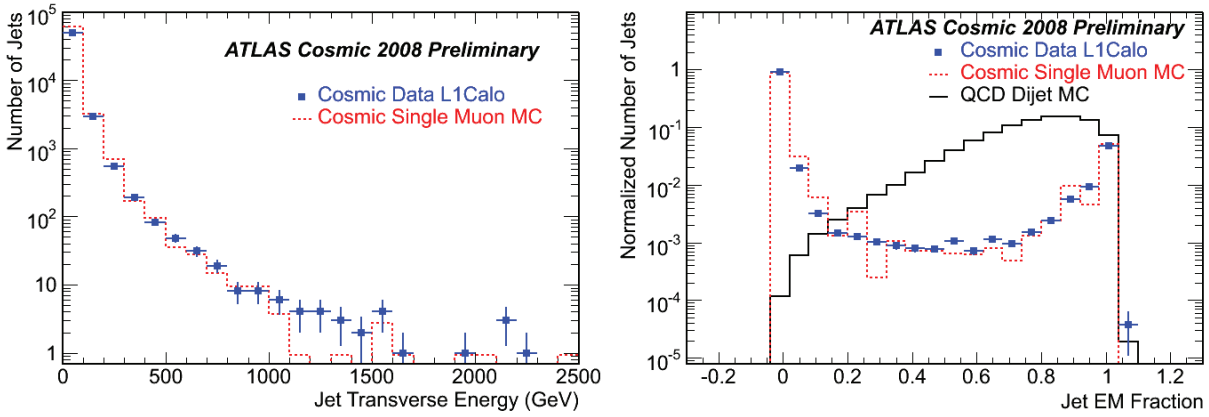


Fig. 40: Left: distribution of the jet energy for data (dots) and Monte Carlo simulation (dotted histogram). Only events with at least one jet that exceeds 20 GeV transverse energy are included. The Level-1 trigger inefficiency and cosmic air showers are not simulated. Right: electromagnetic fraction of jets for data (dots) and Monte Carlo (dotted), where the fraction is defined by the ratio of energy deposited in the electromagnetic calorimeter divided by the total deposited energy. The distributions are normalised to unity. Only jets with $E_T > 20$ GeV are included. Shown by the solid histogram is the expected distribution for QCD di-jet events as they originate from proton–proton collisions.

calorimeter responses to be combined. It can be studied with highly energetic cosmic muons releasing a Level-1 calorimeter trigger-accept signal. Jets from muon showers with energies exceeding the TeV scale are found in the data. Figure 40 (left) shows the distribution of the jet energy for calorimeter triggered events for data and Monte Carlo simulation. Because the simulated data do not include the Level-1 trigger inefficiency, the Monte Carlo distribution is normalised to data in the 100–300 GeV range. Only events that have a jet with $E_T > 20$ GeV are included in the figure. Good agreement between data and simulation is observed. A small excess at large transverse energy in data may be due to air-showers, not included in the simulation. The right-hand plot in Fig. 40 shows the electromagnetic (EM) fraction of jets for data and Monte Carlo, where the fraction is defined by the ratio of energy deposited in the electromagnetic calorimeter divided by the total deposited energy. The distributions are normalised to unity. As before, only jets with $E_T > 20$ GeV are included. Also shown for comparison is the distribution expected for QCD di-jet events as they originate from proton–proton collisions. The most likely value for the EM fraction is 0 or 1 for fake jets from cosmics, because the high energy deposition from photons originating from highly energetic muons will localise either in the electromagnetic or the hadronic calorimeter. QCD jets have a broad distribution of the EM fraction with a maximum at around 0.8. Electromagnetic fractions less than 0 or larger than 1 are due to small negative energy contributions from noise. One concludes from the plot that good separation between QCD jets and fake jets from cosmic rays can be obtained by vetoing jets with EM fractions close to 0 and 1.

7 Commissioning with single proton beam data

A lucky period between September 10 and 13, 2008, with — for the first time — single beams of 450 GeV LHC injection energy circulating in both directions of the LHC, gave the experiments the opportunity to commission the detector and the data taking chain with proton-beam background in synchronisation with the LHC clock. A single 'pilot' bunch containing approximately 3 billion protons — radio-frequency captured and not, with closed and open collimators, stably circulating or lost — travelled through the injection chain, transfer lines and finally the LHC. The single-beam exercise at injection energy was briefly repeated in 2009, at the restart of the LHC after an accident that caused a one-year delay in the commissioning and physics schedule.

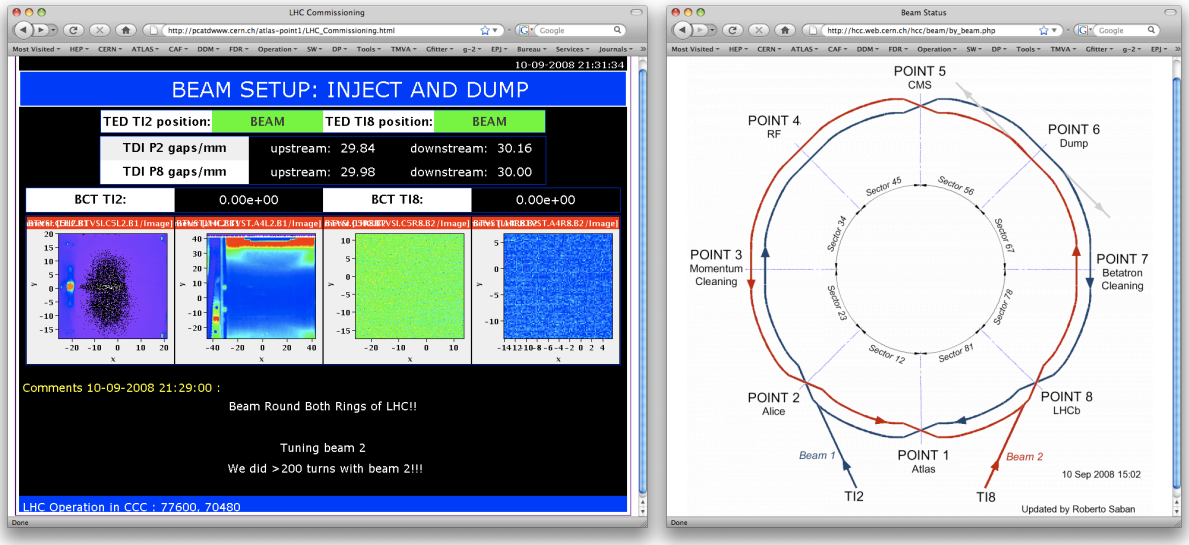


Fig. 42: Main LHC information displays sent from the CERN Control Room (CCC, 'Triple-C') to the experiments and the interested world. The left picture displays basic quantities such as the currents (in number of protons per bunch) passing through the two transfer lines serving to inject the LHC beam lines. Apart from displaying sometimes cryptic information displays and plots, it features useful operator comments on the bottom of the display: "Beam Round Both Rings of LHC !!" (one notices the capital letters and the abundant use of exclamation marks, which appropriately reflect the mood of the day). The right panel is a sketch of the two LHC beams. The colour codes are important: Beam Blue (1) must *always* be blue, and Beam Red (2) must *always* be red (source: Steve Myers, LHC coordinator). The detectors are located at four out of eight straight sections: Point 1 (ATLAS), Point 2 (ALICE), Point 5 (CMS) and Point 8 (LHCb). The remaining four straight sections serve beam acceleration, beam cleaning and dump purposes (see Section 2).

Figure 42 shows two of the most important information panels provided to the experiments (and the general public) by the LHC operators. One notices the particular location of Point 1 (ATLAS cavern) on the right panel: both beams need to make a full turn before reaching ATLAS. It was hence the last experiment to see beam, and it is affected by any problem along the beam line. A few photographs taken on 10 September in the LHC, ATLAS, CMS, and LHCb control rooms are shown in Fig. 43.

Fig. 41: The Google search page at 'Jour J' — the LHC start-up, 10 September 2009.

7.1 Beam-on-collimator events

Somewhat unexpectedly and all of the sudden, events where the entire detector was lit appeared on the event displays. A few typical events are collected in Fig. 44. The reaction in the ATLAS control room upon the arrival of the first event is witnessed by the photo in Fig. 47. *What happened?*

The events seen belong to so-called 'beam splash' type events, which originate from pilot-beam-on-collimator dumps. Collimators are placed at a distance of about 140 m on both sides of the experiments. If they are closed, the beam dumps on them, producing an avalanche of scattered particles that reach the detector. For such an event occurring every 42 seconds during a short period ATLAS typically recorded 300 000 silicon strip tracker hits (on lowered voltage for safety reason, reducing the hit efficiency; the pixel detector was switched off) and 350 000 transition radiation tracker hits, approximately all passing high-threshold discrimination. The sum of all calorimeter cells in these events exceeds 3000 TeV. Moreover 350 000 drift tube hits were recorded in the muon spectrometer and 320 000 (65 000)

Fig. 43: Snapshots taken on 10 September in the LHC (upper left), ATLAS (upper right), CMS (lower left), and LHCb (lower right) control rooms, exhibiting untypical occupancy.

muon trigger hits in the barrel (endcaps). Apart from being spectacular, beam splash events are useful in many ways for the experiments. Their main purpose is to serve timing-in the various detector parts and systems including the trigger with respect to each other. It is also interesting to correlate position and energy response in splash events, and to use them to identify dead channels. In the November 2009 beam splash period, after the LHC restart, it was also possible to exercise, for the first time in realistic conditions, the ATLAS standalone beam abort system using the diamond Beam Condition Monitor (BCM) detectors. By lowering the abort thresholds, a deliberate BCM beam abort was triggered by a beam splash event reaching ATLAS. No fake abort was observed. Beam splash events have also been observed in the forward detectors of the experiments, designed to measure the relative luminosity. In total, ATLAS recorded about 70 beam splash events (of a total of approximately 100 delivered) in September 2008, and another 106 events (all triggered) in November 2009. CMS received and recorded an order of magnitude more beam splash events.

An example for a timing study is given in Fig. 45. Shown in the left plot is the mean hit time (expected minus measured) versus the endcap disk, where a larger absolute number corresponds to a larger absolute pseudorapidity. The measurement corresponding to a single beam-splash event is shown. The event arrives from the A-side ($+z$ side) so that the hit time behaves as expected for a collision event for the far side (C-side), but wrongly for the A-side with respect to the expected collision timing used in the event reconstruction (the event comes from behind and the hit time is thus anticipated). A similar behaviour is observed for all other detector systems.

Beam splash events from both sides can be used to adjust the timing for both far sides. The right plot in Fig. 45 shows the mean time residual along the z coordinate of all ATLAS muon drift tube chambers using the synchronous front of splash particles and the very large particle flux. A linear relation

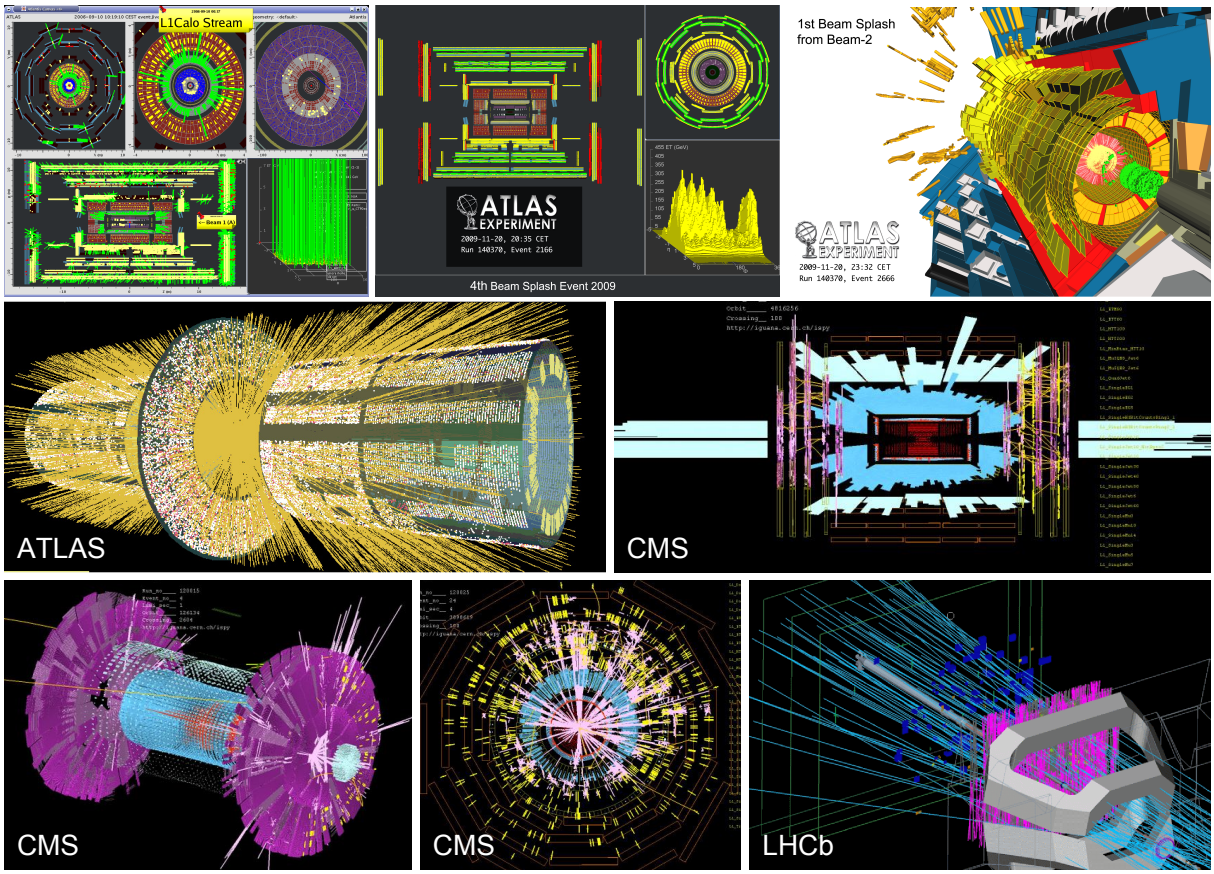


Fig. 44: Event displays of beam–on–collimator 'splash' events recorded by ATLAS (upper plots and centre left), CMS (centre right, lower left and middle), and LHCb (lower right).

is found with a slope determined by the speed of light. A timing study with beam splash events in the CMS hadronic calorimeter is shown in Fig. 46. Drawn are the differences between reconstructed and expected cell times for beam splash events before (left panel) and after timing adjustment (right) using previously measured beam splash events. The large deviations from zero in the left panel are due to collision time settings. CMS also correlated the energy deposits in the hadronic and electromagnetic calorimeters for beam splash events, reproducing nicely the expected linear dependence and a relative coefficient of $E_{\mathrm{HCAL}} \simeq 6.5 \cdot E_{\mathrm{ECAL}}$.

7.2 Beam background events

After the beam splash events, the collimators were all opened allowing the beam to circulate in the LHC and to pass by the experiments. Beam passages without interactions are measured primarily in the beam pickup detectors based on electrostatic current induction. These detectors are installed ± 175 m away from the interaction points of the experiments (many more such beam pickups are installed along the LHC for beam monitoring purposes). They provide input signals to the Level-1 triggers, indicating filled LHC bunches, and also a time reference for the detector systems. In case of beam collisions, the coincidence of signals in the two beam-pickup detectors can be used to identify colliding bunches and, more importantly, their timing difference (measured by an oscilloscope) can be used as input to the beam

Fig. 47: A 'beam splash' event being seen in the ATLAS control room.

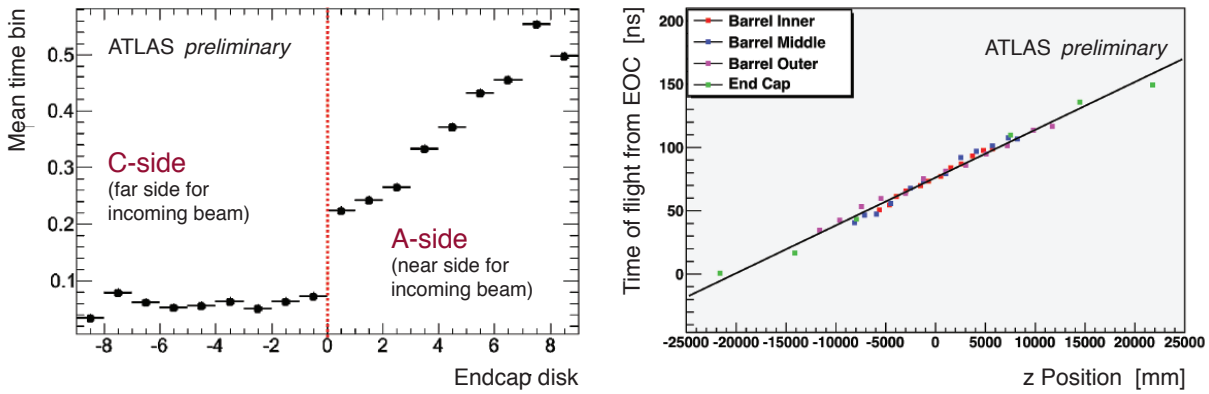


Fig. 45: Left: timing properties of a single beam-splash event originating from the A-side in the ATLAS silicon strip tracker (see text). Right: time residual versus the z coordinate along the ATLAS muon drift tube chambers for a beam splash event. The slope is determined by the speed of light.

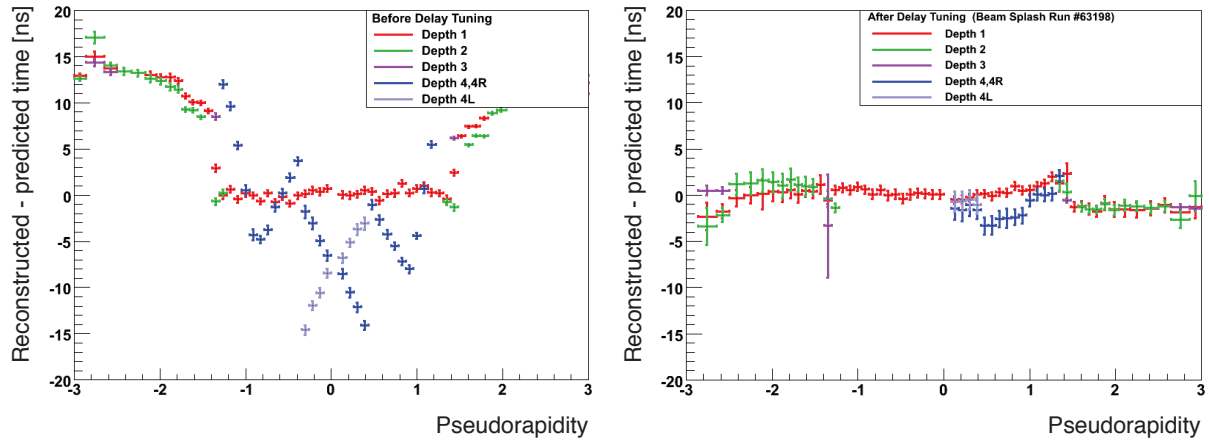


Fig. 46: Difference between reconstructed and expected cell time versus the pseudorapidity for beam splash events in the various layers and geometrical regions of the CMS hadronic calorimeter. Left is uncorrected assuming collision timing, and right is after correction with the use of previously observed events.

'cogging', that is a relative radio-frequency phase adjustment of the bunches to ensure collisions in the interaction point ($z = 0$) without longitudinal shift. In the Level-1 trigger the beam pickup signals are put in coincidence with the other triggers to reduce background from cosmic rays. This requires, however, a proper timing-in of the various trigger signals.

Circulating single-beam bunches can also provide beam-related background particles that are measured by the experiments. At low beam intensities, there are two main sources of beam backgrounds referred to as 'beam–gas interactions', which are interactions of beam particles with residual gas in the beam pipe or with the beam pipe wall. Via the decay of pions such a process also produces muons, which travel with the proton beam in what is called the 'beam halo' (usually referred to as 'beam-halo background', which is what seems to be the primary single-beam background seen so far in the detectors). Such beam related backgrounds originating from fixed-target collisions are strongly boosted in the forward direction. Figure 48 shows the distributions of the track polar angle with respect to the beam axis for single-beam data, simulated beam-halo background, and cosmic ray events taken with no beam present in the LHC. Whereas the beam background peaks at small angles, cosmic ray tracks peak at larger values, which are, however, much below the $\sim$1.5 rad that would be expected, because a forward trigger has been used to select these events. The orange shaded histogram shows the distribution of single-beam

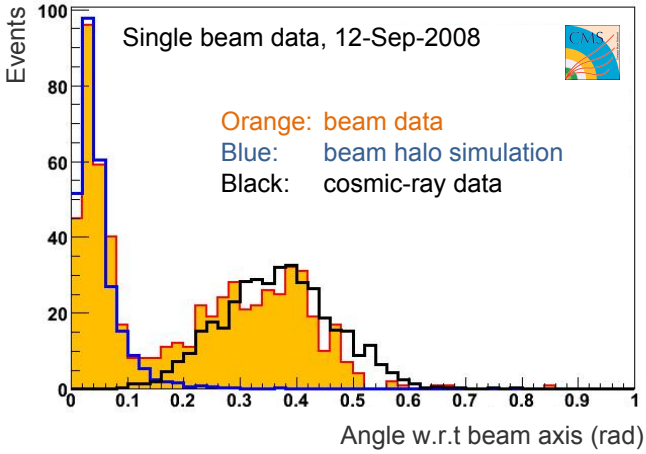


Fig. 48: Distributions of the track polar angle with respect to the beam axis obtained by CMS for single-beam data (orange shaded), beam-halo background simulation (blue line), and cosmic ray data with no beam (black line).

events accepted by the same trigger. One clearly distinguishes the beam-related background from the cosmic muon contamination.

Event displays of beam background events with halo muons taken by ATLAS and CMS are shown in Fig. 49. In ATLAS the toroidal magnetic fields in the muon spectrometer bend the muon tracks longitudinally in the z coordinate.

7.3 Radio-frequency bunch capture

After injection into the LHC, the protons in a bunch start to spread longitudinally and transversely due to their mutual repulsion. Within milliseconds the bunch thus 'debunches'.[26] Debunching can be directly observed by the experiments via a decaying beam pickup signal during circulating beam. An example for this is displayed in the upper plot of Fig. 50 showing the beam pickup signal amplitude in volts versus the time in nanoseconds as measured by ATLAS. The spikes represent the induced signal when a bunch passes nearby an ATLAS beam pickup detector. The time difference between adjacent spikes amounts to 89 μs, which corresponds to an LHC revolution period. The signal weakens while the bunch disintegrates. The lower panel of Fig. 50, sketches the radio-frequency field bucket structure of the LHC. A bunch filled with protons is captured within a bucket of 2.5 ns length (precisely: 2.495 ns, i.e., a radio frequency of 400.79 MHz).[27] Only every tenth bucket is filled providing the design bunch period of 25 ns.

Figure 51 shows a series of attempts in September 2008 to capture a bunch in the LHC within a radio-frequency bucket. The horizontal lines represent a measured beam pickup signal after 10 LHC turns. The leftmost plot shows the decaying bunch in absence of a radio-frequency (RF) field. The signal induction from the debunched beam becomes unmeasurable after 250 turns. The centre-left plot shows a first capture attempt, at a wrong injection phase, so that the bunch is split into two by the RF field, leading to a fast decay. For the centre-right plot the injection phase has been improved, but is still shifted with respect to the RF phase, leading to a moving proton package and a fast decay. Finally, the rightmost

[26] Debunching can also be useful. For example, controlled debunching and rebunching can be used to split and multiply bunches in the injection chain of an accelerator. This is, however, a delicate technique which is not used for bunch splitting in the LHC injector (PS).

[27] The radio-frequency electrical field together with the relativistic contraction provide a stronger longitudinal constraint on the bunch size than the bucket length. At 450 GeV beam energy, the longitudinal RMS of the bunch is expected to be around 8 cm, (the measured values were found to be significantly lower than that) decreasing to approximately 6 cm at 7 TeV design energy.

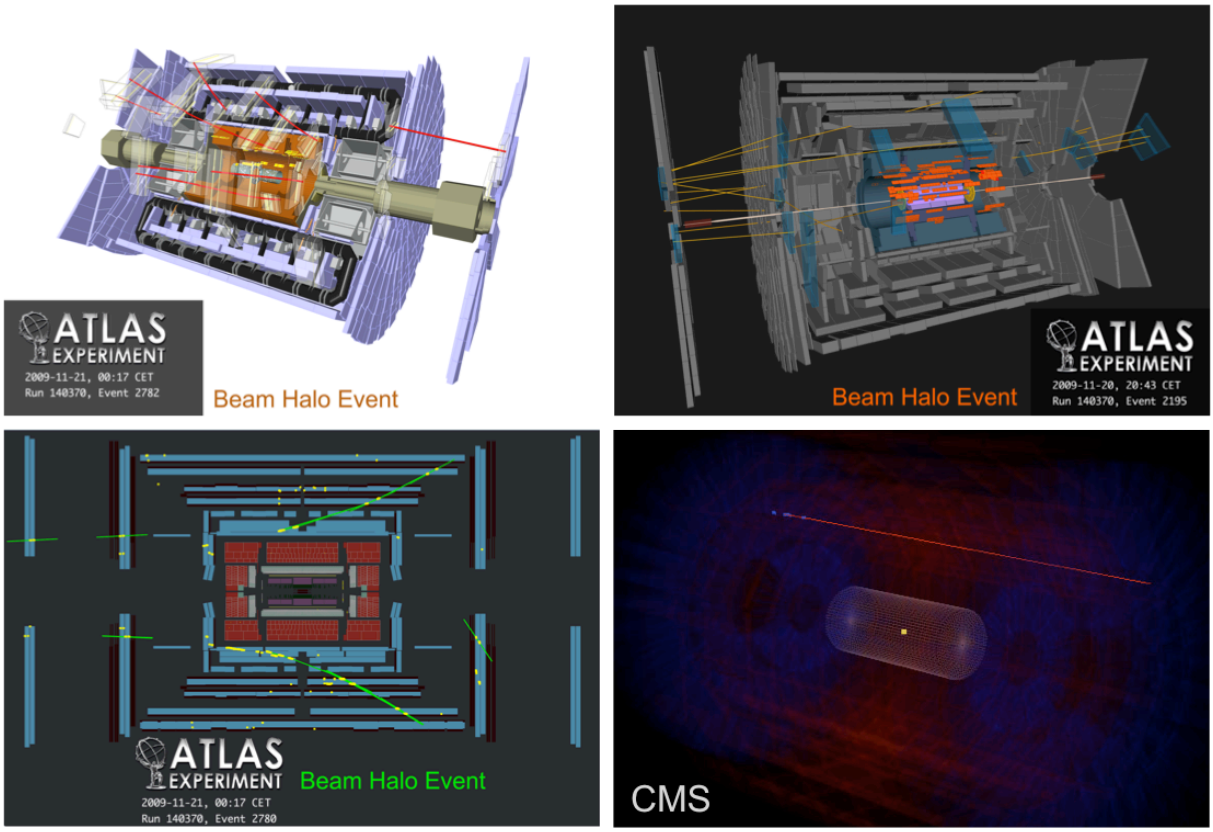


Fig. 49: Beam-related background events with halo muons taken by ATLAS and CMS (lower right) in November 2009.

plot shows an accurate injection phase and a properly captured bunch. No decay of the signal due to limited lifetime can be noticed.

Since the experiments record events triggered by the beam pickup signals and, by running synchronously with the LHC clock, also store the bunch crossing number that led to the trigger accept, it is possible to measure the beam debunching *and* its capture in an RF bucket. Such a measurement has been performed by CMS and the result is shown in Fig. 52. Before the RF capture the bunch crossing number of the triggered events is spread over many bunches. After successful RF capture all triggered events have the same bunch crossing number 831 as seen by the spike in the distribution at that point.

8 Early physics at the LHC — Overview

The major part of the LHC proton–proton physics programme can be grouped under the following grand themes.

1. **Mass** — search for the Higgs Boson.
2. **Electroweak unification** — precision measurements (W and top masses) and tests of the Standard Model.
3. **Hierarchy in the TeV domain** — search for supersymmetry, extra dimensions, new symmetries in the TeV domain, and other exotic phenomena.
4. **Flavour** — B meson mixing, rare decays and CP violation as tests of the Standard Model.

This programme is also reflected in the ATLAS and CMS physics organisation, separated into so-called 'physics objects groups' (CMS) or 'combined performance groups' (ATLAS), and 'physics analysis

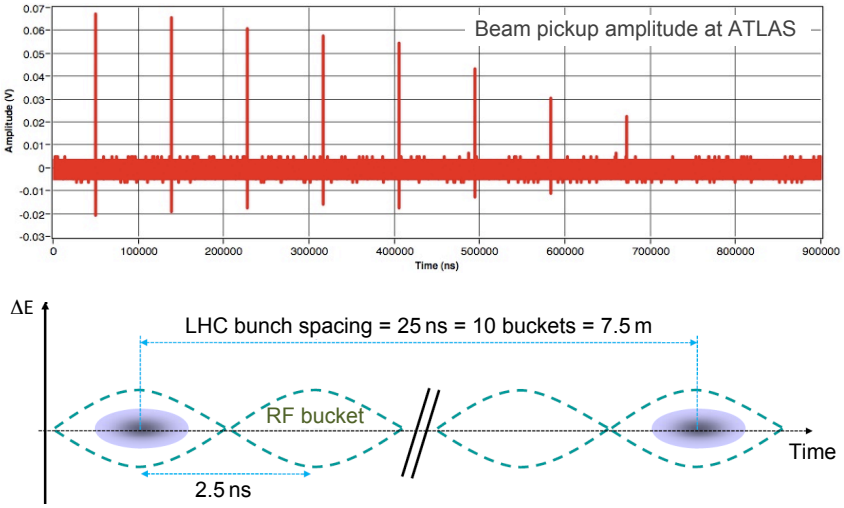


Fig. 50: Top: decaying circulating beam signal in an ATLAS beam pickup detector due to beam debunching. Bottom: bucket and bunch structure in the LHC.

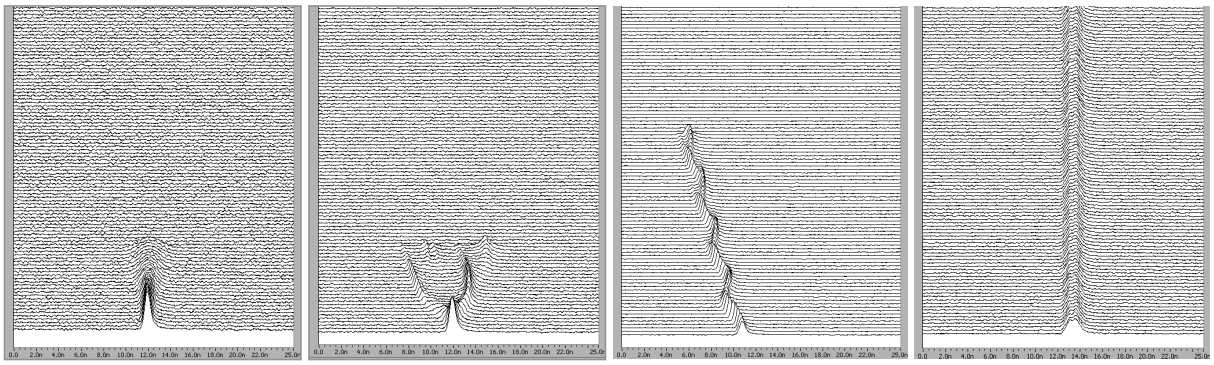


Fig. 51: Attempts and successful (rightmost plot) radio-frequency capture of a bunch in the LHC. Each horizontal line represents 10 LHC turns. See text for a discussion of the plots.

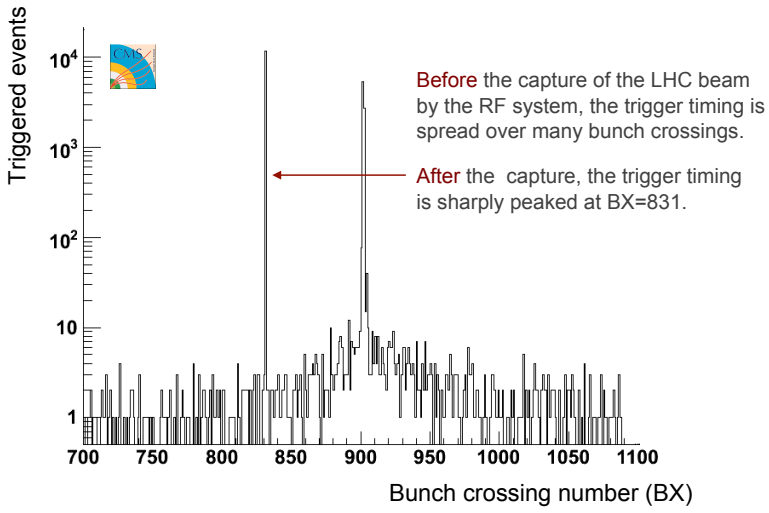
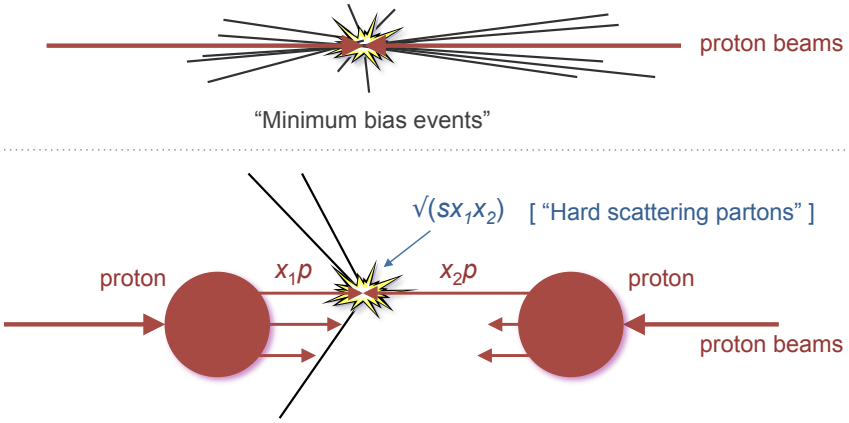


Fig. 52: LHC bunch decay and radio-frequency capture as measured by CMS.

groups'. The former groups provide the reconstruction of the objects that combine various detector systems and that are common input for physics analysis. They are subdivided into 'e/gamma', 'jets/missing transverse energy', 'hadronic tau', 'muons' and 'flavour tagging' groups. The physics groups are organised in 'Standard Model' containing QCD, electroweak and diffraction physics, 'B physics', 'Top', 'Higgs', 'SUSY', 'Exotics', 'Heavy ions', 'Luminosity', and 'Monte Carlo generators' subgroups.

Since protons are made out of quark and gluon constituents ('partons'), collisions of protons are complex scattering processes involving elastic, diffractive (single and double), inelastic non-diffractive and central diffractive interactions (pomeron–pomeron scattering). The large majority of the proton–proton events are due to interactions at large distances. The inclusive sum of single and double diffractive, and non-diffractive processes are called 'minimum bias' events, in allusion to lowest transverse momentum events that can be selected by a trigger, and in contrast to 'zero-bias events', which can only be obtained if all events or a random sample of events are selected. The total minimum bias cross section at 14 TeV centre-of-mass energy at the LHC is approximately 70 mb. It dominates by orders of magnitude the primary physics channels of interest. Minimum bias events are characterised by tracks with small transverse momenta of $\langle p_T \rangle = 0.5$ GeV on average.



The constituents of the protons participating in the interaction carry only a fraction of the proton's momentum. The fraction is governed by parton distribution functions that cannot be predicted from first principles and are taken from experiment. The complexity of describing proton–proton interactions includes, besides the hard scattering as described by parton-level perturbative QCD, the parton distribution functions of the proton, the underlying event (describing the possibility of multiple parton interactions in the same proton–proton collision), initial- and final-state radiation, the definition of jets, and the minimum bias event properties.

Figure 53 (taken from Ref. [23]) illustrates the structure of a proton–proton collision event as it occurs in the LHC. Hard subprocesses between partons need to be convolved with parton densities, the decays of the hard subprocesses, initial- and final-state radiation, and multiple parton interactions (and their initial- and final-state radiation), as well as beam remnants and other outgoing partons (not shown) to arrive at a realistic description. All parton-level processes are connected through colour confinement, leading to a primary hadronisation, with many primary hadrons being unstable and further decaying.

To reconstruct such an event in ATLAS or CMS it first needs to be triggered, i.e., the event must pass several trigger levels with increasing rejection power. Once accepted, the event is written to disk and promptly reconstructed on large offline computer farms comprising several thousand central processing units. The reconstruction program reconstructs tracks of charged particles in the inner tracker and the muon systems, electromagnetic clusters in the electromagnetic calorimeter, hadronic clusters and jets in the combined electromagnetic and hadronic calorimeters, missing transverse energy in the calorimeters, and identifies particles and objects: muons, electrons, photons, taus, jets, and heavy quark flavour. All these steps in the reconstruction chain involve tremendous challenges regarding efficiency, purity, accu-

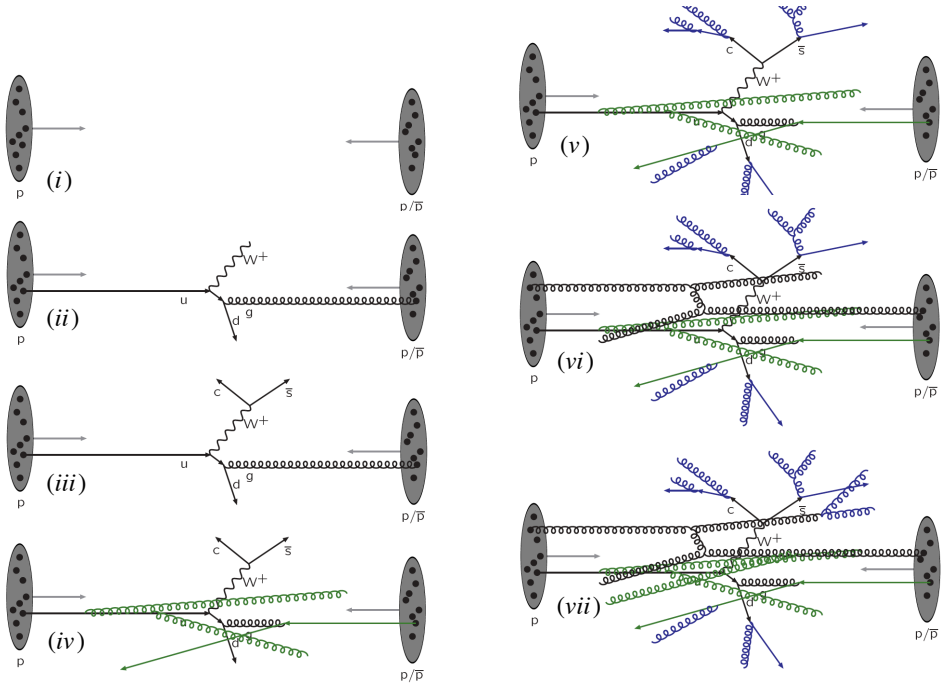


Fig. 53: Schematic Feynman graphs for proton–proton collisions corresponding to: (*i*) incoming proton beams: parton distributions; (*ii*) hard subprocess: described by matrix elements; (*iii*) resonance decays: correlated with hard subprocess; (*iv*) initial-state radiation: spacelike parton showers; (*v*) final-state radiation: timelike parton showers; (*vi*) multiple parton–parton interactions; (*vii*) multiple parton–parton interactions with its initial- and final-state radiation. Pictures and legend taken from Ref. [23].

racy and resolution (calibration). The extensive commissioning work performed by the experiments will surely pay off when analysing the first collision data and comparing them with Monte Carlo simulations.

With increasing statistics data-driven analysis and calibration methods will take over and the experiments will achieve the performance they have been designed for.

After the reconstruction of the primary physics objects, the events are selected according to topological criteria that characterise the physics channel of interest. *Inclusive analyses* count events with leptons, photons, jets or missing transverse energy. For example, a QCD analysis may select events with high-energetic (or many) jets. A combined QCD and electroweak analysis may select events with leptons or photons in the final state. A search for supersymmetry with R-parity conservation will select events with large missing transverse energy, and may also require leptons to reduce the contamination from Standard Model QCD events. *Exclusive analyses* kinematically combine reconstructed objects. For example, an analysis using $W \to \mu\nu$ decays will identify a muon and compute the transverse W mass using the muon momentum and the transverse missing energy vector. To select top–antitop events, where, for example, one top decays to $be\nu$ and the other to bqq, one must identify the electron and two b-jets, and compute the top mass from the invariant mass of one of the b jets and two hard light-quark jets, which originate from a W decay. To identify Higgs decays into two photons one must identify two photons in the event and compute their invariant mass, which needs to accumulate at the same value within the experimental errors to create a significant Higgs signal over backgrounds from random two-photon or misidentified photon-jet combinations. Similarly, to search for Higgs decays into two electrons and two muons, once must identify the corresponding leptons and compute their invariant mass. Intermediate on-shell resonances with known mass can be used as additional kinematic constraints. Finally, to search for new high-mass resonances such as Kaluza–Klein graviton states decaying into lepton pairs, as predicted in models with extra spatial dimensions, one must identify the leptons and compute their mass to obtain a signal over the dominant Drell–Yang di-lepton background. For many of these analyses it is beneficial to combine all the available object-level and event-level information using multivariate statistical pattern recognition techniques.

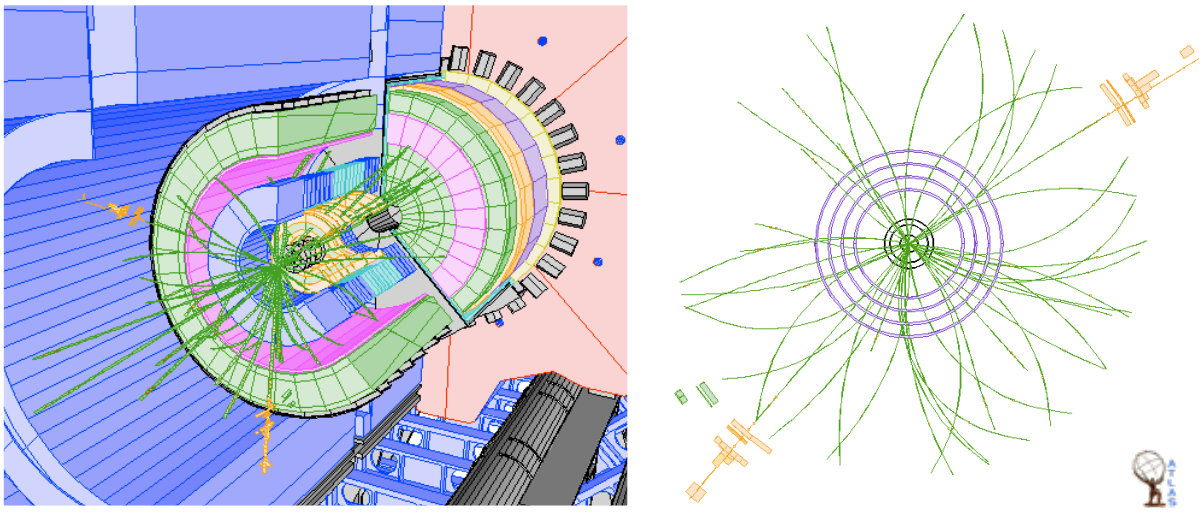


Fig. 54: Event display of a simulated $Z \to ee$ event in ATLAS. The final-state electrons have tracks in the inner tracker and large energy depositions in the electromagnetic calorimeter. Their invariant mass is consistent with that of a Z boson.

9 Physics commissioning

With emphasis at the beginning of the collision data taking, but also throughout the whole lifetime of the experiments, physics commissioning such as the calibration and alignment of detector systems and physics objects, as well as the data-driven ('*in-situ*') measurement of efficiencies, purities, calibration biases and resolutions, will represent a large part of the experimental work. We discuss in the following the *in-situ* calibration of the electromagnetic calorimeter, the determination of material in the inner tracking detector, and jet and missing transverse energy calibration and reconstruction.

9.1 *In-situ* electromagnetic calorimeter calibration

Among the primary measurements driving the performance requirements for the ATLAS and CMS electromagnetic calorimeters is the search for $H \to \gamma\gamma$. Since this channel is important at low Higgs mass where the intrinsic width of the Higgs is negligible,[28] the measured width of the di-photon invariant mass, and hence the sensitivity for discovery, will be determined by the energy resolution of the electromagnetic calorimeter. We have already mentioned the importance of the constant term in the calorimeter energy resolution for Higgs searches in Footnote 17. We can extend this by a back-of-the-envelope exercise. Let us consider a data sample for an integrated luminosity of $20\,\text{fb}^{-1}$ containing 690 $H \to \gamma\gamma$ and $\sim$170 000 background events with di-photon invariant mass $110 < m_{\gamma\gamma} < 150\,\text{GeV}$. With the nominal (design) ATLAS electromagnetic calorimeter resolution, assuming a constant term of 0.7%, a fit to the di-photon mass would yield a signal significance of 2.9σ. Worse constant terms of 1.0% or even 2.0% would reduce this significance to 2.4σ and 1.8σ, respectively.[29]

It is hence mandatory to keep the constant term, originating from non-uniformities in the calorimeter response due to inhomogeneities and non-linearities, as small as possible by *intercalibrating* the calorimeter with physics events. Calorimeter intercalibration (which is *not* absolute scale calibration) can be performed with any physics events that provide a predicted or smooth energy deposition.

The most favourable channel for *in-situ* intercalibration is $Z \to ee$. The Z mass being precisely

[28] A Higgs of mass 120 GeV has an intrinsic width of 4 MeV, while at 200 GeV the Higgs has a width of 1.4 GeV due mainly to the opening of the di-weak-boson channels.

[29] Note that this test assumes the simplest possible $H \to \gamma\gamma$ analysis approach. A more sophisticated fit using more discriminating variables and detector-specific 'categories' boosts the fit performance significantly.

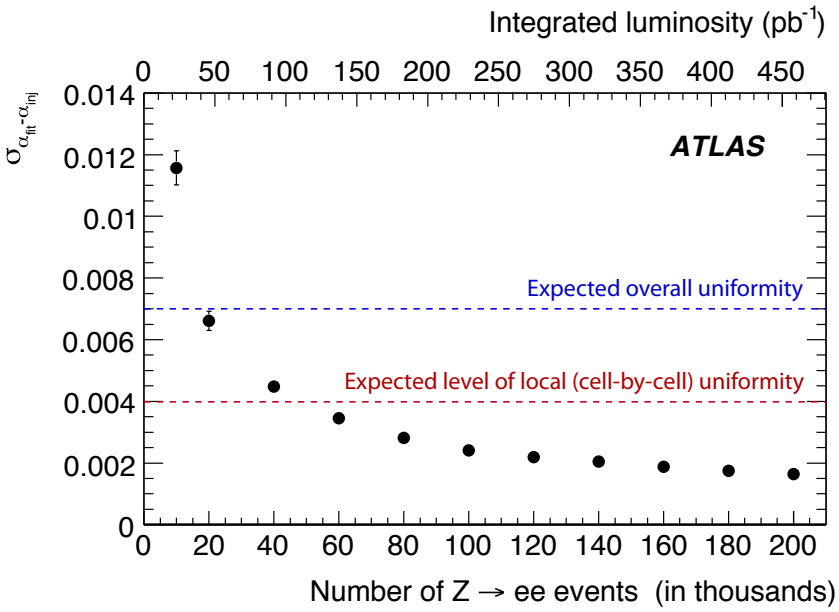


Fig. 55: Statistical yield of the $Z \to ee$ electromagnetic calorimeter intercalibration in ATLAS. Shown is the expected accuracy achieved for the constant term versus the number of events used in the intercalibration fit. The corresponding integrated luminosity is given on the upper abscissa.

measured at LEP to (91.1875 ± 0.0021) GeV, the average reconstructed di-electron mass in the detector after calibration must reproduce it (per event, the detector resolution and the natural width of 2.5 GeV will lead to a natural smearing). With sufficient statistics, the mass-constrained intercalibration can be done per geometrical detector units, which are suitably chosen regions in pseudorapidity and azimuth, typically $\Delta\eta \times \Delta\phi = 0.2 \times 0.4$. ($Z \to ee$ decays also allow one to calibrate the absolute energy scale, which is required to be known at the per mil level or less for most analyses, and should be at the 0.02% level for the high-precision W mass measurement.) For a given intercalibration region i, it is assumed that long-range non-uniformities, encoded in a parameter α_i, have modified the measured electron energy as $E_i^{\text{reco}} = E_i^{\text{true}} \cdot (1 + \alpha_i)$. Neglecting correlations between the electrons and postulating that the opening angle between the two electrons is correctly measured on average, the effect on the di-electron invariant mass is $M_{ij}^{\text{reco}} = M_{ij}^{\text{true}}(1 + (\alpha_i + \alpha_j)/2)$. The α_i can be extracted from a maximum-likelihood fit to $Z \to ee$ candidates, which must also incorporate a background component from events other than $Z \to ee$.

Figure 54 shows the event display of a simulated $Z \to ee$ event in ATLAS. The electrons leave large energy deposits in the electromagnetic calorimeter and their invariant mass is consistent with that of a Z boson. Approximately 10 000 of these events (and approximately 10 times more $W \to e\nu$) will be recorded in $10\,\text{pb}^{-1}$ integrated luminosity (reconstruction efficiency not subtracted). Figure 55 depicts the expected statistical yield of the $Z \to ee$ electromagnetic calorimeter intercalibration in ATLAS. Shown is the expected accuracy achieved for the constant term versus the number of events used in the intercalibration fit. Indicated by the dashed horizontal lines are the design value of 0.7% for the constant term and the level of the local non-uniformity from cell-by-cell variations, estimated to be 0.4%. Design calibration performance is expected to be reached with $20\,\text{pb}^{-1}$ integrated luminosity.

9.2 Inner detector material mapping

The high-precision and redundant inner tracking systems of ATLAS and CMS come at the price of a significant amount of material the particles must traverse. Figure 57 shows the material in the inner tracking system of ATLAS (left) and CMS (right) in terms of radiation lengths. It is remarkable that only a small part of it stems from active detector material, whereas the main contributions are due to services.

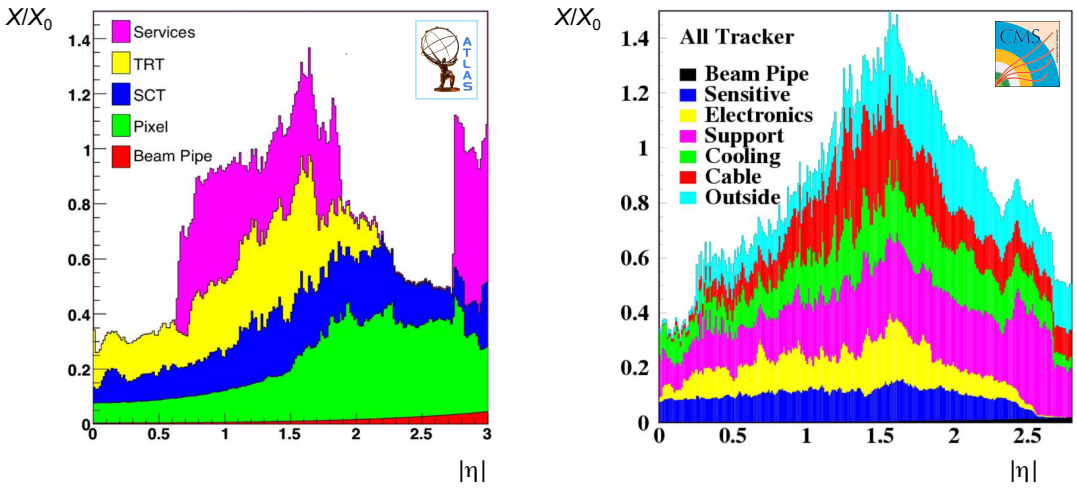


Fig. 57: Material in the inner tracking system of ATLAS (left) and CMS (right) in terms of numbers of radiation lengths X/X_0. Solenoid and calorimeter cryostat add roughly $2X_0$ before the electromagnetic calorimeter presampler in ATLAS.

The amount of radiation lengths in these services needed to be systematically reevaluated throughout the planning and construction phases of both detectors. While the technical proposals in 1994 estimated about (in units of X_0) 0.2 (0.6) at $\eta = 0$ ($\eta = 1.7$ corresponding to about 20° polar angle) for both ATLAS and CMS, it became 0.2 (1.5 for ATLAS and 0.85 for CMS) at the time of the TDRs in 1997, to finally converge to 0.3 (1.3 for ATLAS and 1.5 for CMS) at the time of the construction in 2006. Note that in ATLAS objects need to traverse approximately an additional $2X_0$ before reaching the presampler (available for $|\eta| < 1.8$), and roughly another X_0 before the electromagnetic calorimeter.

A good understanding and simulation of the inner detector material is crucial for precision measurements such as the W mass, where the accurate calibration at the Z mass needs to be transferred to the W mass using Monte Carlo simulation. Many other physics analyses benefit from a precise material mapping. The best method to perform a radiography of the inner tracking detector is to use photon-to-electron–positron-pair conversion, which occurs only in the vicinity of a nucleus that recoils against the e^+e^- system and thus ensures momentum conservation (*cf.* Fig. 56). The conversion needs to happen not too far from the interaction point so that sufficient tracking layers remain to reconstruct the electron and positron tracks and their common vertex position, which indicates matter. A photon-conversion-based radiography of the ATLAS inner tracking detector, obtained from Monte Carlo simulation, which implements a detailed modelling of the active and passive components, is shown in Fig. 58. The photons originate from π^0 and η decays, and Monte Carlo truth information has been used for the conversion vertices (the measured conversion map will look quite different).

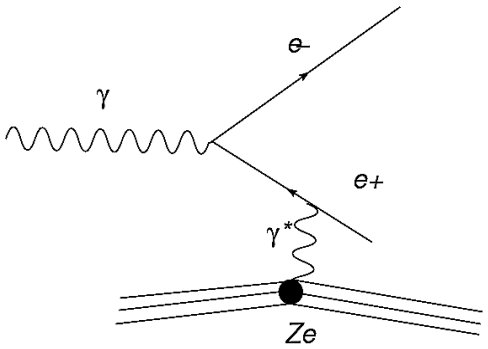


Fig. 56: Feynman diagram for the conversion of a photon to an electron–positron pair in presence of a nucleus.

9.3 Efficiency determination with the tag-and-probe method

Decays of Z bosons to leptons can also be exploited to measure trigger selection and offline reconstruction efficiencies from data. The primary method used for this is denoted ‘tag-and-probe method’, the

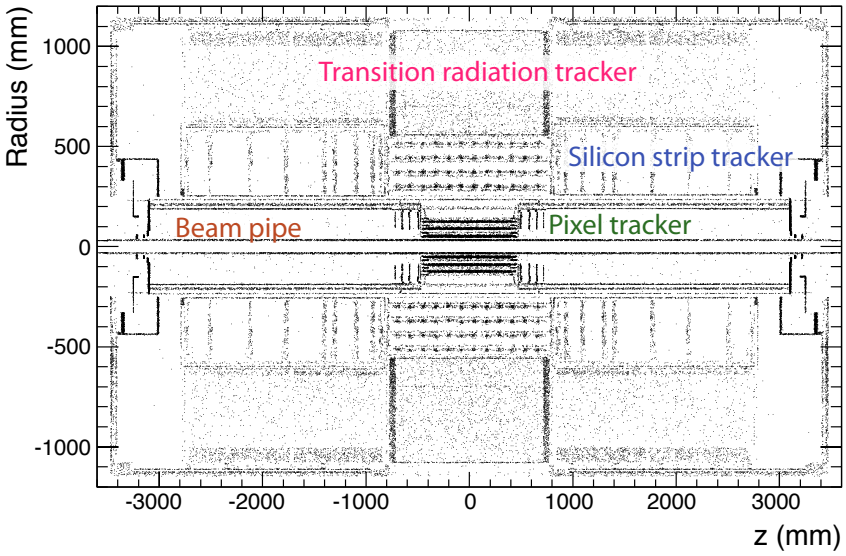


Fig. 58: Mapping of photon to electron–positron conversions as a function of z and radius, integrated over the azimuth angle, for the ATLAS inner tracking detector. The mapping has been made from 500 000 simulated minimum bias events (∼40 minutes of data taking at 200 Hz output rate), using ∼90 000 conversion electrons of transverse momentum larger than 0.5 GeV, originating from photons from π^0 and η decays. Monte Carlo truth information is used for the conversion vertices. The plot shown does not represent the latest version of the ATLAS detector description. In particular the beam condition monitor stations located at $z = \pm 1840$ mm are not yet included.

principle of which is straightforward (see sketch in Fig. 59). Let us consider the example of determining the reconstruction efficiency of muons in the muon system using $Z \to \mu\mu$ candidate events.[30]

The candidate event has been triggered by the 'tag muon', which is a 'golden' muon candidate with an isolated track from combined inner tracker and muon system reconstruction, and transverse momentum larger than 20 GeV. The probe muon is another muon candidate, which is independent of the tag-muon selection. To find the candidate we require a track reconstructed in the inner tracker and an invariant mass of tag and probe muons consistent with that of a Z boson. We now count how often the probe muon has been reconstructed in the muon spectrometer. With sufficient statistics the efficiency of the probe muon reconstruction can be evaluated in bins of p_T, η and ϕ. Usually, the result has to be corrected for combinatorial background under the Z peak. The most powerful approach combines background and efficiency determination in all regions within a single unbinned maximum-likelihood fit. The tag-and-probe method is very flexible, and many versions of the same idea exist. Figure 60 shows an event display of a simulated $Z \to \mu\mu$ event in ATLAS. The minimum ionising muon tracks traverse the calorimeters and leave measured hits in the muon spectrometer. Approximately 10 000 of these events (and approximately 10 times more $W \to \mu\nu$) will be recorded in 10 pb^{-1} integrated luminosity (reconstruction efficiency not subtracted).

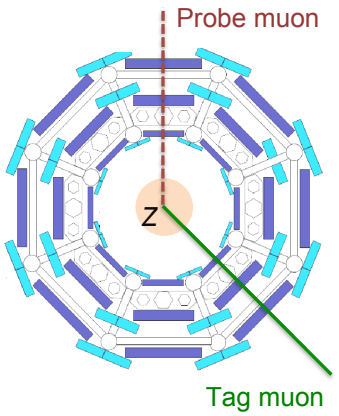


Fig. 59: Sketch illustrating the tag-and-probe method.

[30] The expression 'candidate' refers to the fact that for real data we do not know whether a reconstructed $Z \to \mu\mu$ candidate is indeed the process we believe it to be, or whether it is background from random combinations of muons ('combinatorial background') or objects faking muons. Only a statistical analysis allows us to separate signal from irreducible background.

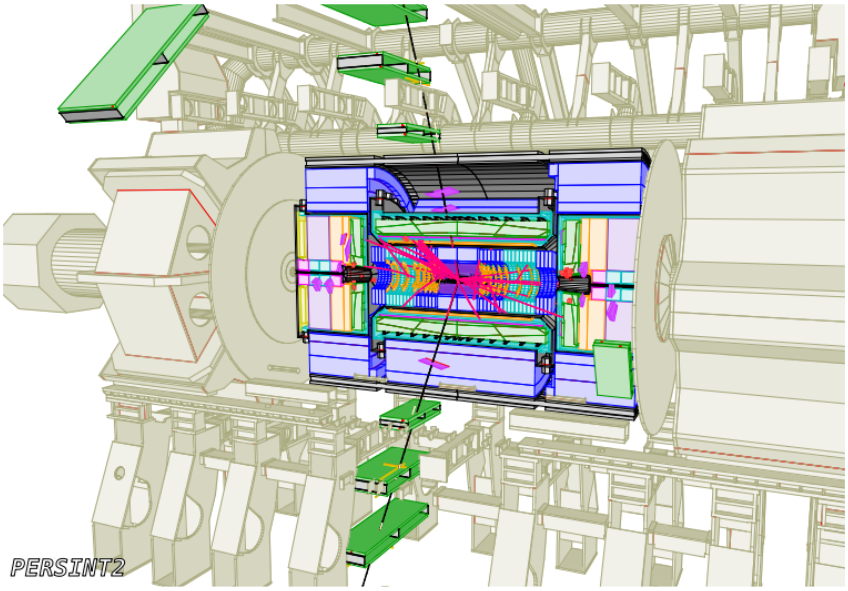


Fig. 60: Event display of a simulated $Z \to \mu\mu$ event in ATLAS. The final-state muons are measured in the muons spectrometer.

9.4 Jet calibration

A precise knowledge of the absolute jet energy scale (JES) is needed by many physics analyses. Typically a calibration of better than 1% is required for the measurement of the top-quark mass, but also for supersymmetry signatures. Jets are complex phenomenological objects, and their reconstruction involves a large number of corrections and calibrations. Only a brief overview is given here.

The jet energy reconstruction and calibration can be divided in four steps:

1. Calorimeter tower or cluster reconstruction.
2. Jet forming (cone, k_t, anti-k_t or other 'jet algorithms').
3. Jet calibration from calorimeter to the particle scale.
4. Jet calibration from particle to the parton scale.

The discussion here concentrates on jet calibration, assuming jets have been formed by an algorithm with suitable experimental and theoretical properties for the physics measurement under study.

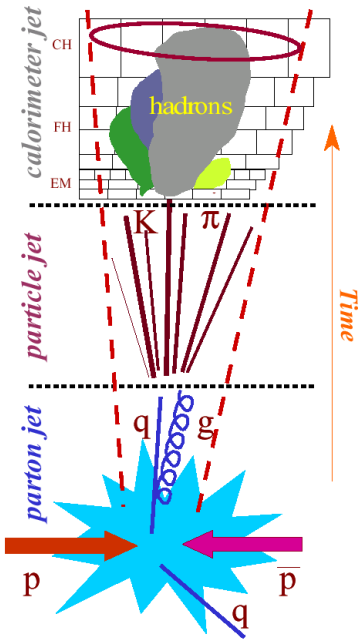


Fig. 61: Illustration of the various jet reconstruction levels from partons over hadrons to the calorimeter.

Several and conceptually quite different calibration approaches are considered by the experiments. Monte Carlo based jet calibrations, transforming the electromagnetic energy scale to the hadronic scale, can be distinguished according to the level of detail with which the jet constituents are treated and separately corrected. The 'global jet calibration' uses as input clusters that have been properly calibrated at the electromagnetic scale, and which are matched in energy to the Monte Carlo truth particle jet for bins of E_T and η. This calibration returns the jet energy at the hadronic scale (*cf.* sketch in Fig. 61). On the contrary, the 'local hadron calibration' calibrates clusters independently of the jet algorithm by making an assumption on their electromagnetic or non-electromagnetic nature. Jets are then formed out of calibrated clusters, and the jet energy is given at the hadronic scale. Finally, *in-*

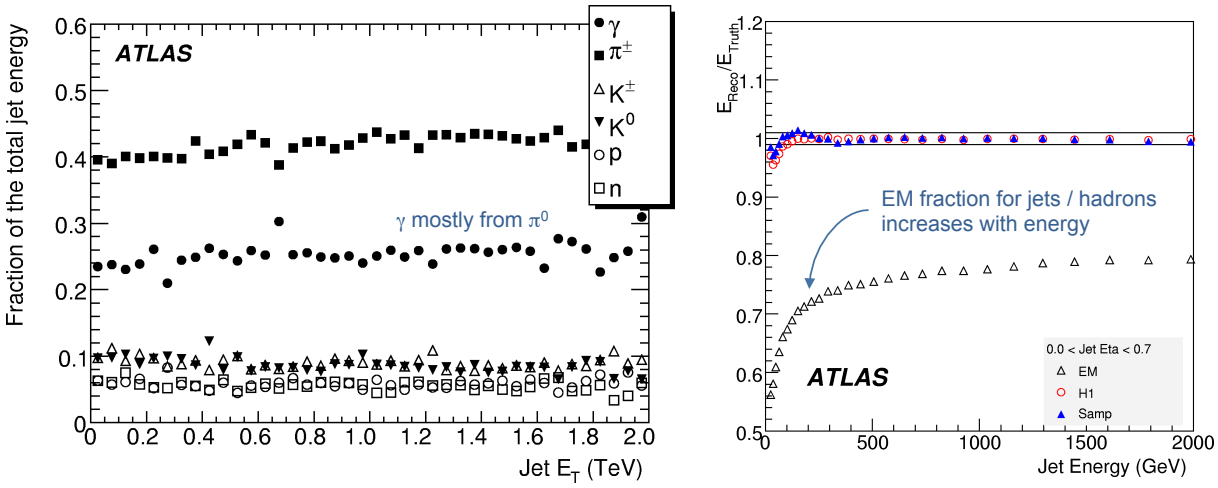


Fig. 62: Left: fractional energy carried by different particle types as a function of the jet energy (ATLAS simulation). Right: jet energy linearity as a function of jet energy (ATLAS simulation). Shown are jets reconstructed at the electromagnetic (EM) scale (open triangles), and using global jet calibration algorithms (open circles and full triangles). The jets have a large cone radius of 0.7.

situ calibration methods are used to match the hadron to the parton levels of the jet using known physics processes.

A large amount of contributions to the jet signal at the various jet levels must be considered in the calibration process. The parton level is governed by the physics process of interest. At the hadron level (particle jet), one must take into account the jet reconstruction algorithm efficiency, added tracks from in-time event pileup from minimum bias scattering interactions, added tracks from the underlying event, and lost soft tracks due to the magnetic field. At the calorimeter jet level one must account for longitudinal energy leakage, detector signal inefficiencies (e.g., dead channels, dead HV boards) background from pileup events, electronic noise, the definition of the calorimeter signal (cluster algorithm, noise suppression, etc.), dead material losses (front material, geometrical cracks in the active material, transition regions, etc.), the detector response characteristics ($e/h \neq 1$), and the jet reconstruction algorithm efficiency. The left panel of Fig. 62 shows the fractional energy that is carried by different particle types in a jet as a function of the jet energy. The largest contributors are charged pions, followed by photons originating mostly from π^0 decays, so that the total pion component amounts to roughly 70% of the jet energy, with no significant jet energy dependence. The right plot shows the jet energy linearity and the electromagnetic fraction versus the jet energy. The electromagnetic fraction for jets or hadrons increases with the jet energy, asymptotically reaching 80% for very hard jets. After calibration, the energy response is accurate above 300 GeV, whereas softer jets are more difficult to calibrate due to the stronger impact of calorimeter noise fluctuations and other effects.

The ultimate goal of the jet reconstruction is to match the calibrated hadronic scale to the initial parton momentum with the use of physics events, i.e., to perform *in-situ* calibration. Several approaches exist.

- Directly verify the hadronic energy scale with **isolated prompt hadrons** from minimum bias events, or hadrons from τ decays, by comparing the reconstructed hadron energy with the momentum of the hadron track measured in the inner tracker. Another possibility is to use track balancing in ϕ (energy and momentum conservation of proton–proton collisions requires the event to be transversely balanced, but not longitudinally) to intercalibrate the hadronic scale with respect to different hadron energies.
- Use **transversely balanced γ-jet or $Z(\rightarrow \ell\ell)$-jet events**. This method assumes that electromag-

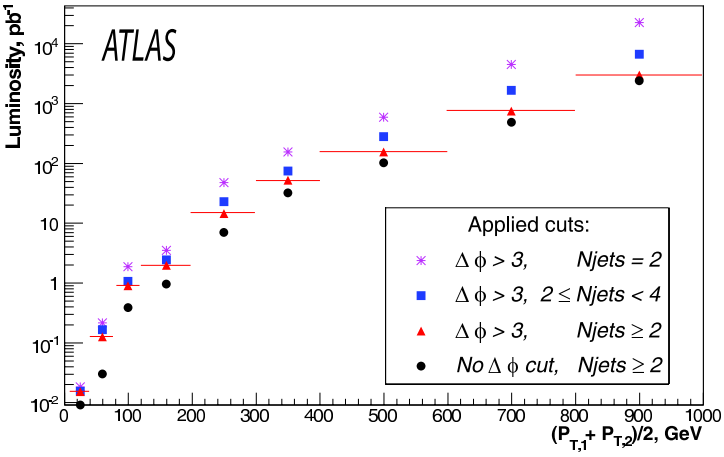


Fig. 64: Integrated luminosity required to reach 0.5% precision on the jet energy scale with the multi-jet calibration method for various p_T ranges in the region $0.7 < \eta < 0.8$, and with different sets of selection cuts (ATLAS simulation): all Pythia di-jet events (circles), requiring $\Delta\phi > 3$ rad between the two leading jets (triangles), requiring in addition fewer than four reconstructed jets in an event (squares), and requiring exactly two reconstructed jets (stars).

netic objects have been properly calibrated beforehand. The jet energy calibration is performed with respect to the average transverse momentum of photon (or Z) and jet. Owing to the large cross section of 180 nb for γ-jet processes[31] this method can be applied with initial data. The statistical yield corresponding to an integrated luminosity of $10\,\text{pb}^{-1}$ would allow a jet calibration of better than 1% statistical precision for $p_T < 200$ GeV. However, the determination of systematic uncertainties is tricky, and requires careful studies. For example, initial- and final-state radiation, underlying events and in-time event pileup, but also out-of-jet particles have the potential to contribute to the γ-jet imbalance, and these effects must be disentangled from miscalibration. Monte Carlo studies by ATLAS have shown that systematic imbalances of non-calibration origin contribute at the 10% level for 20 GeV jets, whereas the effect is below 1% for jets above 100 GeV.

– Use **QCD di-jet and multi-jet events** for $\Delta\eta \times \Delta\phi$ intercalibration. Di-jet events cannot constrain the absolute jet energy scale, but allow one to intercalibrate the calorimeter response. In case of more than two jets in the event, the leading jet dominates the energy resolution of the event, so that one may assume that the error in the vector sum of the 'soft' jets is negligible with respect to the hard jet, and hence 'calibrate' E_T versus η and ϕ (*cf.* sketch in Fig. 63). This method benefits from huge statistics (the di-jet cross section exceeds by a factor 100 to 5000 the γ-jet cross section), but sizable systematic effects arise from soft jets, in particular for the multi-jet approach, requiring detailed studies. Figure 64 shows the integrated luminosity required to reach a precision in the jet energy scale of 0.5% with the multi-jet calibration method for various p_T ranges in the region $0.7 < \eta < 0.8$ and for different sets of selection cuts (see figure caption). Requiring the jets to be back-to-back (i.e., applying a tight $\Delta\phi$ cut) reduces systematic effects from initial- and final-state radiation and the underlying event. The figure has been obtained by ATLAS with the use of simulated events.

– Absolute jet energy scale calibration is possible by means of W **decays into a pair of jets**, for clean W from top decays. However, this calibration applies to soft jets only (jet energies below

[31] The leading parton level processes contributing to the γ-jet cross section are t-channel quark–quark scattering via fermion exchange into $g + \gamma$, and quark–gluon scattering via fermion exchange into $q + \gamma$, gluon–gluon scattering via a box diagram into $g + \gamma$, and the s-channel quark–gluon-to-quark annihilation into $q + \gamma$.

200 GeV). In addition, the W boson does not carry colour charge, which makes it differ from QCD jets.

The LHC will explore energies that have never been reached before. Above 500 GeV, neither measurements nor test beam results are available for jet calibration. Multi-jet balancing should allow a few per cent jet energy scale accuracy in that range with $1\,\mathrm{fb}^{-1}$ integrated luminosity.

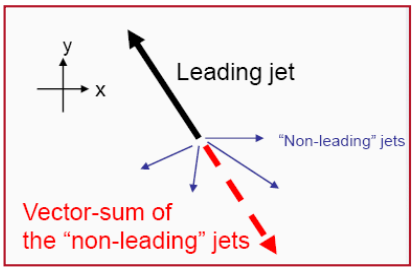


Fig. 63: Illustration of multi-jet energy calibration.

9.5 Missing transverse energy reconstruction

A precise reconstruction of missing transverse energy (MET) in terms of energy scale, linearity, and resolution is essential for the ATLAS and CMS physics programme. Large MET is predicted in many new physics scenarios, notably in supersymmetric extensions of the Standard Model respecting R-parity, where a stable weakly interacting neutral particle is produced that — just as neutrinos — escapes the detector without measurable interaction with the active material. Figure 65 shows a simulated SUSY candidate event in CMS that exhibits significant MET of 360 GeV. MET is also an ingredient of precision Standard Model measurements, such as semileptonic top reconstruction and the W mass, and also of the search for $H \to \tau\tau$ decays, the cross section of which may or may not be enhanced by beyond Standard Model contributions. The MET measurement is particularly sensitive to systematic effects in the detector response and the reconstruction. Understanding MET in early data is therefore one of the primary physics commissioning challenges.

The conceptually simplest way to reconstruct MET is to compute the transverse vector sum of all the electromagnetic and hadronic calorimeter cells and to correct for unaccounted contributions. In the case of ATLAS, one has

$$\not\!\!E_T = \sqrt{\not\!\!E_x^2 + \not\!\!E_y^2}\,, \tag{5}$$

$$\not\!\!E_{x,y} = \not\!\!E_{x,y}^{\mathrm{Calo}} + \not\!\!E_{x,y}^{\mathrm{Cryo}} + \not\!\!E_{x,y}^{\mathrm{Muon}}\,, \tag{6}$$

where the symbol $\not\!\!E$ denotes missing energy. The calorimeter term

$$\not\!\!E_{x,y}^{\mathrm{Calo}} = -\sum_{\text{EM \& Had cells}} E_{x,y}\,, \tag{7}$$

is calibrated at the hadronic energy scale. The electromagnetic scale would underestimate MET by roughly 30% because the largest contributions to it stem from hadrons and jets.

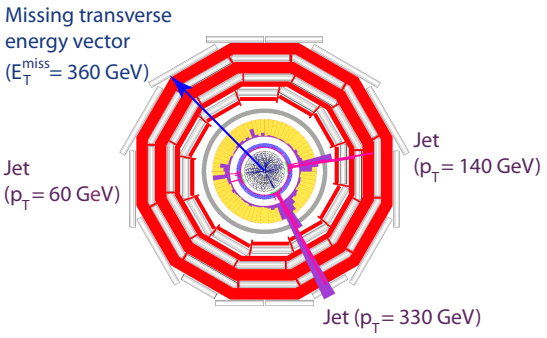


Fig. 65: Display of a simulated SUSY event in CMS. The arrow indicates the missing transverse energy vector.

The 'cryostat' term in Eq. (6) corrects for energy loss (leakage) in the cryostats between the electromagnetic and hadronic calorimeters and becomes important for jets with large transverse momentum (representing a 5% contribution per jet with $p_T > 500\,\mathrm{GeV}$). It is given by

$$\not\!\!E_{x,y}^{\mathrm{Cryo}} = -\sum_{\mathrm{Jets}} w^{\mathrm{Cryo}} \cdot E_{x,y}^{\text{Jet-at-cryo}}\,, \tag{8}$$

where w^{Cryo} is a calibration weight determined empirically from Monte Carlo simulation, and $E_{x,y}^{\text{Jet-at-cryo}}$ is the average of the jet energies deposited in the third layer of the electromagnetic calorimeter and in the first layer of the hadronic calorimeter.

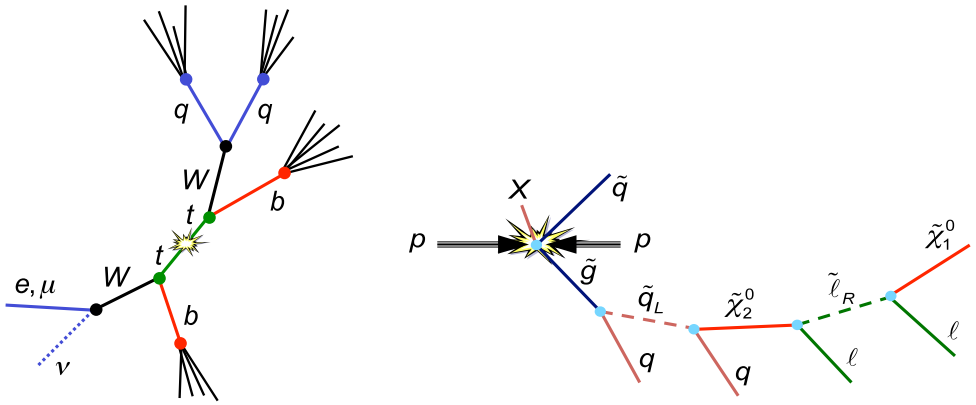


Fig. 67: Schematic graphs of processes generating true MET. The left graph corresponds to a $pp \to t\bar{t}(+X)$ Standard Model event (only the top part of the event is shown), where one top-quark decays fully hadronically and the other semileptonically. The neutrino generates MET. The right graph depicts a typical decay cascade as obtained in R-parity conserving supersymmetry. An initial gluino decays into a left-handed squark and a quark (giving a jet), the squark decays into a heavy neutralino and a quark (giving another jet), the heavy neutralino further decays into a slepton and a lepton, and the slepton finally decays into the lightest neutralino, which escapes detection, and a second lepton of opposite charge with respect to the previous lepton. Note that the initial supersymmetric particles are created in pairs, but only one decay cascade is shown here.

Finally, the muon term sums over measured muon momenta within the muon spectrometer acceptance ($|\eta| < 2.7$)

$$\not{E}_{x,y}^{\text{Muon}} = -\sum_{\text{Muons}} E_{x,y}\,. \tag{9}$$

The MET reconstruction can be refined by associating reconstructed electrons, photons, muons, hadronically decaying τ leptons, b-jets and light jets to calorimeter cells, and replacing for these cells the global calibration by one that takes into account the nature of the identified objects.

It is apparent from the above equations that all detector systems contribute to the MET measurement, which makes it vulnerable to hardware, reconstruction, and calibration problems. One distinguishes between 'true' and 'fake' MET. For example, weakly interacting neutral particles generate true MET (*cf.* Fig. 67). Even without systematic effects, MET is created by the detector response resolution, giving rise to fake MET. Fake MET can also be introduced by detector problems or misreconstruction, such as dead and noisy channels, particles falling out of the detector acceptance (e.g., muons for $|\eta| > 2.7$), unaccounted pile-up contributions to resolution effects, backgrounds from beams or cosmic rays, 'punch-through' of hadron showers into the muon system faking a muon signal, and many more effects.

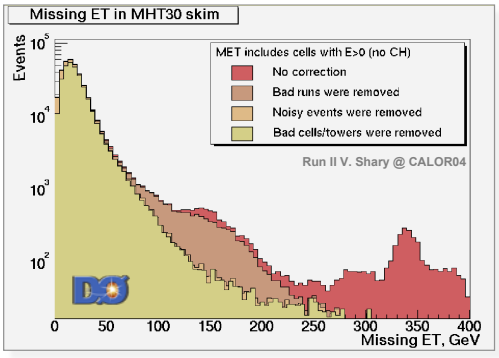


Fig. 66: Missing transverse energy distribution measured by the D0 experiment at the Tevatron $p\bar{p}$ collider. Shown are the various correction stages leading to the removal of fake MET tails that could be misinterpreted as new physics.

The suppression and — if not possible — proper simulation of fake MET is crucial to increase the sensitivity to the true MET. This requires the best possible jet energy resolution and absolute scale, and a thorough classification through data quality bookkeeping and the simulation of varying detector problems. Figure 66 shows an extreme case of MET distortions due to detector noise and bad channel effects,

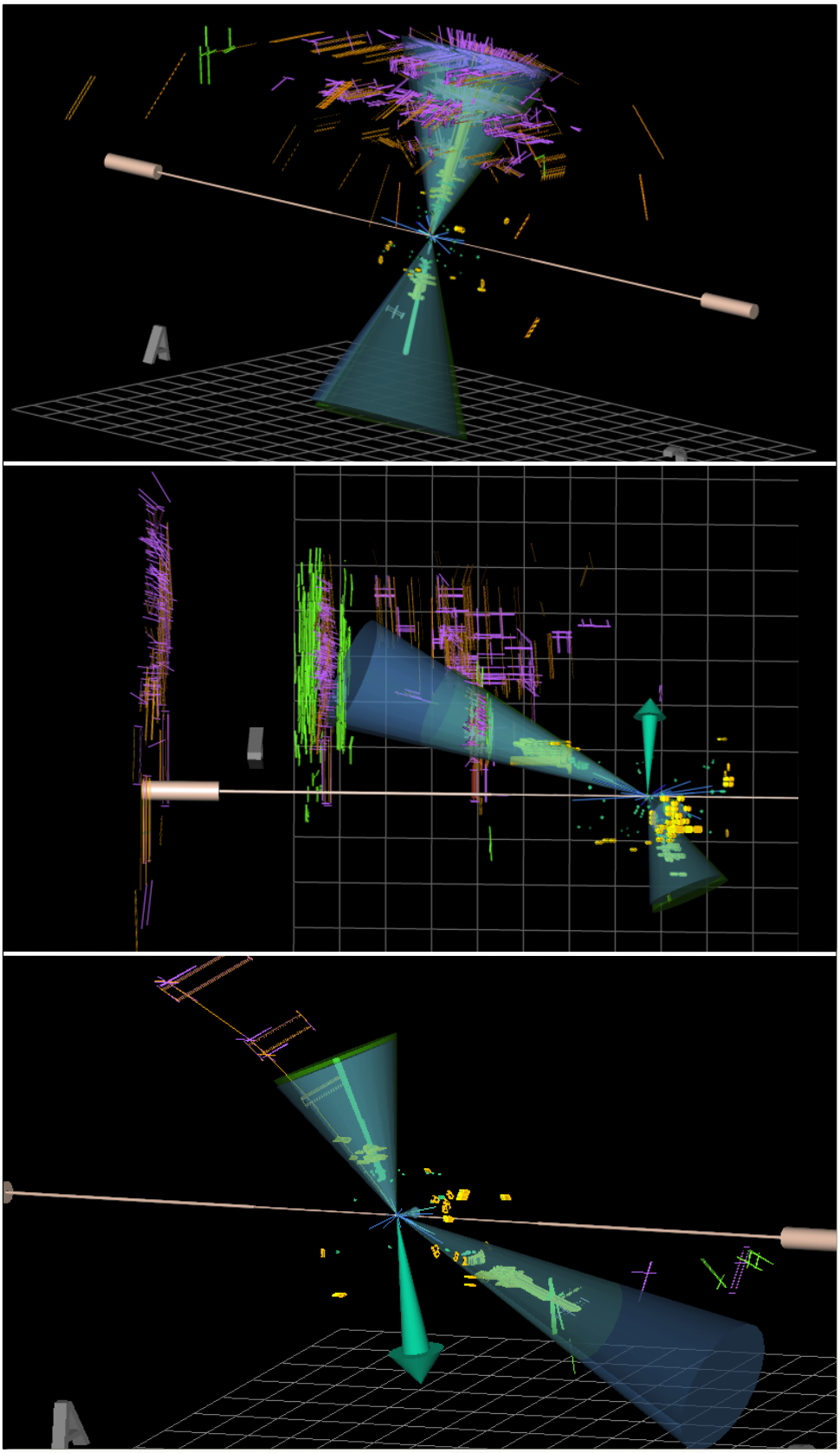


Fig. 68: ATLAS event displays of simulated rare di-jet events creating large amounts of fake MET (represented by the round arrow). The upper two displays show hadron punch-through from the calorimeter into the muon spectrometer. The lower display shows large MET found in an event populating the jet energy resolution tails.

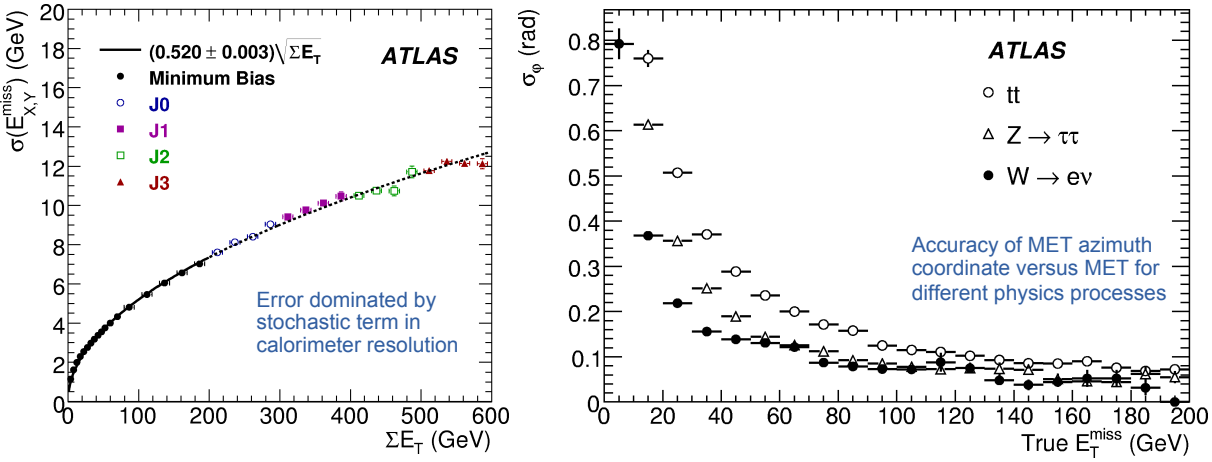


Fig. 69: Left: expected MET resolution for ATLAS versus the transverse energy sum for minimum bias events and various jet samples. Right: accuracy of the MET azimuth angle versus MET for $t\bar{t}$, $Z \to \tau\tau$ and $W \to e\nu$ events, obtained from Monte Carlo simulation in ATLAS.

provided for illustration purposes by the D0 experiment. The effects lead to large tails that — if not properly corrected or simulated — could be misinterpreted as a new physics signal. Figure 67 depicts schematically two processes that generate true MET. The left one is a Standard Model $t\bar{t}$ event where one of the tops decays semileptonically, and the right one is a supersymmetric event with its typical decay cascades ending with two invisible lightest stable supersymmetric particles (only one of two decay cascades is shown). Figure 68 shows event displays from simulated events in ATLAS that were selected for featuring pathologically large fake MET due to hadron punch-through (upper two displays) and jet mismeasurement (lower display). Both types of fake MET usually point towards a jet, which allows such backgrounds to be reduced by eliminating events where the MET vector lies on a jet axis.

Fake MET tails can be studied with early data using minimum bias events. Ample statistics will be available thanks to the large minimum bias cross section,[32] allowing the experiments to select clean data samples. Because the expected true MET is negligible (∼0.06 GeV) the measured MET will be dominated by fake effects from single hadron and jet energy resolution (82%), and acceptance (18%). ATLAS expects an MET average value of 4.3 GeV. The left panel in Fig. 69 shows the expected MET resolution for ATLAS versus the measured transverse energy sum for minimum bias events and various jet samples providing increasing transverse energies. The resolution is dominated by the stochastic term in the jet energy resolution, giving a square-root dependence on the transverse energy sum with the expected coefficient of approximately 50% (see Section 5).

True MET in early data can be measured in leptonic W decays, which have good statistics, but also in $Z \to \tau\tau$ events. In case of one τ decaying semileptonically (hadron(s) plus neutrino) and the other leptonically (electron or muon plus two neutrinos), one can reconstruct the τ mass by assuming that the τ decay products were emitted collinear with the τ flight direction in the lab frame. This is a useful conjecture since the τ exhibits a strong boost. With this one finds

$$m_{\tau\tau}^2 \approx 2 \cdot \left(E_h + E_{\nu(h)}\right) \left(E_\ell + E_{\nu(\ell)}\right) \left(1 - \cos\theta_{h\ell}\right) , \tag{10}$$

where the neutrino energies are approximated by MET. Simulated $Z \to \tau\tau$ decays in ATLAS showed that this method allows the Z mass to be reconstructed with an average resolution of 12 GeV. The right panel of Fig. 69 gives the expected accuracy of the MET azimuth angle versus the true MET for $t\bar{t}$, $Z \to \tau\tau$

[32] At 14 TeV centre-of-mass energy, a minimum bias event rate of 70 kHz is expected to be produced at $10^{30}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$ peak luminosity, and owing to the logarithmic $\sqrt{s}$ dependence similar orders of magnitude are expected at lower centre-of-mass energy. For example, at 900 GeV the minimum bias cross section is reduced by a factor of only 1.8 with respect to 14 TeV.

and $W \to e\nu$ events simulated by ATLAS. With larger true MET the signal-to-calorimeter-noise ratio increases and hence the quality of the MET reconstruction. Moreover, the more hadronic activity in the detector, the worse the MET reconstruction.

10 Early physics with ATLAS and CMS

Early physics measurements will be performed while the detectors are still being commissioned. Some of the commissioning tasks will thus have to take priority to allow systematic uncertainties to be evaluated. An example is the determination of the absolute tracking efficiency, which is an important ingredient of first QCD measurements such as the average number of produced tracks per pseudorapidity region, and which depends on basic detector properties such as the hit efficiency, the alignment of the inner tracking systems, and low-transverse-momentum track finding. Likewise, any physics measurement requires the determination of at least the relative trigger efficiency, and in case of cross section measurements also the absolute trigger efficiency as well as the integrated luminosity. The latter quantity requires either an absolute luminosity detector or, more importantly at the beginning of data taking, an LHC beam scan ('Van der Meer scan'[33] [24]).

The following paragraphs present a very brief and incomplete overview of initial measurements that will be performed at ATLAS and CMS after the collection of approximately $100\,\mathrm{pb}^{-1}$ integrated luminosity. Most of the prospective studies shown here are taken from ATLAS [3]. The CMS studies, documented in Ref. [4], are very similar. All results shown are based on Monte Carlo simulation at 14 TeV centre-of-mass energy. This is, however, not the energy at which the LHC will start. Because of problems with the magnet quench protection, the startup centre-of-mass energy in 2010 will be 7 TeV, after a pilot run at LHC injection energy of 0.9 TeV in 2009. The decision whether or not to raise the energy to 10 TeV in the course of the year 2010 will depend on the running experience. The design energy of 14 TeV can only be reached after a shutdown of approximately one year, which may be scheduled in 2011 or 2012, when vulnerable parts of the quench protection system are exchanged.

10.1 Minimum bias studies

Minimum bias events will dominate the first triggered data samples of all LHC experiments. The total minimum bias cross section receives contributions from inelastic non-diffractive and diffractive collisions,[34] where whether or not single diffractive events are included is subject to the experiment's definition. Experimentally, it is not possible to distinguish these classes of events on an event-by-event basis. Minimum bias triggers have usually large (medium, small) efficiencies for non-diffractive (double diffractive, single diffractive) events. If coincident hits in both forward regions of the detector are required, the efficiency of single diffractive events becomes small. In-time coincidence is a useful requirement to eliminate beam related backgrounds (beam gas and beam halo events, *cf.* Section 7.2). These backgrounds are, however, also eliminated when requiring the reconstructed tracks in the event to form a primary vertex. The minimum bias analysis will most likely be the first paper published by ALICE, ATLAS and CMS. Apart from the physics measurement, it will represent a first proof that the detectors (mainly the inner tracking systems) work and the data including the trigger and tracking efficiencies are understood.

Multiparticle production is successfully described by phenomenological models with pomeron

[33] The beam scan is used to measure the beam sizes and positions in a collider, which, together with the known currents, can be used to compute the absolute luminosity. The beams are scanned across each other at the collision point and, using beam position measurements, the amount of motion is correlated with detectors monitoring the relative luminosity of the collisions at each scan point. This method has been successfully applied at the heavy-ion collider RHIC [25].

[34] Diffraction denotes the excitation of the proton(s) participating in the inelastic scattering. One distinguishes single, double and central diffraction. While single and double diffractive events produce activity in only one and both forward regions of the detector, respectively, central diffractive events, (which are described by double pomeron exchange, and have small cross sections), give activity at small pseudorapidities.

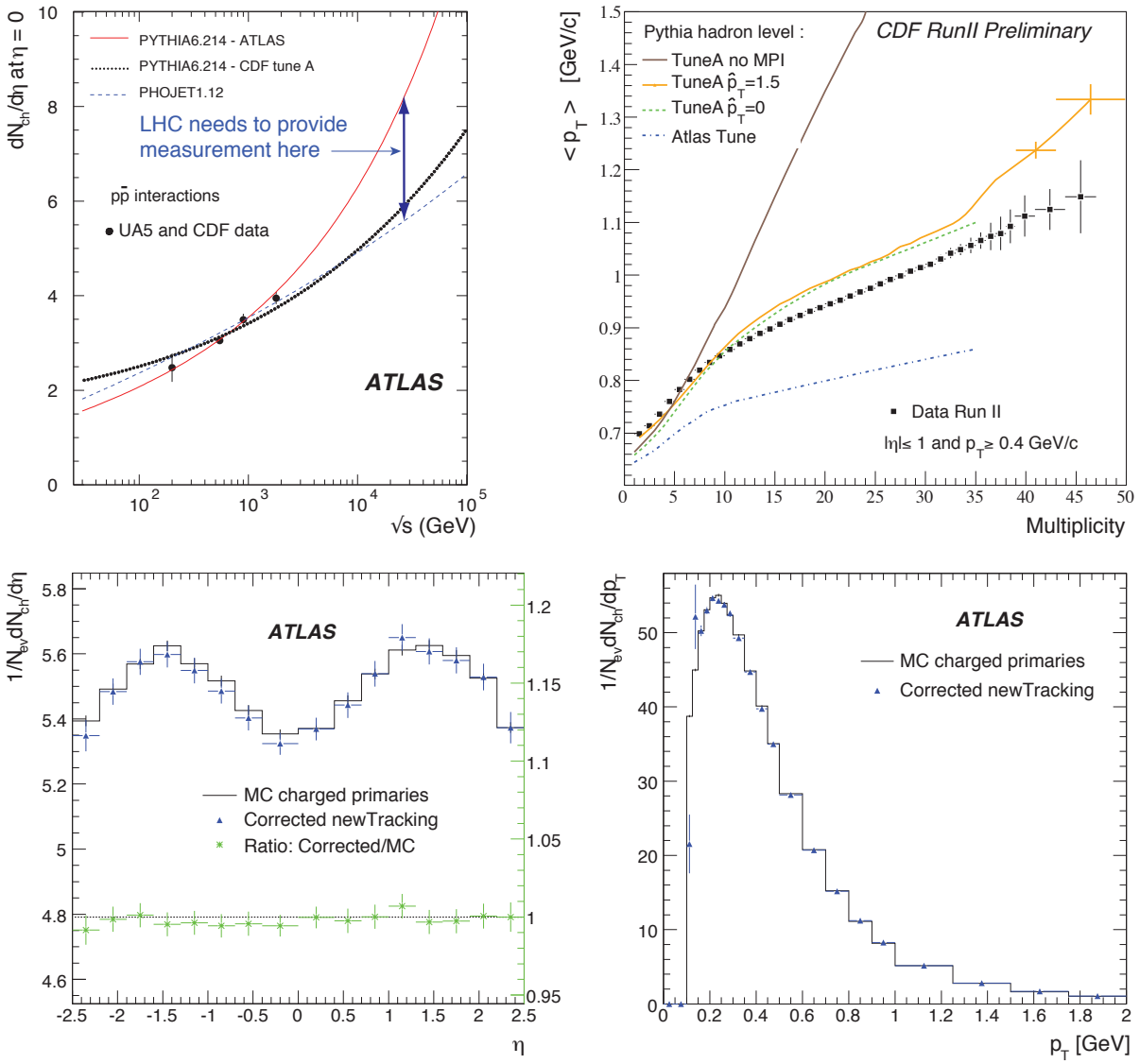


Fig. 70: Top left: central charged particle density for non-single diffractive inelastic events in $p\overline{p}$ collisions as a function of energy, extrapolated to large centre-of-mass energies. Shown are available measurements and Monte Carlo generator predictions. Top right: correlation between average track transverse momentum and the charged particle multiplicity for $\eta < 1$ as measured by CDF [26], and compared with various Pythia generator tunings. 'No MPI' means that multiple parton interactions have been switched off in the generator. Bottom plots: particle density in non-diffractive minimum bias events versus the pseudorapidity (left) and p_T (right) in ATLAS with special low-momentum track reconstruction enabled. Systematic errors on track reconstruction are not included in the right plot.

exchange, which dominates at high energies. These models relate the energy dependence of the total cross section to that of the multiplicity production using a small number of parameters, and are the basis for several Monte Carlo event generators describing soft hadron collisions. Minimum bias multiplicity measurements between 200 GeV and 2 TeV centre-of-mass energies at the CERN ISR, CERN Sp$\overline{\text{p}}$S, Fermilab's Tevatron, and BNL's RHIC colliders have been used to tune these generators for predictions of multiplicities at LHC energies.

The top left panel in Fig. 70 shows a comparison of model predictions for the central charged particle density in non-single-diffractive $p\overline{p}$ events for a wide range of centre-of-mass energies compared

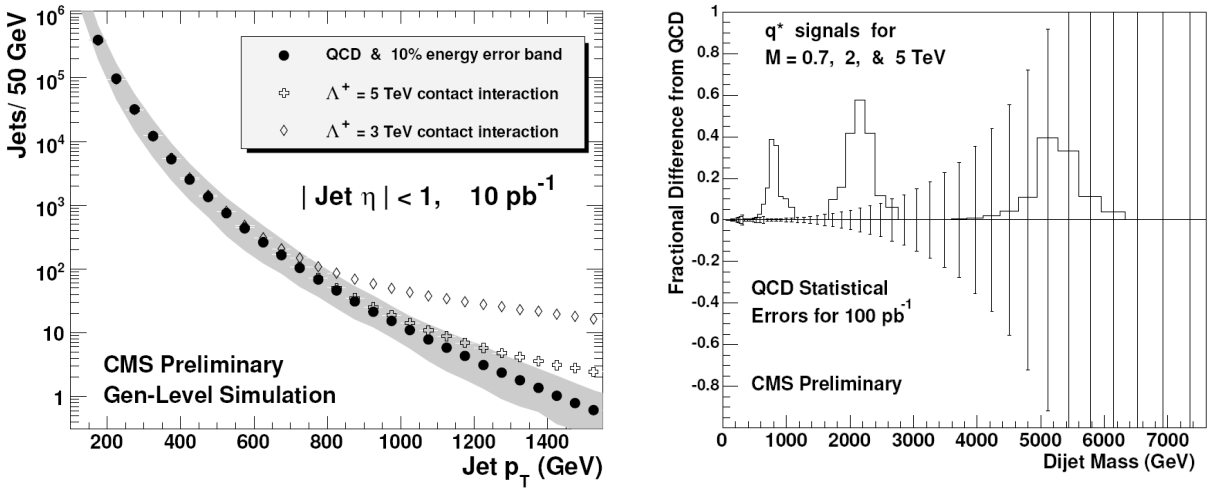


Fig. 71: Left: distribution of the transverse momentum of the hardest central jet in simulated di-jet events for $10\,\mathrm{pb}^{-1}$ integrated luminosity (CMS study). The shaded band indicates the systematic uncertainties. The cross and diamond curves indicate the distortions in the high-p_T spectrum expected from contact interactions at the scales $\Lambda^+ = 3\,\mathrm{TeV}$ and 5 TeV, respectively. Right: fractional difference from the QCD expectation of the di-jet invariant mass (CMS). Also shown are the contributions to the difference from heavy excited quarks decaying into jet pairs.

with measurements that have been corrected for detector acceptance. Large extrapolation uncertainties exist that must be overcome by LHC measurements. Improved generator tunings at LHC energies will directly feed into Monte Carlo predictions of many primary physics channels. A good minimum bias multiplicity description is also important because event pileup from minimum bias interactions is background to hard scattering processes at high luminosity. The top right panel in Fig. 70 is taken from CDF [26]. It shows the measured dependence of the average track transverse momentum on the charged particle multiplicity per event for $|\eta| < 1$, compared with various Pythia generator tunings. Without multiple parton interactions the average predicted p_T multiplicities above 6 is grossly overestimated. The bottom plots in Fig. 70 show the particle density versus pseudorapidity (left) and transverse momentum (right) in ATLAS for simulated minimum bias events. Special low-p_T tracking reconstruction has been enabled for these plots, which allows one to lower the track measurement down to $p_T = 150\,\mathrm{MeV}$ (standard cut is 500 MeV), at the price of larger systematic uncertainties (not included in the error bars). The statistics shown corresponds to 1 minute of data taking with $10^{31}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$ at 14 TeV.

10.2 Di-jet studies

Jet production has a roughly 1000 times lower cross section than non-diffractive minimum bias scattering, but is still an abundant process for early physics measurements and performance studies. Apart from its importance for QCD studies and Monte Carlo generator tuning at yet unexplored centre-of-mass energies, jet production can be used to probe the Standard Model. Inclusive di-jet production ($pp \to 2\,\mathrm{jets}+X$) is the dominant LHC hard scattering process. It is straightforward to observe and has a rich potential of new physics signatures. Restricting the leading jet (the jet with the largest p_T) to the central detector region $|\eta| < 1$ reduces the background from QCD t-channel processes, thus enhancing the sensitivity to new physics contributions to the s-channel at small pseudorapidities. The main variables used for new physics searches are the transverse momentum of the leading jet and the di-jet invariant mass. Prospective studies from CMS show that the highest di-jet masses reached with integrated luminosities of $100\,\mathrm{pb}^{-1}$, $1\,\mathrm{fb}^{-1}$, and $10\,\mathrm{fb}^{-1}$ are respectively 5, 6 and 7 TeV. The current limits from measurements by the Tevatron experiments will be almost immediately extended by ATLAS and CMS.

The left panel in Fig. 71 shows the distribution of the leading central-rapidity jet p_T in simulated

di-jet events as expected by CMS for an integrated luminosity of $10\,\mathrm{pb}^{-1}$. The shaded band indicates the estimated systematic uncertainties. Also shown are the distortions in the spectrum expected from contact interactions[35] at the characteristic scales $\Lambda^+ = 3\,\mathrm{TeV}$ and 5 TeV, respectively. A quantitative sensitivity study shows that contact interactions up to $\Lambda = 3\,\mathrm{TeV}$ can be discovered with the first $10\,\mathrm{pb}^{-1}$. However, the analysis requires excellent understanding of the jet resolution in the tails and the jet energy scale. Systematic errors dominate over the statistical ones and over uncertainties from the parton density functions.

The right plot in Fig. 71 shows a Monte Carlo study by CMS of a search for strongly produced heavy excited quarks decaying into a quark pair. The most sensitive observable here is the di-jet invariant mass. Shown in the plot is the fractional difference between measurement (here: simulation) and Standard Model expectation for $100\,\mathrm{pb}^{-1}$ integrated luminosity. Shown by the resonances are the contributions from excited quarks to that difference, which can be clearly separated below 3 TeV. Other variables can also be looked at. For example, the ratio of di-jet abundances between different regions of pseudorapidity versus the di-jet invariant mass benefits from reduced systematic uncertainties compared with absolute cross section measurements. Also angular distributions exhibit sensitivity to new physics.

10.3 Quarkonia production

Quarkonia ($q\bar{q}$ resonances such as J/ψ, ψ', Υ, Υ', etc.) are abundantly produced at the LHC (see the Feynman graphs in Fig. 72) and excellent sources for early physics commissioning, but also for early physics measurements, e.g., prompt versus non-prompt production distinguished via different lifetimes, ratios of cross sections, polarisation, etc. Examples of Feynman diagrams for the singlet and octet production of a J/ψ resonance are drawn in Fig. 72. The upper diagram describes the leading colour-singlet process, which has a small cross section. The middle diagram, which dominates at low p_T, can be produced through both singlet and octet $c\bar{c}$ states with various quantum numbers. At high p_T, the gluon fragmentation subprocess shown in the lower plot becomes increasingly important.

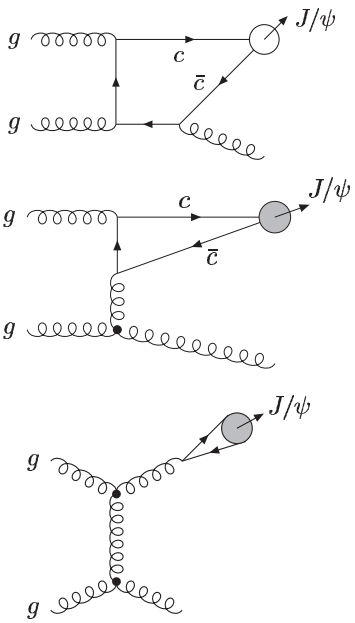


Fig. 72: Examples of Feynman diagrams for the singlet and octet production of a J/ψ resonance (see text).

Quarkonia in ATLAS and CMS are mainly studied through their decays into muon and also electron pairs. Since they are narrow resonances they can be used as commissioning tools for the alignment and calibration of the trigger, tracking, and muon systems. Efficiency studies can employ the 'tag-and-probe method' (see Section 9.3). Owing to the low mass of the resonances, trigger considerations are crucial to estimate the available cross section for analysis. Using a di-muon trigger with 4 GeV thresholds for each muon, the overall rate of events from all quarkonium states is likely to remain below the rate of 1 Hz at a luminosity of $10^{31}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. (The trigger rates may be dominated by background processes.)

The left-hand plot in Fig. 73 shows the cumulative differential cross section of the invariant di-muon invariant mass for J/ψ and $\Upsilon(1S)$ signal events and various combinatorial backgrounds from an ATLAS Monte Carlo study. The plot includes trigger requirements of at least one muon with 6 GeV and

[35] New physics models with fermion substructure ('compositness') at high scale lead to excitations of these fermions which modify scattering cross sections. The interaction can be parametrised by an effective four-fermion contact term

$$\mathscr{L}_{\mathrm{eff}} = \frac{4\pi^2}{\Lambda^2} \sum_{i,k=L,R} \alpha^{ik} \left(\bar{q}_i \gamma^\mu q_i'\right) \left(\bar{f}_k \gamma^\mu f_k'\right) , \tag{11}$$

where Λ is the mass scale of the new interaction. Experimental limits exclude excited fermions up to a few TeV.

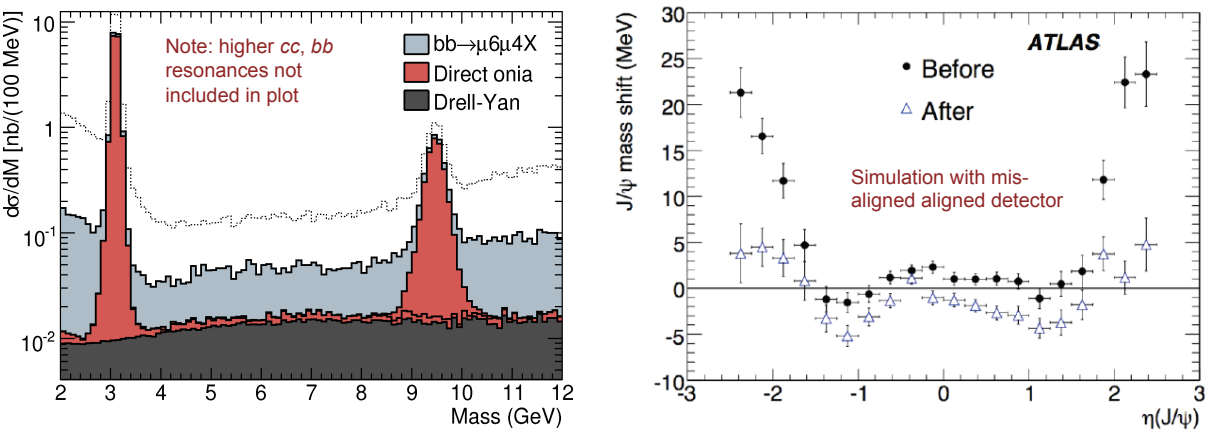


Fig. 73: Left: cumulative differential cross section versus the invariant mass of muon pairs from various quarkonia signal and combinatorial background sources (ATLAS study). A primary vertex and pseudo-proper time requirement of 0.2 ps has been applied. The dotted line shows the cumulative distribution without these cuts. The quarkonia simulation used for the plot does not include higher radial excitations. Right: invariant mass of di-muons from J/ψ decays versus the J/ψ pseudorapidity for simulated ATLAS data where a severe misalignment of the inner tracking system has been introduced. Shown are results before and after alignment.

another one with 4 GeV transverse momentum, and that these muons must originate from a common primary vertex. In addition a lifetime requirement has been applied. (The dotted line shows the cumulative distributions without these latter two requirements). Backgrounds from Drell–Yan processes and leptonic heavy-quark decays are of similar size.

The right panel of Fig. 73 shows a simulated commissioning result from ATLAS. Events of the type $pp \to J/\psi(\to \mu\mu) + X$ with (somewhat unrealistically) severe misalignment in the inner tracker have been simulated and run through the alignment procedure based on hits-on-track residual minimisation. As discussed in Section 6.2, this method suffers from so-called weak modes, which denote misalignments that leave the global χ^2 function, used to minimise the hit residuals, invariant. As seen in the plot, the reconstructed invariant di-muon mass versus the pseudorapidity of the di-muon system exhibits a strong non-uniformity before the alignment, but remaining effects caused by weak modes after the alignment.

10.4 W and Z boson production

Inclusive production of W and Z bosons ($pp \to W(Z) + X$) has large cross sections so that interesting data-driven cross section measurements can be performed with early data (10–50 pb^{-1}). The weak bosons are also important ingredients for commissioning studies: Z bosons are most important for various *in-situ* calibrations (*cf.* Section 9), and Z+jets and W+jets are sensitive probes of higher order QCD calculations. Inclusive weak boson production is also precisely predicted by theory so that a cross section measurement in particular of the more abundant W production can be used to infer the absolute integrated luminosity recorded.

Figure 74 shows the distribution of the W transverse mass for $W \to e\nu$ (left) and $W \to \mu\nu$ (right) decays together with their dominant backgrounds for simulated data corresponding to 50 pb^{-1} integrated luminosity (ATLAS study). The transverse mass is defined by

$$m_T = \sqrt{E_T^\ell \not{E}_T \, (1 - \cos\Delta\phi)}, \tag{12}$$

where $\Delta\phi$ is the angle between the transverse lepton and missing energy vectors, and E_T^ℓ is the transverse energy of the lepton. The transverse W mass is also used as an ingredient for the precision measurement

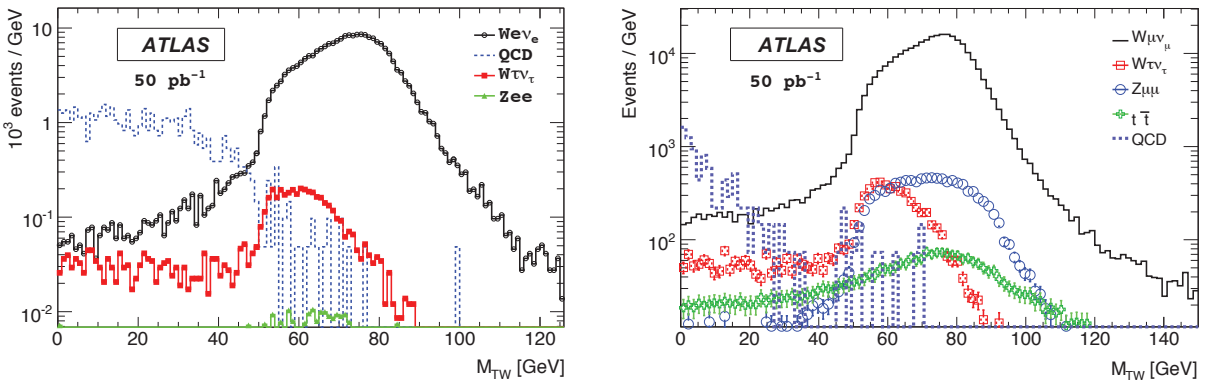


Fig. 74: W-boson transverse mass distribution for $W \to e\nu$ (left) and $W \to \mu\nu$ (right) and backgrounds after full selection except for the M_T cut, for simulated data corresponding to an integrated luminosity of $50\,\mathrm{pb}^{-1}$.

of the W mass, which, however, requires much larger data samples for a competitive measurement, because of the required mass calibration with respect to the accurately known Z boson, which has a ten times smaller leptonic cross section.

Figure 75 shows the Z-boson transverse momentum distribution as measured by D0 at the Tevatron for a centre-of-mass energy of 1.96 TeV. It is compared with Monte Carlo generator models including next-to-leading order QCD calculations. Good control of the transverse momentum of weak bosons is important for many physics studies. Specifically in multivariate Higgs searches, the Higgs transverse momentum can be used as a discriminating variable since Higgs production is expected to have a harder spectrum than QCD backgrounds.

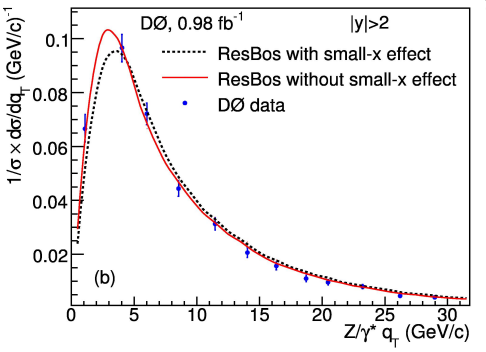


Fig. 75: Z-boson transverse momentum distribution, measured by the D0 experiment at the Tevatron, compared with Monte Carlo generator models.

10.5 Top-quark production

The roughly 100 times larger $t\bar{t}$ production cross section of $\sim$830 pb at the LHC (at 14 TeV centre-of-mass energy) compared with $\sim$7.5 pb at the Tevatron, makes it possible to observe top quarks in early data. Also the electroweak single-top production cross section of $\sim$300 pb is similarly enlarged. Apart from having important physics potential, top quarks represent excellent objects for data-driven commissioning and calibration analyses, notably b-tagging and jet energy scale fits. The leading processes contributing to $t\bar{t}$ production are gluon–gluon scattering (s and t-channels) and quark–antiquark annihilation (s-channel). Single-top production is dominated by W–gluon fusion (t-channel), W exchange between b quarks (t-channel), associated production of top and W, and quark–antiquark annihilation (s-channel, smaller cross section).

Data corresponding to an integrated luminosity of $100\,\mathrm{pb}^{-1}$ should allow the experiments to measure the $t\bar{t}$ production cross section, with events where both W bosons decay leptonically, to an accuracy of 3% statistical and 5% systematic error (dominated by the uncertainty in the integrated luminosity value). The measurement provides an important probe of the validity of the Standard Model at unexplored centre-of-mass energy. Figure 76 shows on the right panel the reconstructed hadronic top mass from a combination of three jets as found in a simulated signal and background sample corresponding to $100\,\mathrm{pb}^{-1}$ integrated luminosity after full event selection. The left panel shows the corresponding di-jet mass formed by light-flavour jets, representing the W signal and combinatorial background. This plot

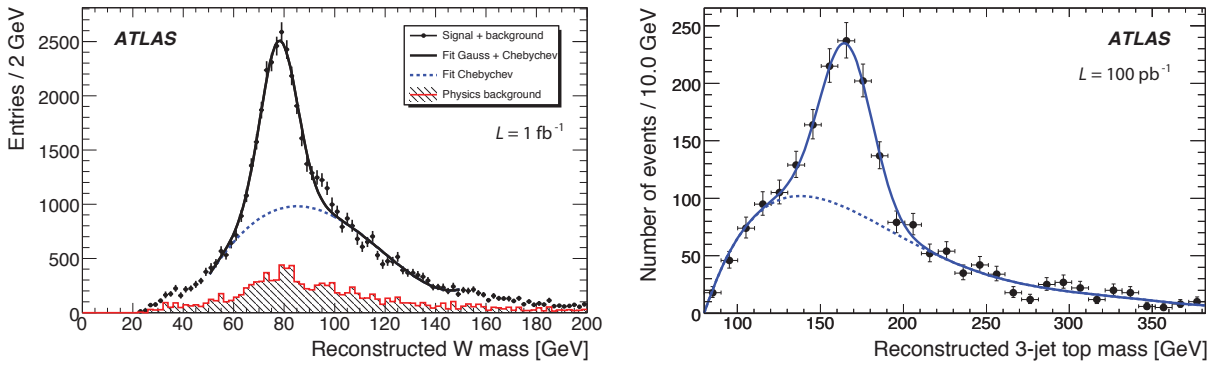


Fig. 76: Left: distribution of the invariant di-jet mass of light-flavoured jets as an estimate of the hadronically decaying W-boson mass in a simulated data sample corresponding to $1\,\mathrm{fb}^{-1}$ integrated luminosity for signal $t\bar{t}$ and background processes (ATLAS study). Right: reconstructed hadronic top mass from the combination of three jets in simulated data corresponding to $100\,\mathrm{pb}^{-1}$ integrated luminosity signal $t\bar{t}$ and background events after full selection.

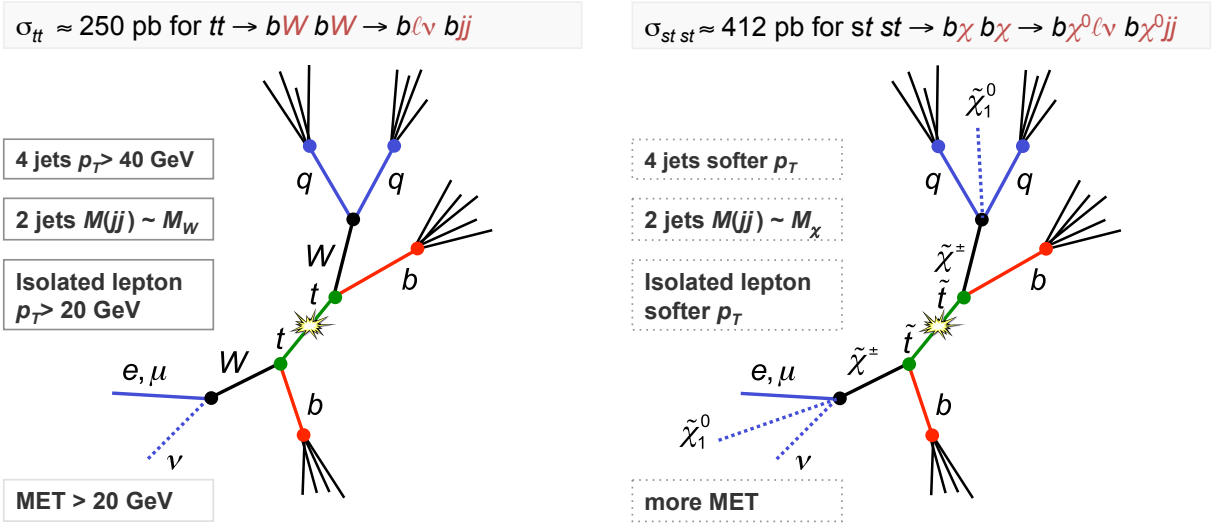


Fig. 77: Left: schematic graph of a $pp \to t\bar{t}(+X)$ process where one top decays fully hadronically and the other semileptonically. The neutrino generates missing transverse energy. Right: schematic graph of a supersymmetric stop–antistop production. The $W^{\pm}$ propagators in top–antitop production are replaced by charginos that decay into three bodies of which one (the neutralino) is a stable weak interacting neutral particle.

can be used to determine and adjust the jet energy scale. A kinematic fit to the true W mass can be used to improve the accuracy of the three-jet top-mass reconstruction.

Single-top production is of particular interest due to its sensitivity to charged new physics fields, such as a charged Higgs replacing the W in the weak propagator as occurs in two-Higgs-doublet models. Single-top production has been observed by the Tevatron in 2009 with the use of advanced multivariate analysis techniques [27]. The measured cross section of $(2.3^{+0.6}_{-0.5})$ pb (CDF), (3.94 ± 0.88) pb (D0), is in agreement with the Standard Model expectation.

Figure 77 shows on the left diagram a schematic drawing of a Standard Model top–antitop event, where one top decays fully hadronically and the other semileptonically. The right diagram shows the production and decay of a light supersymmetric (R-parity conserving) stop–antistop pair, which follows a similar decay cascade with, however, an additional weak interacting neutral particle in the final state

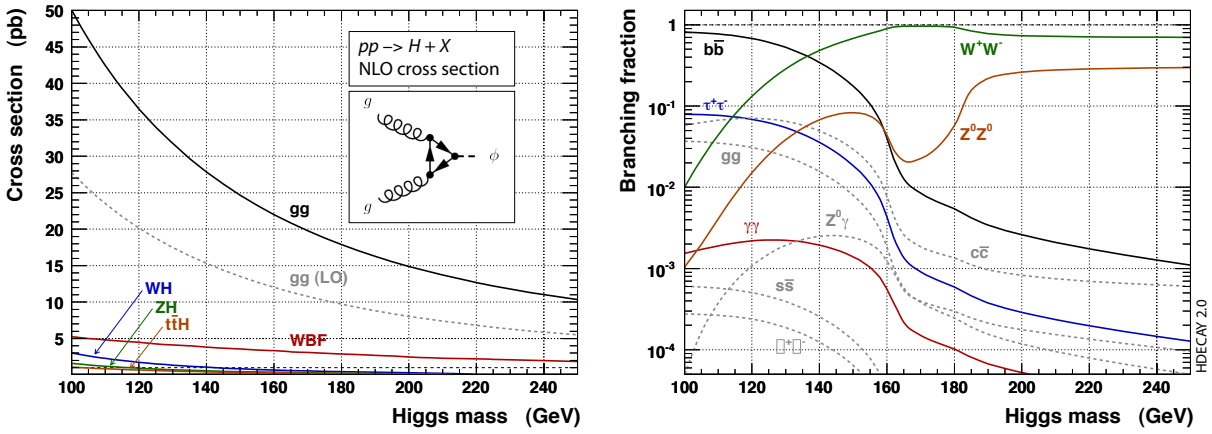


Fig. 79: Expected inclusive Standard Model Higgs boson production cross section for the various production modes (left) and Higgs branching fractions (right) versus the Higgs mass at 14 TeV centre-of-mass proton–proton collisions.

that escapes the detector. As a consequence, the four particle jets and the isolated lepton are softer than in the $t\bar{t}$ case, the two light jets originating from the heavy neutralino decay form the invariant mass of a neutralino, instead of that of a W, and significantly more missing transverse energy is produced in the supersymmetric event. The experimental separation of the $t\bar{t}$ and $\tilde{t}\bar{\tilde{t}}$ processes is difficult and requires more statistics than available in early data taking. The analysis requires b-flavour tagging to be commissioned and proceeds by plotting the minimum three-particles invariant mass that can be formed of a b-jet and the two light-flavoured jets. Subtracting from it the expected $t\bar{t}$ Standard Model contribution a $\tilde{t}\bar{\tilde{t}}$ contamination would show up by a peak below the top (and below the stop mass, due to the escaped neutralino). A study performed by ATLAS shows that with $1.8\,\text{fb}^{-1}$ and a stop mass of 137 GeV, for which the $\tilde{t}\bar{\tilde{t}}b\chi b\chi \to (b\chi^0\ell\nu)(b\chi^0 qq)$ cross section amounts to 412 pb (depending also on other model parameters), exceeding by a factor 1.6 the corresponding $t\bar{t} \to bWbW \to (b\ell\nu)(bqq)$ cross section, a clear signal can be derived.

10.6 Standard Model Higgs boson search

The observation of a Standard Model Higgs boson is inverse femtobarn rather than picobarn physics, and hence not of primary importance for early physics. However, new physics may enhance Higgs-like signals and the experiments must be prepared for surprises. It is also important to begin early with the understanding and improvement of electron muon, tau and photon selection efficiencies and purities, and the study of b-jet and forward-jets tagging, and a thorough categorisation of the relevant Higgs backgrounds to tune the multivariate analyses that will be used to extract a signal.

Figure 79 shows the dependence of the inclusive Standard Model Higgs boson production cross section and branching fractions on the Higgs mass. The dominant production mode is the fusion of two gluons into a scalar Higgs via triangular top loop. Next-to-leading order (NLO) corrections give a sizable K-factor (the factor with which the leading order result needs to be multiplied to include higher orders) in this process. The second most important process is weak boson fusion that is accompanied by two forward jets. Since it is a weak process, next-to-leading order corrections are less important. Following are the associated Higgs production with a W or a Z boson, or with a $t\bar{t}$ pair. The strong rise in the branching fractions to the heavy weak boson pairs is

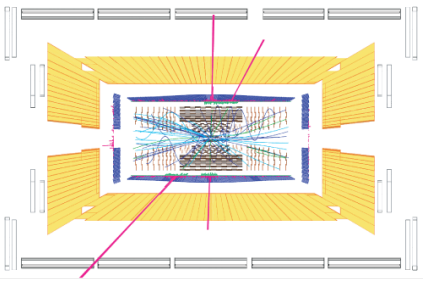


Fig. 78: A simulated $H \to ZZ^\star \to 4e$ event with $m_H = 150\,\text{GeV}$ in CMS.

due to the kinematic opening of these channels, which are favoured because the Higgs couples to the masses of the particles (if the Higgs boson were heavy enough to be able to decay into a top–antitop pair (not shown in the plot), it would reach a branching fraction of up to 20% at around $m_H \sim 500\,\text{GeV}$).

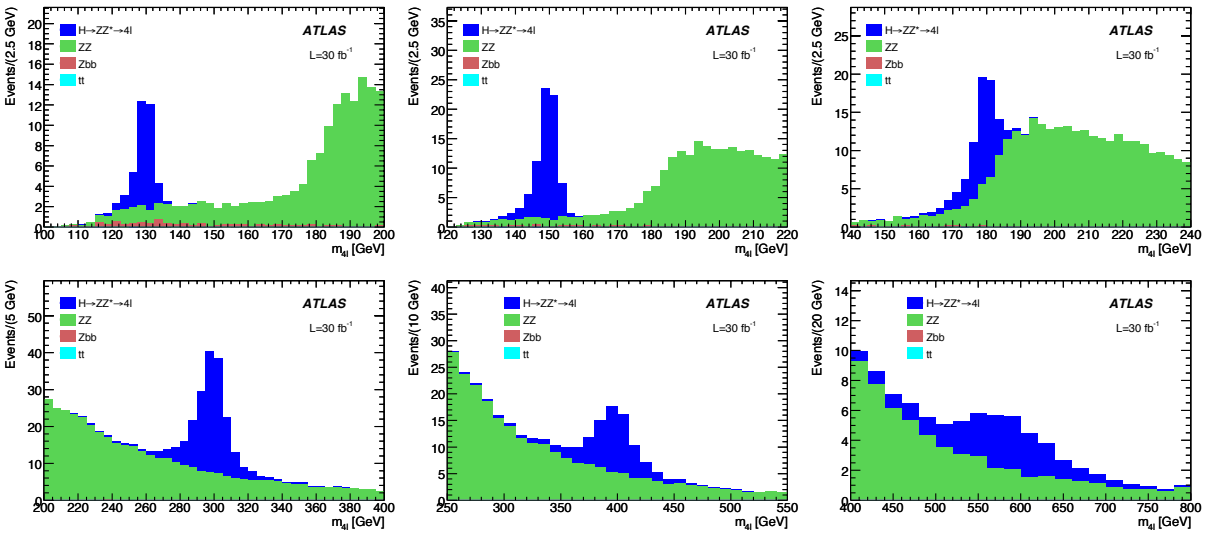


Fig. 80: Reconstruction of the four-lepton invariant mass for simulated $H \to ZZ^{(\star)} \to 4\ell$ ($\ell = e, \mu$) signal and background events corresponding to an integrated luminosity of $30\,\text{fb}^{-1}$ (ATLAS study). From upper left to lower right are shown analyses for the true Higgs masses 130, 150, 180, 300, 400, and 600 GeV, respectively.

Figure 80 illustrates the results of a simulated search for $H \to ZZ^{(\star)} \to 4\ell$ in ATLAS with $30\,\text{fb}^{-1}$ for different true Higgs masses. Because the Higgs partial width into two vector bosons increases with m_H^3, but only linearly for a Higgs decaying into two fermions,[36] the total Higgs width grows fast beyond the $H \to WW$ and $H \to ZZ$ openings. For example, while the Standard Model Higgs width is only 3.6 MeV at $m_H = 120\,\text{GeV}$ and 76 MeV at 160 GeV, it grows to 1.4 GeV at 200 GeV and 8.5 GeV at 300 GeV. In the region favoured by the electroweak fit (see below) the Higgs intrinsic width is much smaller than the experimental resolution and hence negligible.

Electroweak precision observables, measured by experiments at the LEP (CERN), SLC (SLAC) and Tevatron (FNAL) accelerators, can be used in a global Standard Model fit to derive a constraint on the Higgs mass. The resulting $\Delta\chi^2$ curves versus the Higgs boson mass, without and with results from direct Higgs boson searches at LEP and the Tevatron included in the fit, are given in Fig. 81. The result including all the available information yields the allowed range $114 < m_H < 157\,\text{GeV}$ at 95% confidence level. Although this represents an important indication, experimentalists cannot afford to disregard the high-mass region. The analyses must cover all Higgs masses that are not yet excluded by direct searches.

[36] The leading order width of the Higgs boson decay into a fermion–antifermion pair is given by

$$\Gamma^{(\text{LO})}(H \to f\bar{f}) = \frac{G_F N_C}{4\sqrt{2}\pi} m_H m_f^2 \beta_f^3\,, \tag{13}$$

where $G_F = 1.16637 \cdot 10^{-5}\,\text{GeV}^{-2}$ is the Fermi constant, $\beta_f = \sqrt{1 - 4m_f^2/m_H^2}$ is the fermion velocity in the Higgs rest system, and $N_C = 3(1)$ is the number of colours for quarks (leptons). Large next-to-leading order corrections can occur in the case of quarks. The leading order width of the decay into two on-shell weakly interacting vector bosons reads

$$\Gamma(H \to VV) = \frac{G_F m_H^3}{16\sqrt{2}\pi} \cdot \delta_V \cdot A(x)\,, \tag{14}$$

where $\delta_V = 2(1)$ for $V = W(Z)$, and $A(x) = \sqrt{1-4x} \cdot (1 - 4x + 12x^2)$ with $x = m_V^2/m_H^2$. For masses much larger than $2m_Z$ the width $\Gamma(H \to WW)$ is twice as large as $\Gamma(H \to ZZ)$. Very roughly one finds $\Gamma(H \to WW + ZZ) \approx 0.5\,\text{TeV} \cdot (m_H/1\,\text{TeV})^3$, so that for a Higgs mass of 1 TeV the Higgs width becomes of the same order of magnitude.

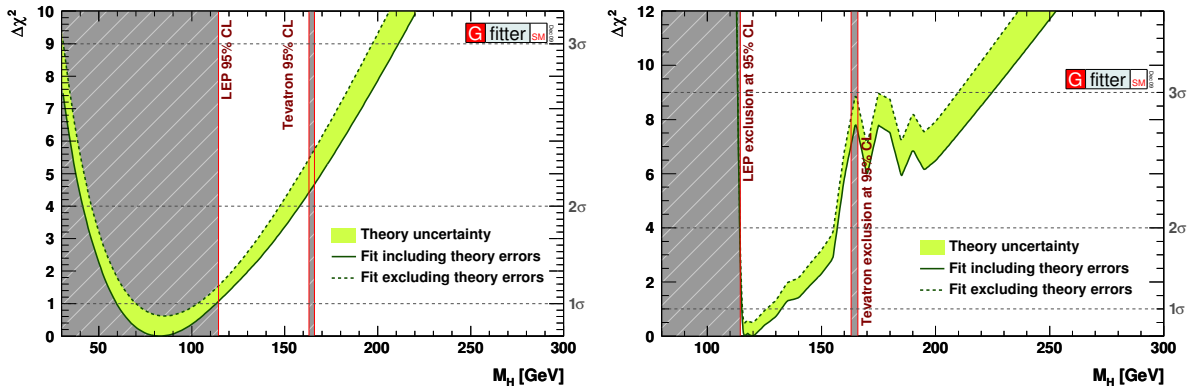


Fig. 81: Curves of $\Delta\chi^2$ obtained from the global fit to electroweak precision data. The right plot includes in addition the results from the direct Higgs boson searches at LEP and Tevatron. The plots are taken from Ref. [6].

From Fig. 81 it becomes clear that very different experimental search strategies need to be pursued depending on the Higgs mass hypothesis. The golden discovery modes are $H \to \gamma\gamma$ for masses below $\sim$150 GeV (grand maximum), which is a very rare channel (branching fraction of about 0.2%) with a clean signature, $H \to WW^{(\star)} \to \ell\nu\ell\nu$ for high masses, which is an abundant but not a clean mode, and $H \to ZZ^{(\star)} \to 2\ell 2\ell'$ which has a sizable branching fraction above $m_H \simeq 130$ GeV, and which is clean at relatively low mass. We have no space here to discuss all these measurements. Early searches will concentrate on the high-cross-section modes leading to a successive exclusion (or discovery) of smaller and smaller Higgs masses. ATLAS and CMS have performed studies to evaluate the discovery reach of the various Higgs search analyses as a function of the Higgs mass. Figure 82 shows an ATLAS study for the Higgs boson discovery (left panel) and exclusion potential (right panel) for given integrated luminosity versus the Higgs mass. At 1 fb^{-1} a Higgs of mass between 150 GeV and 170 GeV could be observed with five standard deviations significance, and Higgs masses above $\sim$127 GeV can be excluded to at least 95% confidencel level.

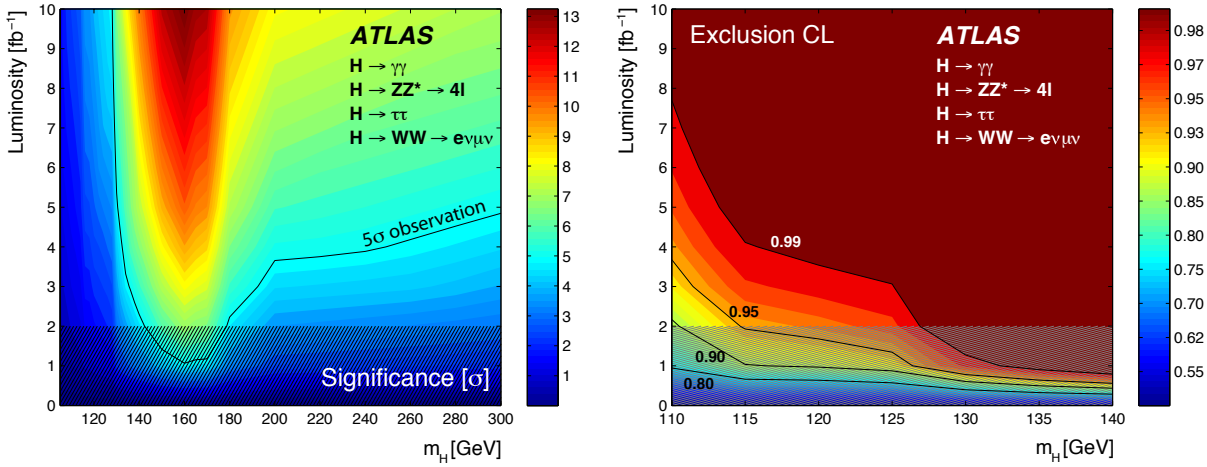


Fig. 82: Standard Model Higgs boson discovery (left) and exclusion potential (right) for given integrated luminosity versus the Higgs mass (ATLAS study). The shaded bands at low integrated luminosity indicate that the results are less accurate (but are expected to be pessimistic).

10.7 Search for phenomena beyond the Standard Model

The primary motivation for the LHC construction is — beyond the discovery of the Higgs boson — the search for signatures from unknown physics at the high-energy frontier, which it is hoped will provide answers to at least part of the current unknowns and problems outlined in Section 1. There is a wealth of models introducing new physics, which is also driven by the relatively few constraints that the high-energy sector must comply with. At any order of magnitude beyond the TeV scale may lurk new symmetries, the breaking of which creates partners of the known Standard Model fields, but which also may lead to a profusion of new particles at ever higher mass scales. Alternatively, in case we live in an apparently severely fine-tuned world, no new physics exists at least in the quark sector up to the reduced Planck scale, leaving a desert of 16 orders of magnitude all described by the Standard Model interactions. This latter picture must probably be regarded as disfavoured, not only by the fine-tuning argument, but since it also contradicts our experience: up to now, each ascent of an order of magnitude in energy has afforded new phenomena in particle physics.

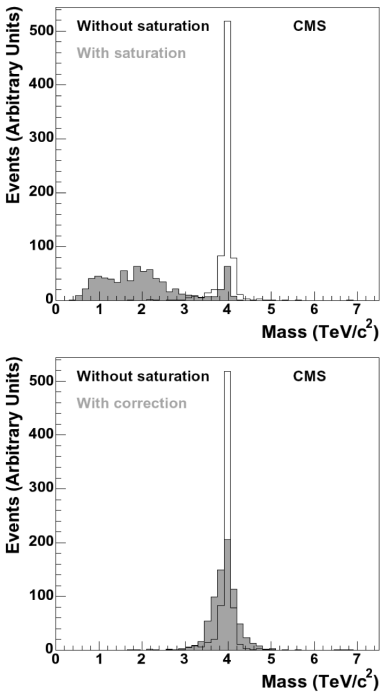


Fig. 83: High energy saturation effect (upper plot) and its correction (lower) in the CMS electromagnetic calorimeter for 4 TeV Randall–Sundrum gravitons decaying to e^+e^-.

Di-lepton resonances at high mass

Popular early searches for new physics involve di-lepton invariant mass spectra, which may exhibit peaks originating from generic Z' resonances present in many beyond the Standard Model scenarios, such as grand unified theories, little Higgs models, Technicolour, and models featuring extra spatial dimensions. The widths of the new resonances may be narrow (such as for Randall–Sundrum gravitons), or broad enough so that they may be resolved in the detector (for example heavy resonances in grand unified theories and little Higgs models, as well as in models with small extra dimensions where the gauge fields are allowed to propagate into the extra-dimensional bulk). The most rigorous direct limits on the existence of heavy neutral particles decaying into di-leptons come from direct searches at the Tevatron, excluding mass scales until approximately 1 TeV (model dependent).

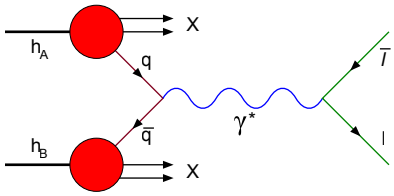


Fig. 84: Feynman graph of a Drell–Yan process (quark–antiquark annihilation to a virtual photon or Z boson) producing a final-state lepton pair.

Contrary to searches with missing transverse energy, which usually do not exhibit clear-cut kinematic signatures, the observation of a di-electron mass peak over (mostly) irreducible Drell–Yan background (Fig. 84) does not require the design calorimeter performance (this is somewhat different for heavy di-muon resonances, where the alignment of the muon system must be well understood to reach good resolution and charge measurement). In case of the search for very high-mass resonances (not

early physics), electromagnetic calorimeter saturation must be corrected (Fig. 83 for CMS). It is also not required to predict the background shapes with Monte Carlo simulation. It can be determined from data by means of a parametrised maximum-likelihood fit with parameters determined simultaneously with the signal abundance by the fit.

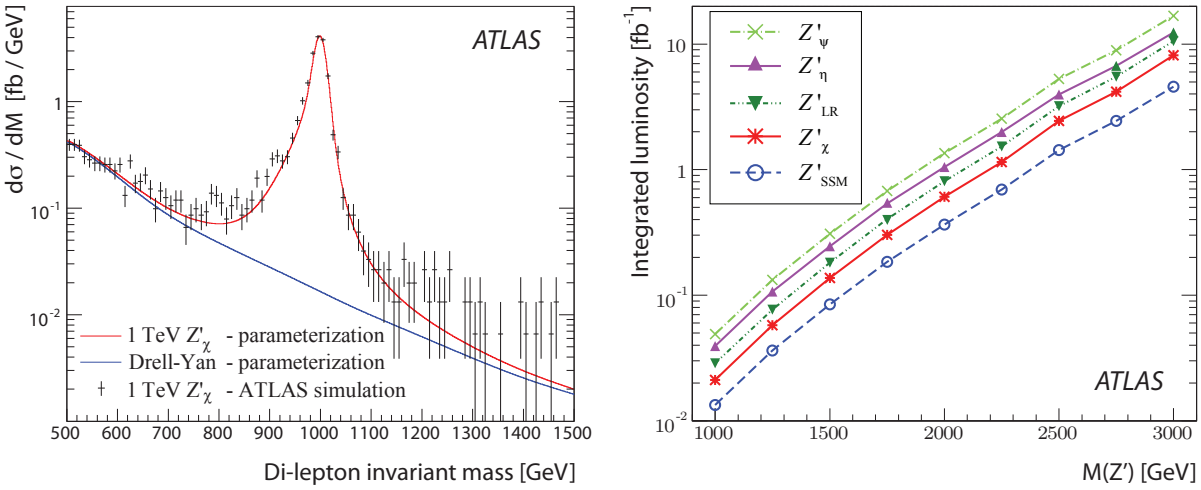


Fig. 85: Left: distribution of the di-electron mass for fully simulated ATLAS data (dots) in presence of a 1 TeV Z'_χ (solid line) and Drell–Yan background (dashed line). The statistics used correspond to 21 fb^{-1}. Right: Required luminosity versus the Z' mass for a 5σ observation according to various Z' models (ATLAS study).

The left panel in Fig. 85 shows a $Z'_\chi \to ee$ peak in ATLAS for a simulated Z' with mass 1 TeV, over Drell–Yan background. The right panel gives the luminosity that is required for a 5σ observation according to various Z' models, as a function of the Z' mass. With 100 pb^{-1} of data, and 14 TeV centre-of-mass energy, Z' (and also W') resonances until a mass of roughly 1 TeV could be discovered. The ultimate goal for ATLAS and CMS reaches about 7 TeV (SLHC prospective).

Statistical considerations

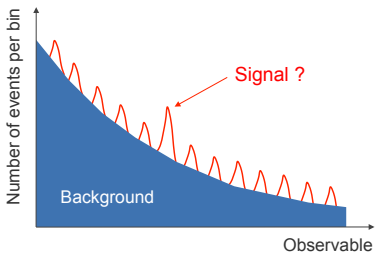


Fig. 86: The p-value quantifying the statistical significance of an observation must be corrected for the statistical trials factor.

The search for di-lepton resonances in a mass range that is large compared with the experimental resolution, and without using prior knowledge about which mass the resonance should have, introduces a statistical 'look-elsewhere effect'. The probability of finding a 6 when playing a dice is 1/6. The probability of finding at least one 6 when playing 2 dice is $2/6 \cdot (1 - 5/6)$. In case of a small single-occurrence probability p and at least one occurrence required, the binomial probability for an occurrence with n trials can be approximated by $n \cdot p$. What counts in the case of the di-lepton invariant mass is the size of the search range in terms of the mass resolution (assuming a negligible intrinsic width of the resonance that is searched for). The number of trials is thus roughly the number of times the resonance 'fits' in the given mass range. Assuming the 'discovery' of a mass peak with a single p-value[37] of 3.0σ (5.0σ) somewhere in the allowed mass range, and assuming 13 independent trials fit into the mass range, the p-value must be corrected by the corresponding *trials factor*. In this case, we find a corrected p-value of 2.1σ (4.5σ). Because of the non-linearity in the rela-

[37] Terminology: the *significance level* of a statistical hypothesis test is the fixed probability of wrongly rejecting the null hypothesis, if it is true. It is the probability for a *Type-I error* to occur. The *p-value* is compared with the significance level and, if it is smaller, the result is significant. It is hence the significance of a single trial.

tion between probability and number of standard deviations, the effect of the correction appears larger at smaller significance of the observation.

The above exercise is a very rough approximation. In practice the evaluation of the trials factor is complex, and the conceptually simplest way to take it into account is via toy Monte Carlo simulation. A natural way to proceed is to perform an unbinned maximum-likelihood fit by describing the background by a simple parametrised function, with parameters determined by the fit, and the signal by a Gaussian or crystal-ball shaped function with predetermined width (obtained from Monte Carlo simulation, but taking into account the mass dependence of the calorimeter or tracking resolution) of which only the mean mass parameter is free to vary in the fit. Also determined by the fit are the signal and background abundances. The fit will converge towards 'some' signal yield at 'some' mass value. To obtain a relative likelihood estimator, the fit is repeated by fixing the signal yield to zero, and the difference between the log-likelihood estimators of the two fits is computed (the fit with free signal yield and mean mass always has a larger log-likelihood value, so that the difference is positive). The p-value of the observed log-likelihood difference is obtained by repeating the same exercise many times with a background-only Monte Carlo model faithfully describing the data. This Monte Carlo model is obtained by using the results from the background parametrisation obtained by the fit to data. The p-value is given by the ratio of the number of cases in which the log-likelihood difference in the Monte Carlo is found to be larger than the one in the data, divided by the total number of trials.

Supersymmetry

In spite of the many creative and interesting new physics models that have appeared in recent years, supersymmetry remains the most popular Standard Model extension. It features an elegant solution of the hierarchy problem by cancelling the diverging weak boson radiative corrections to all orders (where, however, a logarithmic divergence remains due to supersymmetry breaking), a dark matter candidate, natural elementary scalar particles, the democratisation of the fermionic and bosonic degrees of freedom, and grand unification of the electroweak and strong forces. The minimal supersymmetric Standard Model introduces a conserved supersymmetry-parity, denoted *R-parity*, which is even for all Standard Model particles (including a Higgs doublet), and odd for all supersymmetric partners of these.[38] A consequence of R-parity conservation is that the lightest supersymmetric particle (LSP) is stable. Since we have not observed any strongly or electromagnetically interacting particles in the universe that are not included in the Standard Model, and because we need a cold dark matter candidate, it is assumed that the LSP is weakly interacting only (as are neutrinos). The primary LSP candidate is the lightest neutralino, a linear combination of gauginos. In much of the supersymmetry parameter space the neutralino is a mixture of photino and zino, but could also be a gravitino. R-parity conservation also implies that supersymmetric particles can only be produced in pairs. Hence, to produce supersymmetry in a hadronic interaction the centre-of-mass energy of the colliding partons must be twice the characteristic supersymmetric mass scale.

A typical decay cascade of a supersymmetric squark or gluino is depicted in the right-hand plot of Fig. 77. From the diagram one notices that supersymmetric events produce many high-p_T jets, sometimes leptons, and always missing transverse energy due to the escaping LSP (unless it escapes along the beam pipe). Since squarks and gluinos are produced by strong interactions with $\mathcal{O}$(picobarn) cross sections if their masses are well below a TeV, and because supersymmetric events have a clear experimental signature, supersymmetry could be detected quite early. An integrated luminosity of $100\,\mathrm{pb}^{-1}$ is expected to be sufficient for a discovery of relatively low-mass supersymmetry, provided that the Standard Model backgrounds can be well controlled.

Figure 87 shows distributions of missing transverse energy in ATLAS for simulated supersymmetry signal and Standard Model background events, and for analyses with (right panel) and without (left

[38] R-parity, defined by $R = (-1)^{2S+3B+L}$, was originally introduced to avoid the proton decay $p \to e^+\pi^0$, which is possible in supersymmetry.

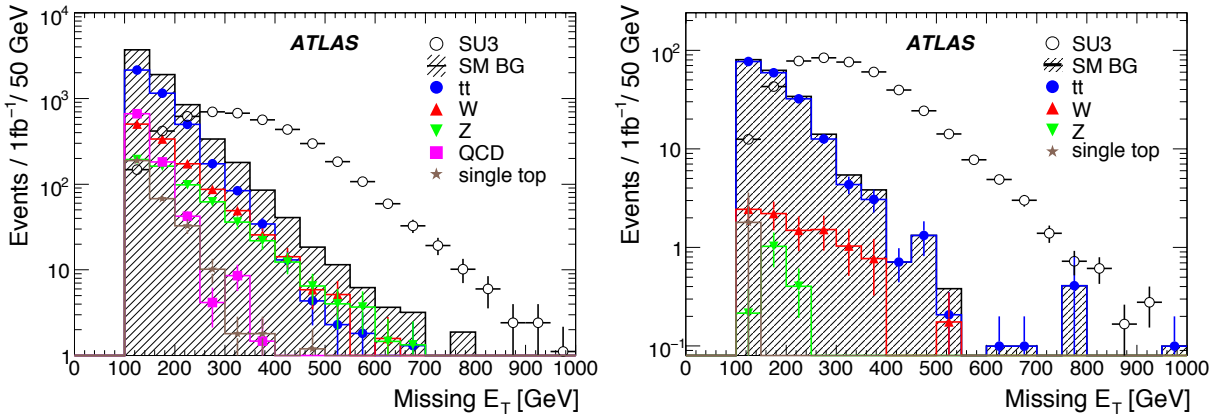


Fig. 87: Simulated distributions of missing transverse energy in ATLAS for analyses without (left) and with (right) requiring a reconstructed lepton (electron or muon) in the detector. Shown are the contributions from Standard Model processes and for R-parity-conserving supersymmetry using a minimal supergravity model (open circles) with the parameters $m_0 = 100\,\text{GeV}$ and $m_{1/2} = 300\,\text{GeV}$. The number of events corresponds to $10\,\text{fb}^{-1}$ integrated luminosity.

panel) requiring a reconstructed electron or muon. A clear signal excess is perceptible in both analyses, but the main Standard Model backgrounds differ significantly between the two. Whereas without lepton requirement, $t\bar{t}$, and W and Z plus jets backgrounds are of similar size in the large-$\not\!\!E_T$ tails, and background from jets (QCD) is also present, the background in the one-lepton analysis is entirely dominated by $t\bar{t}$, with some small contributions from W and jets, but no QCD jets background. This makes the one-lepton analysis particular interesting for the initial running period, when the understanding of the inclusive QCD background is still immature.

Other discriminating variables used in supersymmetry searches are the 'effective mass', which is the scalar sum of the transverse momenta of all jets and leptons (other variations of this variable also include $\not\!\!E_T$, or do not include the lepton momentum), and the transverse mass [see Eq. (12)] which is particularly useful to reduce background from events with a W. Figure 88 shows an event display of a typical supersymmetric event with jets, muons and large $\not\!\!E_T$ in ATLAS.

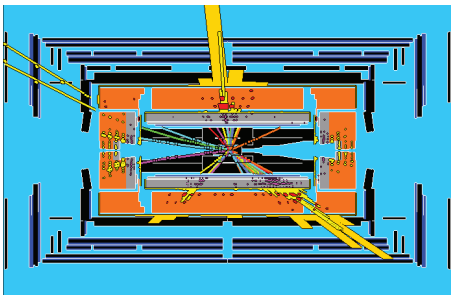


Fig. 88: Simulated supersymmetric event in ATLAS with six particle jets and two muons with opposite charge in the final state, and with large missing transverse energy.

Figure 89 shows the expected discovery potential for the minimal supergravity model as a function of the GUT mass parameters m_0 and $m_{1/2}$ (ATLAS study). The zero-lepton analysis has the best discovery reach. However, taking into account the experimental difficulties of this mode, the one-lepton mode may become competitive. Squarks and gluinos with masses up to 0.75, 1.35, 1.8 TeV can be discovered with integrated luminosities of 0.1, 1 and $1\,\text{fb}^{-1}$, respectively, using the four-jet, zero-lepton analysis.

We should note that supersymmetry could also break R-parity. The signature could be taus originating from $\chi_1^0 \to \tilde{\tau}\tau$ decays. Moreover, signals due to other phenomena could be seen like supersymmetry so that a (challenging) neutralino spin analysis needs to be performed to reveal their fermionic nature. Experimentalists should proceed with the search for supersymmetry as model-independently as possible, and watch out for anomalies, e.g., the occurrence of photons, taus, or strange tops

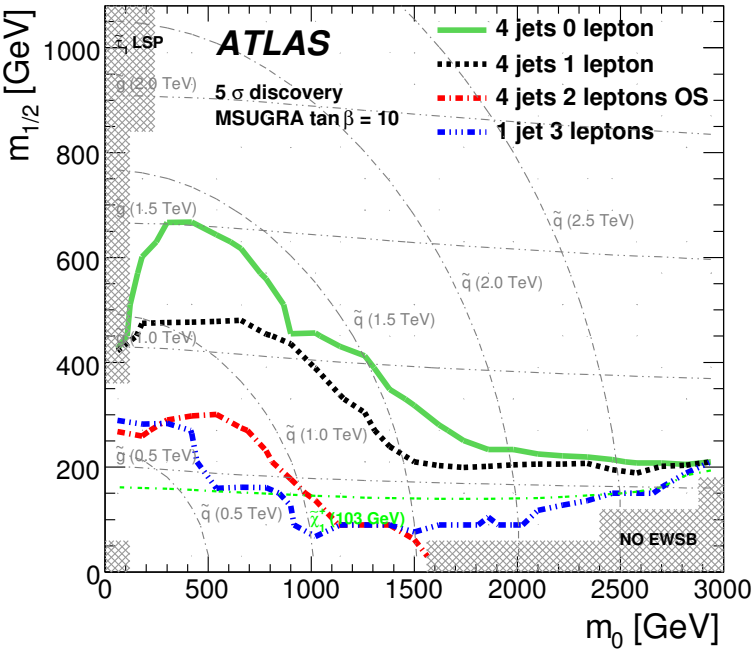


Fig. 89: Expected contours for 5σ minimal supergravity discovery versus the GUT mass parameters $m_{1/2}$ and m_0, for analyses with various lepton requirements and for an integrated luminosity of $1\,\mathrm{fb}^{-1}$. The grey dashed contour lines indicate the corresponding squark and gluino masses. The zero-lepton analysis has the best discovery reach.

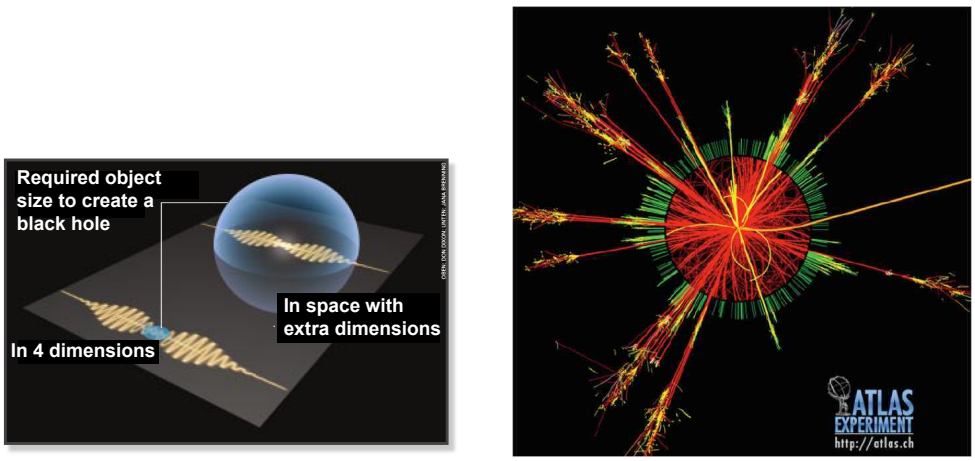


Fig. 90: Left: Schwarzschild radius in 4 and $4+d$ dimensions. Right: Simulation of a black hole decay in ATLAS.

Strong gravity

Finally, if we are allowed to enter trans-Planck scales, that is, gravity in compact extra spatial dimensions is strong enough to reduce the Planck scale to energies reached by the LHC, hard-scattering proton–proton collisions may produce microscopic black holes. An object becomes a black hole if it is smaller than the Schwarzschild radius $r = 2GM/c^2$. In $4+d$ spatial dimensions the Schwarzschild radius[39] becomes $r = 2G^{(4+d)}M_D/c^2$, where $G^{(4+d)}$ is a gravitational constant in the full-dimensional space. The four-dimensional constant G is thus only a reflection of the real gravitational constant $G^{4+d)}$, reduced

[39]The Schwarzschild radius is the radius below which the gravitational attraction between the particles of a body is so strong that the body undergoes gravitational collapse. For a typical star such as the Sun, the Schwarzschild radius is about 3 km.

('diluted') by the extra dimensions (see Fig. 90, Left). The Planck scale is no longer fundamental. If $M_D \approx M_{\text{Planck}}^{(4+d)} \approx 1$ TeV, a black hole can be produced by the LHC if the momentum transfer of the hard scattering reaction exceeds M_D. The cross section of the black hole production is $\sigma_{\text{BH}} \approx \pi r^2$. With $M_D \sim 2$–3 TeV one finds $\sigma_{\text{BH}} \sim \mathcal{O}(\text{pb})$ allowing a fast discovery for $M_{\text{BH}} < 4$ TeV, and $d = 2$–6.

The black hole undergoes a fast ($\tau_{\text{BH}} \sim 10^{-27}$ s) thermal decay via Hawking radiation of temperature $T_H \sim M_D \cdot (M_D/M_{\text{BH}})^{1/(d+1)}$ (a microscopic black hole is not black at all!). The life cycle of a 10 TeV black hole could be sketched as follows: (i) $\Delta t = 0$, $M_{\text{BH}}(\Delta t) = 10$ TeV: *creation* — the micro black hole is created in a proton–proton collision: it is asymmetric, may vibrate and rotate, and may be electrically charged; (ii) $\Delta t = 0 - 1 \cdot 10^{-27}$ s, $M_{\text{BH}}(\Delta t) 10 - 8$ TeV: *'baldness phase'* — emission of gravitational and electromagnetic waves, and charged particles, the black hole is solely characterised by mass and angular momentum; (iii) $\Delta t = 1 - 3 \cdot 10^{-27}$ s, $M_{\text{BH}}(\Delta t) = 8 - 6$ TeV: *slowing down* — the black hole radiates by reducing its angular momentum, its form becomes spherical; (iv) $\Delta t = 3 - 20 \cdot 10^{-27}$ s, $M_{\text{BH}}(\Delta t) = 6 - 2$ TeV: *Schwarzschild phase* — after losing its angular momentum, the micro black hole evaporates its mass via Hawking radiation; (v) $\Delta t = 20 - 22 \cdot 10^{-27}$ s, $M_{\text{BH}}(\Delta t) = 2 - 0$ TeV: *Planck phase* — the black hole shrinks down to the Planck mass (M_D) and fully decays into all particles with probabilities according to their degrees of freedom (Fig. 90, Right). The spectacular decay signature cannot be missed by the experiments.

11 Conclusions and outlook

Commissioning such tremendously complex apparatus as the LHC high-p_T experiments ATLAS and CMS is a continuous challenge. It starts far earlier than with the installation of the experiments in their underground caverns. To some extent it already begins with the design phase, when prototypes are drawn, simulated, and eventually built for the purpose of testing and optimisation. Commissioning continues in dedicated test beams where parts or even complete slices of the detectors, modelling as accurately as possible the final geometry, are assembled. While installing the detectors at their final locations, commissioning campaigns with cosmic ray events are undertaken. Hundreds of millions of cosmic rays have been recorded by both experiments in roughly three years of data taking with more and more complete detectors. Finally, with the start of the LHC commissioning, single beams with 900 GeV injection energy are sent through both LHC beam pipes, circulating or as beam-on-collimator 'splash' dumps, radio-frequency captured or not. Later two beams are injected, again at injection energy, radio-frequency captured, and brought to collision. These collisions produce for the first time so-called minimum bias events, producing roughly 20 tracks in the inner tracking systems, some photons from π^0 and η decays, and electrons from photon conversion, as well as rare jet events and muons from pion and kaon decays. The beams will not be squeezed at this initial stage so that owing to the large beam spot, the small number of bunches in the machine, and the low bunch intensity, the peak luminosity will not exceed $10^{27}\,\text{cm}^{-2}\text{s}^{-1}$. However, once the LHC energy is ramped up, the relativistic contraction of the beam will lead to an increase in the luminosity, and the experiments will see an increase in jet rates, as well as electrons and muons mainly from heavy quark decays and quarkonia. Moreover, beam squeezing (i.e., the reduction of the beam envelope by the magnet optics) and a crossing angle between the colliding beams will further allow an increase in the luminosity of the LHC at higher energy.

With the data taken during these commissioning phases, the experiments have gained experience and obtained a good initial understanding of the detector response, and improved the quality of the data by calibrating and aligning the detector subsystems, which will pay off when analysing the first collision data for physics and detector performance. With the arrival of physics data it is very important to continue improving the detector understanding, and the faithfulness of its description by the detector

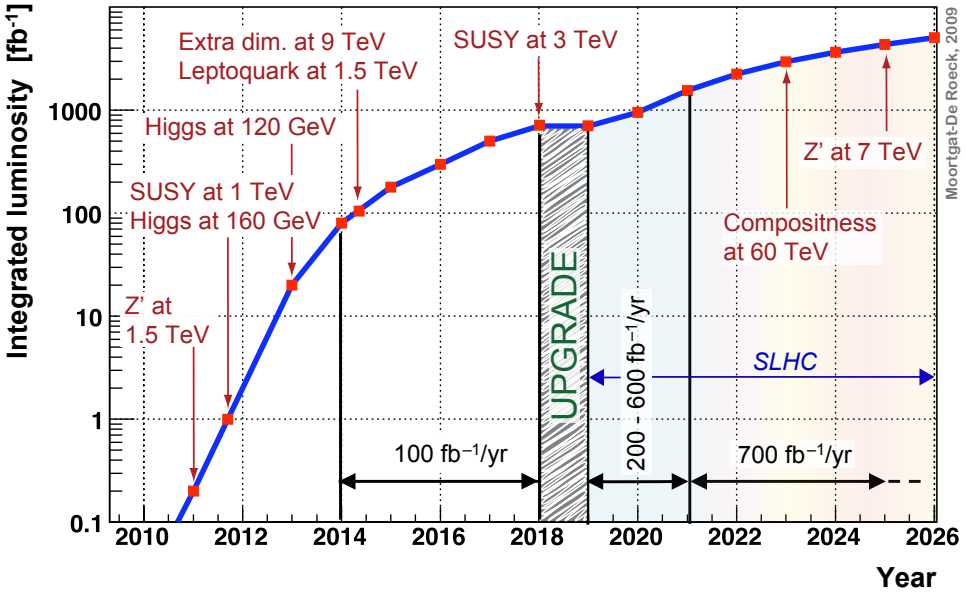


Fig. 91: The LHC programme in a nutshell [28] (see text for discussion).

response simulation. It is the key to a longterm success of the experiments, and to physics results with the smallest possible bias and systematic errors. It is also important that the experiments optimise the fraction of useful data taken, by steadily improving the data-taking efficiency of all detector systems, and aiming at the best achievable data quality.

Figure 91 gives an exploratory view of the expected LHC performance versus year of (design) operation at 14 TeV centre-of-mass energy [28], and the corresponding sensitivity for discovery of various phenomena by the ATLAS and CMS experiments. After accumulating $1\,\mathrm{fb}^{-1}$ integrated luminosity, minimal supersymmetry with up to 1 TeV characteristic mass scale could be discovered. The Standard Model Higgs boson is expected to be observed at any mass with $30\,\mathrm{fb}^{-1}$. With the ultimate integrated luminosity of possibly $500\,\mathrm{fb}^{-1}$ around the year 2018, the discovery reach for many new physics models can be pushed deep into the TeV scale, and properties of earlier discoveries may be studied. Among these are the coupling strengths of the Higgs boson in various production and decay channels. If the Higgs is observed to decay into either $\gamma\gamma$ or $ZZ^{(\star)}$, one will know that it cannot have spin 1. Observations of angular distributions and correlations in $ZZ^{(\star)}$ decays will enable the spin and *CP* properties of the Higgs to be determined. It should also be possible to constrain masses of supersymmetric particles, possibly even the spin of a heavy neutralino. A spin analysis of heavy resonances decaying to di-leptons could be performed in case of a discovery. After four years at the highest peak luminosity with approximately $100\,\mathrm{fb}^{-1}$ of data recorded each year, the increase in sensitivity becomes asymptotic (recall the $1/\sqrt{\mathscr{L}}$ scaling of statistical errors), which is the opportunity to undertake an upgrade of machine and detectors to the Super-LHC (SLHC). The SLHC programme proposes to increase the LHC peak luminosity to $1.5{\cdot}10^{35}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$, i.e., 10 times the nominal LHC peak luminosity [29]. At nominal bunch pattern, it will compel the experiments to cope with 250 pile-up minimum bias interactions occurring in time with the hard-scattering event. This requires many changes to the detectors: (*i*) reduce background rates by changing the beam pipe and improving the shielding, (*ii*) improve the radiation and occupancy tolerance of the detectors and electronics, in many cases by replacing entire subsystems, and (*iii*) increase the bandwidth of front-end and readout electronics to minimise pile-up and handle a 10 times increase in the event rate. A successful SLHC upgrade would allow the experiments to extend their discovery reach for supersymmetry and Z' bosons to 4 TeV and 7 TeV, respectively.

The current situation, however, is that, owing to limitations in the quench protection system, the LHC will begin in 2010 with 7 TeV centre-of-mass energy, which later in the year may or may not be

increased to a maximum of 10 TeV. ATLAS and CMS have performed indicative studies to evaluate the impact of the reduced energy on their physics programme. It is expected that up to half an inverse femtobarn of data will be delivered in 2010. (If the run is continued through 2011 a total of one inverse femtobarn of data could be delivered.) Between 14 TeV and 10 TeV the number of selected $Z \to ee$ events will decrease from roughly 5000 per $10\,\mathrm{pb}^{-1}$ integrated luminosity to 3600 (linear relationship). The number of produced $t\bar{t}$ events will drop by roughly a factor of 2, so that the sample size will attain that of the Tevatron after approximately $100\,\mathrm{pb}^{-1}$ at 10 TeV. The exclusion of a Higgs boson requires about twice more integrated luminosity at 10 TeV than at 14 TeV. A 5σ discovery of a Higgs with mass of 160 GeV (which is unlikely) would require roughly $1\,\mathrm{fb}^{-1}$ of recorded physics data. To challenge the Tevatron Higgs searches, a sample of about $200\,\mathrm{pb}^{-1}$ at 10 TeV is needed. The sensitivity of the search for a heavy Z' is reduced by a factor of roughly 3 at 10 TeV. A 5σ observation of a 1 TeV (Tevatron limit) weighing Z' would require roughly $100\,\mathrm{pb}^{-1}$ of 10 TeV collision data. To achieve an equivalent discovery reach for supersymmetry, a factor of 2 more integrated luminosity is required at 10 TeV centre-of-mass energy. Nevertheless, the current Tevatron limits can be improved with as little as $20\ \mathrm{pb}^{-1}$ of 10 TeV data. What would be the impact of a 7 TeV centre-of-mass energy compared to 10 TeV? The number of $Z \to ee$ will drop by a another factor of 1.4. The $t\bar{t}$ rate will further drop by approximately a factor of 2. The required luminosity for equal search sensitivity for a Z' will increase by a factor of 3, similarly for supersymmetry searches, and a factor of 2–3 for Higgs searches.

References

[1] ATLAS Collaboration, JINST **3**, S08003 (2008).

[2] CMS Collaboration, JINST **3**, S08004 (2008).

[3] ATLAS Collaboration, *Expected Performance of the ATLAS Experiment: Detector, Trigger and Physics*, CERN-OPEN-2008-020, arXiv:0901.0512 (2008).

[4] CMS Collaboration, *CMS Physics: Technical Design Report*, Volume I: CMS-TDR-008-1, CERN-LHCC-2006-001; Volume II: CMS-TDR-008-2, CERN-LHCC-2006-021 (2006).

[5] CDF Collaboration (T. Aaltonen *et al.*), Phys. Rev. Lett. **102**, 031801 (2009); D0 Collaboration, Conference Note 5923-CONF (2009).

[6] H. Flächer *et al.*, Eur. Phys. J. C **60**, 543 (2009) [arXiv:0811.0009], http://cern.ch/gfitter.

[7] ALEPH Collaboration (S. Schael *et al.*), Phys. Rep. **421**, 191 (2005) [arXiv:hep-ex/0506072].

[8] M. Davier, S. Descotes-Genon, A. Höcker, B. Malaescu and Z. Zhang, Eur. Phys. J. C **56**, 305 (2008) [arXiv:0803.0979].

[9] CKMfitter Group (J. Charles *et al.*), Eur. Phys. J. C **41**, 1 (2005) [arXiv:hep-ph/0406184], http://ckmfitter.in2p3.fr.

[10] A. Hoecker and W. Marciano, The muon anomalous magnetic moment (in: Review of Particle Physics 2008), Phys. Lett. B **667**, 1 (2008), updated 2009 on http://pdglive.lbl.gov.

[11] J. Ellis, *Beyond the Standard Model for Montañeros*, lectures at this School.

[12] A.D. Martin, R.G. Roberts, W.J. Stirling, and R.S. Thorne, Eur. Phys. J. C **14**, 133 (2000) [arXiv:hep-ph/9907231].

[13] The Tevatron Electroweak Working Group, FERMILAB-TM-2439-E, arXiv:0908.1374 (Aug 2009).

[14] J. Nash, *Instrumentation for the LHC experiments*, lectures at this School.

[15] D. Froidevaux and P. Sphicas, Annu. Rev. Nucl. Part. Sci. 56, 375 (2006).

[16] N. Ellis, *Trigger and data acquisition*, lectures at this School.

[17] T. Gaisser, *Cosmic Rays and Particle Physics* (Cambridge University Press, 1990).

[18] S. Aefsky *et al.*, J. Instrum. **3**, P11005 (2008).

[19] T. Hebbeker and C. Timmermans, Astropart. Phys. **18**, 107 (2002) [arXiv:hep-ph/0102042].

[20] MINOS Collaboration (P. Adamson *et al.*), Phys. Rev. D **76**, 052003 (2007) [arXiv:0705.3815].

[21] CMS Collaboration [M. Aldaya, P. Garcia-Abia], CMS NOTE-2008/016 (2008).

[22] ATLAS Collaboration, arXiv:0912.2642 (2009).

[23] T. Sjöstrand, *QCD at LHC in pp*, Talk given at 3rd Nordic *LHC and Beyond* Workshop, Lund, Sweden, 2009.

[24] S. Van der Meer, CERN internal report, ISR-PO/68-31, 1968.

[25] K.A. Brown, M. Blaskiewicz, C. Degen, and A. Della Penna, Phys. Rev. Spec. Topics — Accelerators and beams **12**, 012801 (2009).

[26] CDF Collaboration, Phys. Rev. D **79**, 112005 (2009) [arXiv:0904.1098].

[27] CDF Collaboration, Phys. Rev. Lett. **103**, 092002 (2009) [arXiv:0903.0885]; D0 Collaboration, Phys. Rev. Lett. **103**, 092001 (2009) [arXiv:0903.0850].

[28] A. de Roeck and F. Moortgat, *Private communication* (2010).

[29] SLHC web site with references to project papers: http://project-slhc.web.cern.ch/project-slhc.

Student projects

High-energy cosmic-ray acceleration

M. Bustamante[1]*, G.D. Carrillo Montoya*[2]*, W. de Paula*[3]*, J.A. Duarte Chavez*[4]*, A.M. Gago*[1]*, H. Hakobyan*[5]*, P. Jez*[6]*, J.A. Monroy Montañez*[7]*, A. Ortiz Velasquez*[8]*, F. Padilla Cabal*[9]*, M. Pino Rozas*[10]*, D.J. Rodriguez Patarroyo*[11]*, G.L. Romeo*[12]*, U.J. Saldaña-Salazar*[13]*, M. Velasquez*[14] *and M. von Steinkirch*[15]

[1]Pontificia Universidad Católica del Perú, Lima, Peru
[2] EPFL, Switzerland and U. of Wisconsin, USA
[3] Instituto Tecnológico de Aeronáutica, São Josè dos Campos, Brazil
[4] Universidad Nacional de Colombia - Sede Bogota, Colombia
[5] Universidad Tecnica Federico Santa Maria, Valparaiso, Chile
[6] Niels Bohr Institute, Copenhagen, Denmark
[7] Universidad de los Andes, Bogota, Colombia
[8] Instituto de Ciencias Nucleares, Mexico
[9] InSTEC, La Habana, Cuba
[10]Pontificia Universidad Católica de Chile, Santiago, Chile
[11] Universidad Antonio Narino, Bogota, Colombia
[12] Universidad de Buenos Aires, Argentina
[13] UNAM, Mexico
[14] Universidad de Antioquia, Medellin, Colombia
[15] University of São Paulo, Brazil

Abstract

We briefly review the basics of ultrahigh-energy cosmic-ray acceleration. The Hillas criterion is introduced as a geometrical criterion that must be fulfilled by potential acceleration sites, and energy losses are taken into account in order to obtain a more realistic scenario. The different available acceleration mechanisms are presented, with special emphasis on Fermi shock acceleration and its prediction of a power-law cosmic-ray energy spectrum. We conclude that first-order Fermi acceleration, though not entirely satisfactory, is the most promising mechanism for explaining the ultra-high-energy cosmic-ray flux.

A copy of the slides presented during the oral report at the school can be found at the URL below
`http://cern.ch/PhysicSchool/LatAmSchool/2009/Presentations/pDG1.pdf`

1 Introduction

In 1912, Victor Hess, using a balloon flight, measured the intensity of the ionizing radiation as a function of altitude. This date represents the begining of the history of cosmic rays. Since then, we have learned about many of their features, such as their large energy span (1–10^{20} eV), their composition (they are made up of protons, nuclei, electrons and other charged particles), and the behaviour, as a function of energy, of their flux.

However, the source and origin of the highest-energy cosmic rays still elude us [1, 2]. There are two general approaches: in top-down scenarios [3], cosmic rays are produced as secondaries of the decay of heavy particles, while in bottom-up scenarios, the energetic cosmic-ray protons and nuclei are accelerated within regions of intense magnetic fields. During recent years, experiments like AGASA [4, 5] and HiRes [6] have been trying to answer these questions. A newly-built experiment, the Pierre Auger Observatory, has performed observations [7] that hint at active galactic nuclei—galaxies with a supermassive central black hole—as sources of the highest-energy cosmic rays. A plot of the differential cosmic-ray energy spectrum, produced with data from several experiments, is shown in Figure 1.

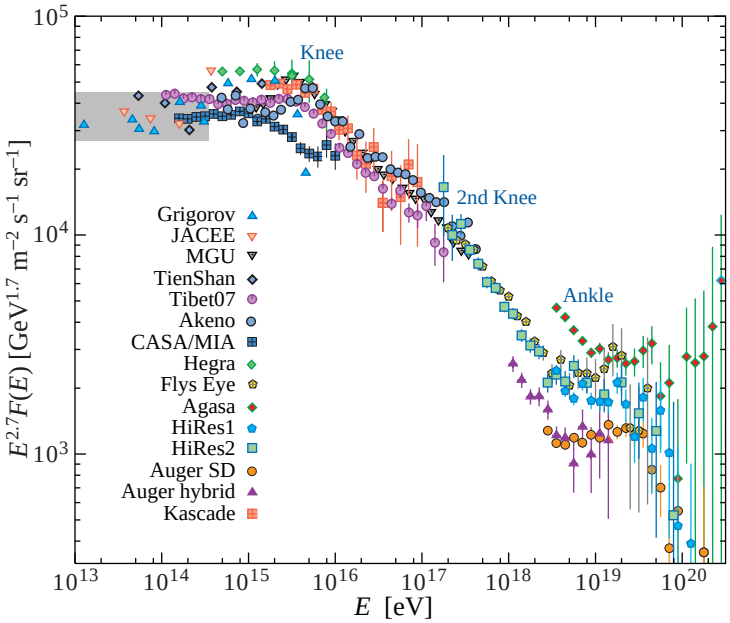


Fig. 1: Cosmic-ray differential energy spectrum, reconstructed from air showers observed by various experiments. The grey box is the region where direct observations of cosmic rays have been made. The spectrum has been multiplied by $E^{2.7}$ to enhance the kinks due to changes in the spectral index: the first one near $10^{15} - 10^{16}$ eV (the *knee*), the second one at 10^{17} eV (the *second knee*) and the last one around 10^{19} eV (the *ankle*). Figure extracted from Ref. [8]

The purpose of this review is to give a brief description of the general constraints on acceleration sites, as well as of the first- and second-order Fermi acceleration mechanism. For a more in-depth review of the theory and observation of cosmic rays, the reader can consult, for example, Refs. [9, 10].

2 General constraints on acceleration sites

In order to be considered as a possible source of ultra-high-energy cosmic rays (UHECRs), an astrophysical object has to fulfil several conditions [11]:

- **geometry**: the accelerated particle should be maintained within the object during the acceleration process;
- **power**: the source should be able to provide the necessary energy for the accelerated particles;
- **radiation losses**: within the accelerating field the energy gained by a particle should be no less than its radiation energy loss;
- **interaction losses**: the energy lost by a particle due to its interaction with other particles should not be greater than its energy gain;
- **emissivity**: the density and power of sources must be enough to account for the observed UHECR flux;
- **coexisting radiation**: the accompanying photon and neutrino flux, and the low-energy cosmic-ray flux, should not be greater than the observed fluxes (this constraint must be satisfied by the flux from a single source and by the diffuse flux).

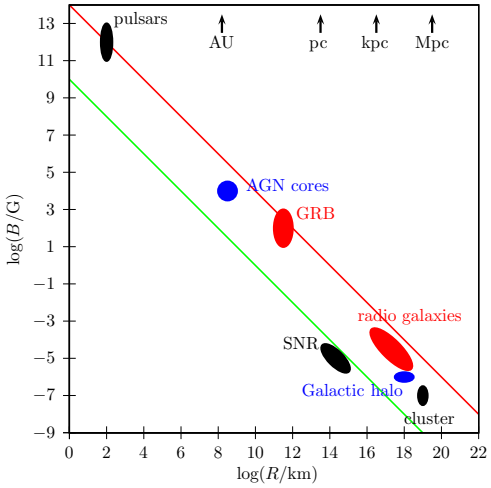


Fig. 2: Hillas plot. Sources above the top (red) line are able to accelerate protons up to 10^{21} eV, while sources above the bottom (green) line are able to accelerate iron up to 10^{20} eV. Figure reproduced from Ref. [12]

2.1 The Hillas criterion

If a particle escapes from the region where it was being accelerated, it will be unable to gain more energy. This situation imposes a limit on its maximum energy that can be expressed as follows:

$$\varepsilon_{\max} = qBR\,, \tag{1}$$

where q is the electric charge of the accelerated particle, B is the magnetic field, and R is the size of the accelerator. Equation (1) is obtained by demanding that the Larmor radius of the particle, $R_L = \varepsilon/(qB)$, not exceed the size of the acceleration region. This is a general geometrical criterion known as the *Hillas criterion*, and is useful in selecting potential acceleration sites.

Figure 2 is an example of a Hillas plot which, for a given maximum energy $\varepsilon_{\max}$ of the accelerated particle, shows the relation between the source's magnetic field strength B and its size R. Sources above the top line are able to accelerate protons up to 10^{21} eV, while sources above the bottom line are able to accelerate iron up to 10^{20} eV.

A more realistic description of particle acceleration takes into account the energy lost during the process. The maximum energy that a particle can obtain in an accelerator if energy losses are accounted for is given by the solution of $d\varepsilon^{(+)}/dt = d\varepsilon^{(-)}/dt$, i.e., the situation where energy lost and gained is equal. The maximum energy of the particle is hence given by the minimum between the value obtained from this equality and the one obtained from the Hillas criterion. Hillas plots for proton and iron taking into account energy losses are shown in Figure 3.

UHECRs are believed to have both a galactic (for energies below the knee) [13] and an extragalactic (above the knee) component [14]. Some potential galactic sources include type II supernovae, pulsars and shock acceleration in supernova remnants, while extragalactic ones include active galaxies and gamma-ray bursts.

3 General forms of acceleration

3.1 Inductive acceleration mechanism

This mechanism is also called *one-shot acceleration* and occurs when a particle is accelerated in a continuous way by an ordered field [see Figure 4(a)]. Radiation losses from accelerated charged particles moving at relativistic velocities are composed of two terms [11], attributed to synchrotron and curvature radiation.

3.1.1 One-shot acceleration with synchrotron-dominated losses

In this regime the maximum energy is given by

$$\varepsilon_s = \sqrt{\frac{3}{2}} \frac{m^2}{q^{3/2}} B^{-1/2} \, , \tag{2}$$

where B is the strength of the magnetic field, and m, q are the mass and charge of the particle, respectively. This notation will be valid for the sections below.

3.1.2 One-shot acceleration with curvature-dominated losses

In the special case when $\overrightarrow{\mathbf{v}} // \overrightarrow{\mathbf{E}} // \overrightarrow{\mathbf{B}}$, curvature losses dominate. This might be the situation in the vicinity of neutron stars and black holes. The corresponding maximum energy is

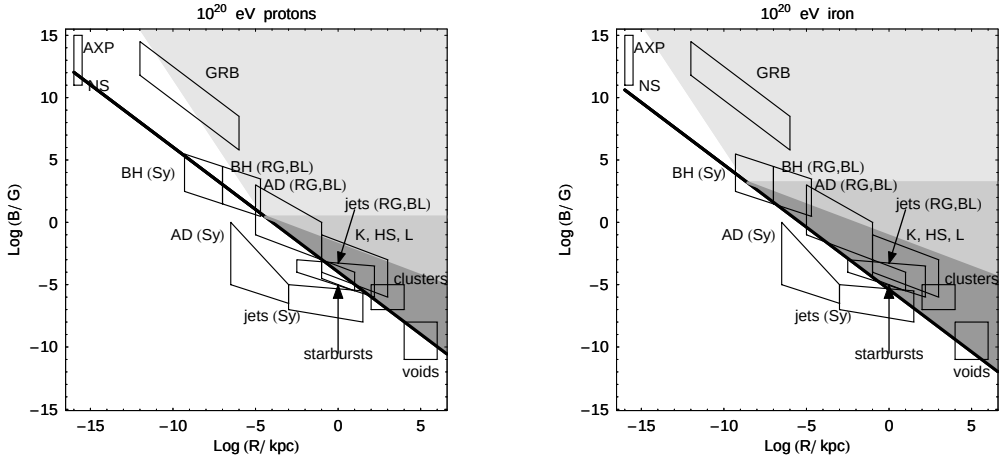
$$\varepsilon_c = \frac{3}{2}^{1/4} \frac{m}{q^{1/4}} B^{1/4} R^{1/2} \, . \tag{3}$$

3.2 Diffusive acceleration

In this mechanism the particle is accelerated in bursts, as a result of its interaction with regions of high magnetic field intensity, as shown in Figure 4(b). The maximum energy, considering synchrotron-dominated losses, is [11]

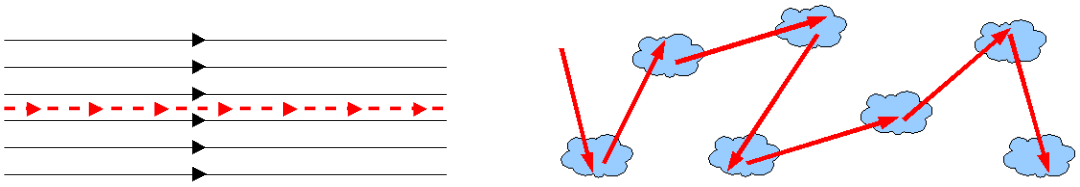
$$\varepsilon_d \simeq \frac{3}{2} \frac{m^4}{q^4} B^{-2} R^{-1} \, . \tag{4}$$

Diffusive acceleration, and in particular Fermi acceleration (see next Section) is the preferred acceleration mechanism in bottom-up scenarios of cosmic-ray production.



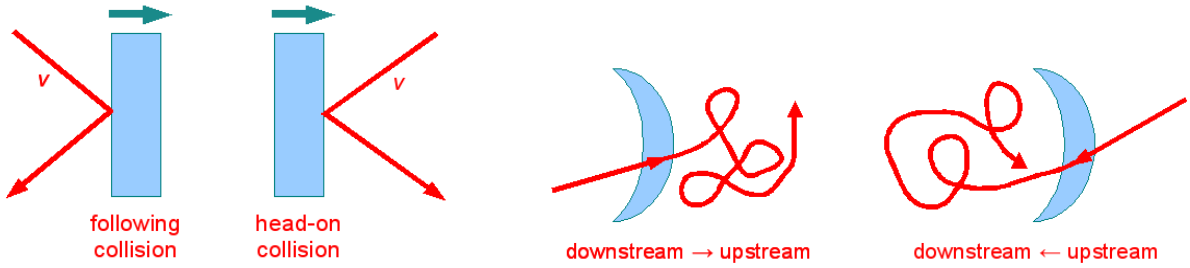
(a) (b)

Fig. 3: (a) Hillas plot for 10^{20} eV protons, including energy losses. The thick line is the lower boundary due to the Hillas criterion. The light grey region is allowed by one-shot acceleration with curvature-dominated losses, the grey region is allowed by one-shot acceleration with synchrotron-dominated losses, and the dark grey region is allowed by both one-shot and diffusive acceleration. **(b)** Same plot for 10^{20} eV iron nuclei. Figures reproduced from Ref. [11]



(a) (b)

Fig. 4: (a) One-shot acceleration. **(b)** Diffusive shock acceleration



(a) (b)

Fig. 5: (a) Second-order Fermi acceleration. **(b)** First-order Fermi acceleration.

4 Fermi acceleration

4.1 Second-order Fermi acceleration

This first version of the Fermi acceleration mechanism (later dubbed *second-order acceleration*) was proposed by Enrico Fermi in 1949 [15] and explains the acceleration of relativistic particles by means of their collision with interstellar clouds. These clouds move randomly and act as 'magnetic mirrors', so that the particles are reflected off them, as shown in Figure 5(a).

After some calculations [12, 16] it can be shown that the average energy gain per collision is

$$\left\langle \frac{\Delta E}{E} \right\rangle = \frac{8}{3} \left(\frac{v}{c} \right)^2 , \tag{5}$$

where v and c are the speed of the cloud and of the particle, respectively. The average energy gain is proportional to $(v/c)^2$: the process is known as "second-order" acceleration owing to the value of the exponent. If we calculate the average time between collisions, an energy rate can be derived from Equation (5):

$$\frac{dE}{dt} = \frac{4}{3} \left(\frac{v^2}{cL} \right) E = \alpha E, \tag{6}$$

where L is the mean free path between clouds, along the field lines. It is possible to find the energy spectrum $N(E)$ by solving a diffusion-loss equation in the steady state and considering this energy rate, plus the assumption that τ_{esc} is the characteristic time for a particle to remain in the accelerating region. In so doing, one finds that

$$N(E)\, dE = \text{const.} \times E^{1+\frac{1}{\alpha \tau_{\text{esc}}}}\, dE\ . \tag{7}$$

Even though second-order acceleration succeeds in generating a power-law spectrum, it is not a completely satisfactory mechanism. First, on account of the observed low cloud density, the energy gain is very slow. Second, the mechanism fails to explain the observed value of 2.7 for the exponent in the power-law spectrum: the value of the exponent is determined by the uncertain value of the combination $\alpha \tau_{\text{esc}}$.

4.2 First-order Fermi acceleration

Before we discuss first-order Fermi acceleration it is convenient to formulate the Fermi mechanism in a more general and simple way, valid for both the second- and first-order versions. For that purpose, we define the average energy of the particle after one collision as $E = \beta E_0$, with E_0 the energy before the collision, and P as the probability that the particle remains, after one collision, inside the acceleration region. After n collisions, we have $N = N_0 P^n$ particles with energies $E = E_0 \beta^n$. Hence the energy spectrum results in

$$N(E)dE = \text{const.} \times E^{-1+\frac{\ln P}{\ln \beta}}\, dE\,. \tag{8}$$

It is clear that in this approach, which exhibits the expected power law, the parameters P and β can be translated into the ones that were found for the Fermi second-order mechanism, and are also going to be applied to the first-order one.

The goal of the first-order acceleration mechanism is to obtain an energy gain that is linear in (v/c), a condition that would make the acceleration process more effective, especially at relatively high values of v. This set-up will occur when the relativistic particles collide with strong shock waves (e.g., like those produced in supernova explosions, active galactic nuclei, etc.), which can reach supersonic velocities (10^3 times the velocity of an interstellar cloud).

Owing to the turbulence behind the shock and the irregularities in front of it, the particle velocity distribution is isotropic in the frames of reference where the interstellar gas is at rest on either side of the shock. Consequently, there is a complete symmetry when a high-energy particle crosses the shock from downstream to upstream or from upstream to downstream; this is illustrated in Figure 5(b).

In both types of crossing, the particle gains energy. It is possible to show [16] that in a round trip the average energy gain is given by

$$\left\langle \frac{\Delta E}{E} \right\rangle = \frac{4}{3}\left(\frac{v}{c}\right)\,. \tag{9}$$

Another quantity that must be considered is the particle escape probability P_{esc} (equivalent to $1 - P$) from the shock. Using kinetic theory, one obtains

$$P_{\text{esc}} = \frac{4}{3}\left(\frac{v}{c}\right). \tag{10}$$

Replacing these two parameters in Equation (8), we get

$$N\,(E)\,dE = \text{const.} \times E^{-2}\,dE\,. \tag{11}$$

In spite of not having obtained the observed exponent of 2.7 yet, the first-order mechanism is very promising, being the most effective and probable one, since shock waves are expected to be present in different astrophysical enviroments. In addition, in contrast to the second-order mechanism, here we find a fixed numerical value for the exponent.

5 Summary

We have presented a brief review of the mechanisms that could accelerate particles up to high energies (10^{20} eV) at galactic and extragalactic astrophysical sites. These mechanisms must fulfil a series of general requirementes, which include geometrical and energetical constraints. Among these, the Hillas criterion, a geometrical constraint on the size of the acceleration region, is most useful in selecting potential sources of cosmic rays. We have also presented two general forms of acceleration: one-shot acceleration, which requires ordered magnetic fields, and diffusive acceleration, in which particles gain energy by bouncing off random magnetic clouds. The latter type of acceleration includes Fermi shock acceleration, which correctly predicts a power-law cosmic-ray energy spectrum, albeit with a different exponent than the one that has been measured. Of the two versions of the Fermi mechanism, the first-order seems to be the most promising one to explain the ultra-high-energy cosmic-ray flux, even though it does not manage to predict the observed spectral index.

Acknowledgements

The authors would like to thank the organizers of the 5th CERN Latin American School of High-Energy Physics.

References

[1] R. Diehl, [arXiv:astro-ph/0902.4795].

[2] M. Ostrowski, Astropart. Phys. **18** (2002) 229 [arXiv:astro-ph/0101053].

[3] N. Busca, D. Hooper, and E. W. Kolb, Phys. Rev. D **73** (2006) 123001 [arXiv:astro-ph/0603055].

[4] G. I. Rubtsov *et al.*, Phys. Rev. D **73** (2006) 063009 [arXiv:astro-ph/0601449].

[5] K. Shinozaki [AGASA Collaboration], Nucl. Phys. Proc. Suppl. **151** (2006) 3.

[6] D. R. Bergman [HiRes Collaboration], [arXiv:astro-ph/0807.2814].

[7] J. Abraham *et al.* [Pierre Auger Collaboration], Astropart. Phys. **29** (2008) 188 [Erratum ibid. **30** (2008) 45] [arXiv:astro-ph/0712.2843].

[8] C. Amsler *et al.* [Particle Data Group], Phys. Lett. B **667** (2008) 1.

[9] J. W. Cronin, Rev. Mod. Phys. **71** (1999) S165.

[10] M. Nagano and A. A. Watson, Rev. Mod. Phys. **72** (2000) 689.

[11] K. Ptitsyna and S. Troitsky, [arXiv:astro-ph/0808.0367].

[12] M. Kachelriess, [arXiv:astro-ph/0801.4376].

[13] S. Gabici, [arXiv:astro-ph/0811.0836].

[14] L. Anchordoqui, H. Goldberg, S. Reucroft, and J. Swain, Phys. Rev. D **64** (2001) 123004 [arXiv:hep-ph/0107287].

[15] E. Fermi, Phys. Rev. **75** (1949) 1169.

[16] M. S. Longair, High Energy Astrophysics, Vol. 2: Stars, the Galaxy and the Interstellar Medium (Cambridge University Press, 2008).

The inert doublet model*

C. Arias[1], *J. Martins*[2], *H. Martinez*[3], *E. Ron*[4], *C. Salzmann*[5], *G. M. S. Vasconcelos*[6], *F. Villalba*[7]
1. Instituto de Física, Universidad de Antioquia, A.A. 1226, Medellín, Colombia
2. Instituto de Física Armando Dias Tavares, Universidade do Estado do Rio de Janeiro, Brazil
3. Departamento de Física, Universidad Técnica Federico Santa Maria, Valparaiso, Chile
4. Universidad Autónoma de Madrid, Madrid, Spain
5. Physik Institut der Universität Zürich, Zurich, Switzerland
6. Instituto de Física "Gleb Wataghin", Universidade Estadual de Campinas, Campinas/SP, Brazil
7. Universidad Nacional de Colombia, sede Bogotá, Colombia
Student Discussion Group

Abstract
Higgs mass divergences require Standard Model extensions such as additional physics or fields. The divergences would be less unnatural for large Higgs masses. However, the electroweak precision tests (EWPT) indicate that the Standard Model Higgs is light ($m_h < 186$ GeV). Nevertheless it is possible to increase the Higgs mass consistent with the EWPT. Here we review how this could be achieved introducing an extra Higgs doublet that has no couplings to leptons and quarks nor a vacuum expectation value. New scalar inert particles are obtained which are good dark matter candidates.

A copy of the slides presented during the oral report at the school can be found at the URL below
`http://cern.ch/PhysicSchool/LatAmSchool/2009/Presentations/pDG2.pdf`

1 Introduction

Experimental data indicates that the mass of the Standard Model (SM) Higgs boson is light, for instance the electroweak precision tests (EWPT) indicate that $m_h < 186$ GeV [1]. If we allow the Higgs to be heavy (~ 500 GeV) the introduction of new physics beyond the Standard Model (SM) becomes necessary in order to fit the EWPT experimental data. In this document we review the Inert Doublet Model (IDM) [2] as an example for required new physics. In this model an inert doublet scalar is introduced without vacuum expectation value (vev), nor couplings with the matter.

In Section 2 we summarize some of the consequences allowing a heavy Higgs in the SM (naturalness, perturbativity and agreement with the EWPT); in Section 3 we introduce the potential of the IDM and summarize the constraints on the parameters of the model mostly given by experimental data of the EWPT. In section 4 we summarize the possible collider signals for this model, and finally in Section 5 some consequences considering the lightest inert Higgs as a possible dark matter candidate are shown.

2 A heavy Higgs

If we allow for larger Higgs masses in the SM it is natural to wonder how much heavier this mass should be in order to preserve the physics where the SM remains unchanged. The authors of the IDM asked for a natural theory up to energies of 1.5 TeV. With this bound the maximum scale at which perturbation theory is useful must satisfy $\Lambda_p > 1.5$ TeV, where Λ_p is defined as the scale where the one-loop correction to the SM Higgs coupling reaches 30% of the tree-level value. This give us an upper limit on the Higgs mass. The values for Λ_p and the Landau pole scale Λ_L (where the self-coupling blows up) are shown in Table 1 for different heavy Higgs masses. These values are calculated considering the renormalization group flow of the heavy Higgs self-coupling (See Appendix A of Ref. [2] for details).

*Work performed as a student project under the supervision of D. Restrepo.

Table 1: Heavy Higgs perturbativity scale Λ_P and Landau pole Λ_L. Taken from Ref. [2]

m_h [GeV]	Λ_P [TeV]	Λ_L [TeV]
400	2.4	80
500	1.8	16
600	1.6	7.5

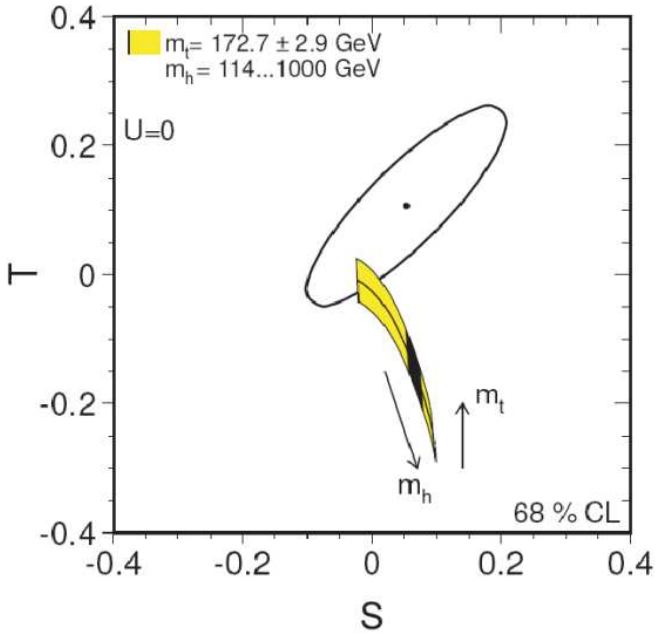


Fig. 1: The Higgs mass value as a function of the parameters S and T. The black region is for a mass between 400 GeV and 600 GeV

The EWPT favours a small Higgs mass m_h. This is only valid in the absence of new physics in the SM. We shall focus our discussion on the S and T parameters given in terms of m_h and the Z boson mass m_Z by

$$T \approx -\frac{3}{8\pi \cos^2\theta_W} \ln \frac{m_h}{m_Z} \tag{1}$$

$$S \approx \frac{1}{6\pi} \ln \frac{m_h}{m_Z} \,. \tag{2}$$

Experimental constraints on these parameters impose an upper bound for m_h ($m_h < 186$ GeV) [1]. Thus new physics allowing for a heavy Higgs will add a ΔS and a ΔT to these parameters. For the IDM the ΔS contributions can be neglected. The new physics contribution to T (in our case the IDM) must contribute with a positive ΔT for the range of Higgs masses given in Table 1. Such contribution must be

$$\Delta T \approx 0.25 \pm 0.1 \,, \tag{3}$$

in order to get the Higgs mass $m_h = 400$—600 GeV inside the 68% CL ellipse (Fig. 1).

3 The inert doublet model

We now present the IDM and summarize the constraints on its parameters.

The general model is invariant under $H_2 \to -H_2$ and is given by

$$V = \mu_1^2|H_1|^2 + \mu_2^2|H_2|^2 + \lambda_1|H_1|^4 + \lambda_2|H_2|^4 + \lambda_3|H_1|^2|H_2|^2 + \lambda_4|H_1^\dagger H_2|^2 + \frac{\lambda_5}{2}[(H_1^\dagger H_2)^2 + \text{h.c.}]. \quad (4)$$

The parity behaviour of H_2 implies that it does not couple to matter.

The physical fields are shown explicitly in the following parametrization:

$$H_1 = \begin{pmatrix} \phi^+ \\ v + (h + i\chi)/\sqrt{2} \end{pmatrix}, \qquad H_2 = \begin{pmatrix} H^+ \\ (S + iA)/\sqrt{2} \end{pmatrix}. \quad (5)$$

In the IDM only H_1 couples to matter and acquires a vev, whereas H_2 does not. This gives the Standard Model Goldstone particles ϕ^+ and χ plus three inert scalars, one charged H^+, and two neutral particles S, A. Expanding this potential around the minimum $H_1 = (0, v), \quad H_2 = (0, 0)$ we get the mass spectrum:

$$m_I^2 = \mu_2^2 + \lambda_I v^2, \quad I = \{H, S, A\} \quad (6)$$

$$\lambda_H = \lambda_3$$
$$\lambda_S = \lambda_3 + \lambda_4 + \lambda_5$$
$$\lambda_A = \lambda_3 + \lambda_4 - \lambda_5 \quad (7)$$

To get a potential V bounded from below we obtain the following constraints on the IDM parameters:

$$\lambda_{1,2} > 0 \qquad \lambda_3, \lambda_{\rm L} \equiv \lambda_3 + \lambda_4 - |\lambda_5| > -2(\lambda_1\lambda_2)^{1/2}. \quad (8)$$

The coupling λ_2 only affects the self-interactions between the inert particles and it is assumed to be small:

$$\lambda_2 \lesssim 1\,. \quad (9)$$

The parameter space can be explored in terms of the four masses m_h, m_H, m_A, m_S, the Z mass (or v), and the quartic couplings λ_2 and λ_3. In the SM the EWPT implies a relation between m_h and m_Z. In this model, there is also a relation among the masses. It follows from the expression for ΔT, that in this

$$\Delta T \approx \frac{1}{24\pi^2\alpha v^2}(m_H - m_A)(m_H - m_S). \quad (10)$$

Requiring that $\Delta T \approx 0.25 \pm 0.1$ the condition found is

$$(m_H - m_A)(m_H - m_S) = M^2, \qquad M = 120^{+20}_{-30}\ \text{GeV}\,. \quad (11)$$

Here m_H should be either bigger or smaller than both m_S and m_A to get a positive contribution to T. Furthermore, if the lightest inert particle is to be a dark matter candidate, it must be neutral, so H must be heavier than m_S and m_A. It can also be found that $\Delta S \lesssim 0.04$ for a wide parameter range. So this contribution may be neglected in the EWPT. The ΔT range is controlled by the constraints on the λ's, perturbativity and naturalness conditions. The resulting range is shown in Fig. 2. In this figure it is apparent that ΔT is of the order needed to raise the Higgs mass on a wide region in the parameter space.

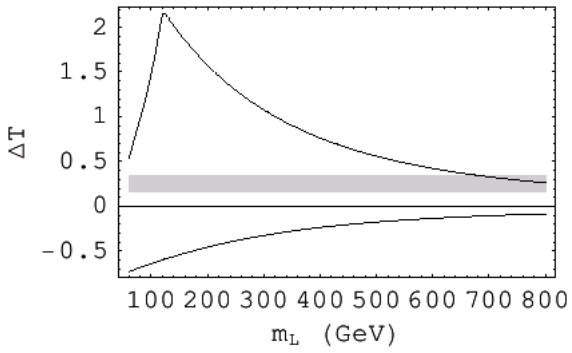


Fig. 2: Maximum and minimum values for ΔT allowed by the constraints vs. m_L. The grey band shows the ΔT needed to raise the Higgs mass.

4 Signals

The inert particles S and A could be produced in pairs. The energy necessary for the production is low enough for these particles to be produced at LEP2 as could happen if, for example, $m_L \approx 70$ GeV and Δm is small. Taking $\Delta m \ll m_L$ and $\sqrt{s} = 200$ GeV, the cross-section for the production of these pairs is:

$$\sigma(e^+e^- \to SA) = \left(\frac{g}{2c_w}\right)^4 \left(\frac{1}{2} - 2s_w^2 + 4s\_w^4\right) \frac{1}{48\pi s} \frac{(1-4m^2/s)^{3/2}}{(1-m_Z^2/s)^2} \approx 0.2 \text{ pb}. \quad (12)$$

Taking A as the heavier state, this decays into S plus Z^*.

These pair productions could be obtained at the LHC in the from

$$\begin{aligned} pp &\to W^* \to HA \quad \text{or} \quad HS \\ pp &\to Z^*(\gamma^*) \to SA \quad \text{or} \quad H^+H^-\,, \end{aligned} \quad (13)$$

followed by

$$H \to AW \quad \text{or} \quad SW \quad (14)$$

$$A \to SZ^*. \quad (15)$$

These new fields could also be detected by measuring the width of the SM Higgs, which is now increased because of the existence of new channels:

$$h \to SS, AA, H^+H^-. \quad (16)$$

5 Dark matter candidate

The IDM provides a suitable candidate for dark matter (DM). The lightest inert particle (LIP) is stable given that the parity symmetry $H_2 \to -H_2$ is respected. In the following we shall discuss some of the constraints available from data on DM imposed on the parameters of the IDM in order to reproduce the DM density and to detect it directly.

5.1 Relic abundance

It is useful to distinguish two cases with respect to the mass of the LIP in order to take into account the dominant annihilation rates for it.

1. $m_L \geq m_W$. This case is interesting since it includes almost the whole of the range allowed by the naturalness constraints. The dominant annihilation mode in this regime is into gauge bosons. It is convenient to consider separately the annihilation into transverse and longitudinal modes, in the former the cross-section can be approximated for big m_L as

$$(\sigma_{LL\to\perp\perp})v_{rel} \approx 130\,\text{pb}\left(\frac{100\text{ GeV}}{m_L}\right)^2, \tag{17}$$

while for $m_L \sim m_W$ it serves as an order-of-magnitude estimate. With this, we can say that for $m_L \sim m_W$ the cross-section is $(\sigma_{LL\to\perp\perp})v_{rel} \approx 400\,\text{pb}$. The longitudinal contribution can be calculated in terms of annihilation into massless Goldstone bosons [2] and a lower bound for the total cross-section of 10 pb is obtained. Using these values as input for the calculation of DM abundances $\Omega_{DM}h^2 \leq 0.02$ in the whole range of m_L, decreasing to 0.002 for $m_L \sim m_W$. This density is much lower than the observed one ($\Omega_{DM}h^2 \approx 0.1$). So we can conclude that in this case the LIP can only provide a subdominant component to the DM.

2. $m_L < m_W$. Taking into account the naturalness and EWPT constraints, we can restrict ourselves to the interval $m_L = (60\text{–}80)$ GeV. In spite of the fact that some additional cancellations in Eq. (6) are needed to enter in this regime, they are not so restrictive. Below the vector boson production threshold the dominant process is the coannihilation $SA \to Z^* \to \bar{f}f$, the cross-section is

$$\sigma v_{rel} = b v_{rel}^2 \tag{18}$$

$$b = \left(\frac{g}{2c_w}\right)^4 \frac{\sum_{\text{fermions}}(g_V^2+g_A^2)}{96\pi m_L^2[1-m_Z^2/(4m_L^2)]^2}, \tag{19}$$

where we supposed that $\Delta m \ll m_L$. In the range of interest $b \approx (250\text{-}60)$ pb for $m_L = (60\text{-}80)$ GeV. Supposing that $\Delta m < T$ the thermally averaged cross-section is $\langle\sigma v_{rel}\rangle = 6bT/m_L$. If we take a temperature $T_f = m_L/25$ and Δm much smaller than it, we get an averaged cross-section $\langle\sigma v_{rel}\rangle \sim (60\text{-}15)$ pb for $m_L = (60\text{-}80)$ GeV; using this value, the relic abundance turns out to be $\Omega_{DM}h^2 \approx (0.5\text{-}2.5)\times 10^{-2}$ which is below the observed value. The remaining possibility to obtain the correct density of dark matter is to take $\Delta m > T_f$, with this the density of the heavier particle is thermally suppressed and the coannihilation rate decreases, given the lower number of partners available. Naively we can expect in this case that the abundance is increased by a factor $\sim (1/2)\exp(\Delta m/T_f)$ with respect to the unsplit case; in this way we can increase the density to fit the observations and find that $\Delta m \approx 8$ GeV. More elaborate calculations [2] confirm these estimations and can be used to find the exact behaviour of Δm.

6 Conclusions

We have illustrated one specific extension of the Standard Model where it is possible to have a heavy Higgs, still compatible with the electroweak precision tests. The extension with an inert Higgs doublet may also have a proper dark matter candidate.

References

[1] The LEP Collaborations: ALEPH, DELPHI, L3, OPAL, and the LEP Electroweak Working Group, A combination of preliminary electroweak measurements and constraints on the Standard Model, arXiv:hep-ex/0511027.

[2] R. Barbieri, L. J. Hall, V. S. Rychkov, Phys. Rev. D **74** (2006) 015007, hep-ph/0603188v2.

Searching for new physics in two-body decays: Ideas and pitfalls

E. Arrieta Diaz[1], *F. Benitez*[2], *A. Büchler*[3], *L.J. Cieri*[4], *A. Florez*[5], *E. Garces-Garcia*[6], *B. Gonçalves*[7], *F. Koetsveld*[8], *K.J.C. Leney*[9], *H. Marquez Falcon*[10], *M. Moncada*[11], *P. Quintero*[11], *D. Romero*[12], *K. Shaw*[13], *J. Swain*[14], *M.P. Zurita*[4]

[1] Michigan State University, East Lansing, Michigan, United States of America
[2] Universidad de la República, Montevideo, Uruguay
[3] Universität Zürich, Zurich, Switzerland
[4] Universidad de Buenos Aires, Buenos Aires, Argentina
[5] Vanderbilt University, Nashville, Tenessee, United States of America
[6] Dpto. de Física, CINVESTAV, México DF, Mexico
[7] Universidade Federal de Juiz de Fora, Juiz de Fora, Brazil
[8] Nikhef, Radboud Universiteit, Nijmegen, Netherlands
[9] University of Liverpool, Liverpool, United Kingdom
[10] Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Mexico
[11] Universidad Nacional de Colombia, Bogotá, Colombia
[12] Pontificia Universidad Católica de Chile, Santiago, Chile
[13] University of Sheffield, Sheffield, United Kingdom
[14] Northeastern University, Boston, Massachussets, United States of America

Abstract

Many new physics processes, and indeed many Standard Model interactions involve two-body decays. Although the kinematics are relatively simple, mistakes can easily be made when applying cuts to data in order to separate the signal from backgrounds. We present a short, but relevant list of possible sources of errors, and discuss the consequences of these.

A copy of the slides presented during the oral report at the school can be found at the URL below
`http://cern.ch/PhysicSchool/LatAmSchool/2009/Presentations/pDG3.pdf`

1 Introduction

There are many interesting two-body decay processes, including those by which the existence of the Higgs boson could be confirmed or denied, or where an indication of new physics processes (beyond the Standard Model) are expected to appear. There are, however, several problems associated with the analysis of this type of process, which are rarely documented. These mostly stem from the fact that once cuts start to be made on kinematic variables (for example transverse momentum or pseudo-rapidity of the decay particles), one may be over-constraining the kinematics, thus biasing the experimental data.

The kinematics of two-body decay processes are covered in Section 2, followed by two examples of possible actual processes — $B_S \to \mu^+\mu^-$ and $H \to ZZ^* \to 4$ leptons — both of which will be well within the reach of the LHC, which is due to start taking data in November 2009. The former is the experimentally simpler of the two analyses, since the final state simply consists of two muons, and the mass of the B_S is well known from data. The Higgs analysis is complicated by the fact that not only is the mass of the Higgs unknown, but also that the two Z bosons themselves subsequently decay, leaving four particles in the final state.

Finally, a summary of some of the general problems and common mistakes associated with two-body decay analyses is made in Section 6, together with examples of common cuts which can adversely affect the experimental results.

2 General kinematics

The general kinematics of two-body decay processes of the type $A \to B + C$ are best described in the centre-of-mass frame, where the decaying particle (A) is at rest. Conservation of 4-momentum implies that B and C are emitted back-to-back, with their 3-momenta being equal and opposite. Furthermore, Lorentz invariance implies no preferred direction for the final 3-momenta, which is reflected in the absence of angular dependence in the kinematics. The initial 4-momentum can then be written in the form

$$p_A = (m_A, \overrightarrow{0}).$$

The quantity p_A is of course conserved, being equal to the sum of the final momenta p_B and p_C, where

$$p_B = (E_B, \overrightarrow{p_B}) \qquad\qquad p_C = (E_C, \overrightarrow{p_C}) \tag{1}$$

with $\overrightarrow{p_B} = -\overrightarrow{p_C} = \overrightarrow{p}$. It can then easily be shown that the energies and absolute values of the final 3-momenta of B and C can be expressed in terms of only the invariant masses of the particles.

$$\begin{aligned} E_B &= \frac{m_A^2 + m_B^2 - m_C^2}{2m_A} \\ E_C &= \frac{m_A^2 - m_B^2 + m_C^2}{2m_A} \end{aligned} \tag{2}$$

and

$$|\overrightarrow{p}| = \sqrt{E_B^2 - m_B^2} = \frac{m_A^2 - (m_B^2 + m_C^2)}{2m_A}. \tag{3}$$

3 Phase space

The phase space for two-body decays is severely constrained, which makes these types of decay conceptually easy to treat. Here, we analyse the basic kinematics in the general case. The differential decay rate of an unstable particle to a given final state in the centre-of-mass frame is [1]

$$d\Gamma = \frac{1}{2m_A}\left(\Pi_f \frac{d^3p_f}{(2\pi)^3}\frac{1}{2E_f}\right)|\mathcal{M}_{fA}|^2\,(2\pi)^4\delta^{(4)}(p_A - \sum p_f), \tag{4}$$

where the matrix element $\mathcal{M}_{fA}$ is the Feynman amplitude related to the quantum probability of the process, $2m_A$ is the incoming flux, E_f is the energy of the final-state particle (e.g., E_B, E_C), p_A is the 4-momentum of the decaying particle, p_f is the 4-momentum of the final-state particle (p_B, p_C), and the δ function accounts for 4-momentum conservation.

For the special case of a two-particle final state, the integration over the phase space takes the simpler form

$$\int \left(\Pi_f \frac{d^3p_f}{(2\pi)^3}\frac{1}{2E_f}\right)(2\pi)^4\delta^{(4)}(p_A - \sum p_f) = \int \frac{d\Omega_{CM}}{4\pi}\frac{1}{8\pi}\Big(\frac{2|\overrightarrow{p}|}{E_{CM}}\Big), \tag{5}$$

where $|\overrightarrow{p}|$ is the magnitude of the 3-momentum of either final particle. Finally, in the special case where particles B and C have the same mass, it can easily be shown that

$$\Gamma = \frac{1}{2m_A}\frac{1}{8\pi}\frac{2|\overrightarrow{p}|}{E_{CM}}\,|\mathcal{M}_{fA}|^2\ . \tag{6}$$

This expression shows that the phase space is severely constrained in the case of two-body decays. As will be shown below, this has to be taken into account at the time of performing background cuts to possible measurements.

The $|\mathcal{M}_{fA}|^2$ factor in all these expressions has to be supplemented by the actual physical process taking place, and can be computed using the relevant Feynman rules, as will be shown in the following two important examples.

4 $B_S \to \mu^+\mu^-$

b-quarks are bound by strong dynamics into colour-neutral hadrons, and the non-perturbative nature of these states makes the extraction of precision information about physics at high energies problematic. To explore possible new physics effects it is necessary to untangle them from non-perturbative QCD effects.

This is, as yet, an unsolved problem, and no unique solution exists. Instead, there are a variety of theoretical approaches and techniques, generally adapted to specific problems. While approaches based directly on QCD are clearly to be preferred, model-dependent methods are often the only option available and thus also play an important role. Effective field theories, such as the heavy-quark expansion or chiral perturbation theory are commonly used too.

These theories are based on the idea that in a given process only certain degrees of freedom may be relevant to understand the physics involved. This is often the case when kinematical considerations restrict the momenta of external particles, effectively constraining the momenta of virtual particles as well.

One can argue that in these cases it makes sense to remove from the theory all intermediate states of high virtuality. Their absence might be compensated for by introducing new (effective) interactions between the remaining degrees of freedom. Using this approach one can recover, for example, the Fermi theory of weak interactions at low energies; starting from the Standard Model Lagrangian and integrating out the massive gauge vectors.

What makes an effective field theory powerful is that the deviation from the limiting behaviour may be organized in a systematic expansion in a small parameter, usually related to the scale up to which the theory makes sense. An effective field theory is then predictive, precisely because it is under perturbative control.

Many quantities of experimental and phenomenological importance cannot be analysed by these methods, however, even if these are systematic and well understood. For the description of exclusive hadronic weak decays, most exclusive semi-leptonic decays, strong decays, fragmentation, and many other interesting aspects of B-physics, only a few model-dependent approaches are available.

4.1 Theoretical framework

The decays $B^0_{s,d} \to l^+l^-$ are dominated by the Z^0 penguin (also called vertical or annihilation penguin) and box diagrams involving top quark exchanges, as shown in Fig. 1.

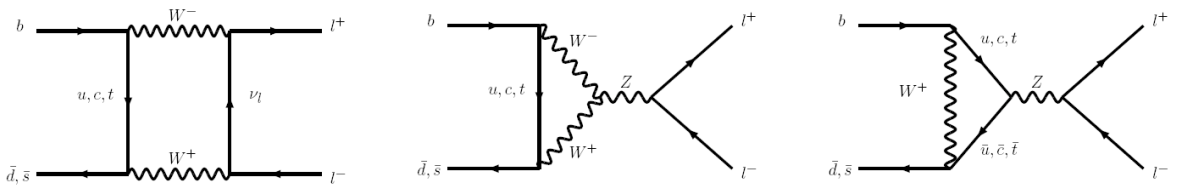


Fig. 1: Decay processes contributing to $B^0_{s,d} \to l^+l^-$ in the Standard Model

The effective Hamiltonian for $B^0_{s,d} \to l^+l^-$ decays is given by

$$H_{eff} = -\frac{4G_f}{\sqrt{2}} V_{tb}^* V_{tq} \left[C_{10} Q_{10} + C_S Q_S + C_P Q_P \right] \ , \tag{7}$$

where q indicates a strange quark for the B_s^0 or a down quark in the case of the B_d^0.

$$Q_S = \frac{e^2}{16\pi^2} (\bar{q}_{L\alpha} b_{R\alpha})(\bar{l}l)$$
$$Q_P = \frac{e^2}{16\pi^2} (\bar{q}_{L\alpha} b_{R\alpha})(\bar{l}\gamma_5 l)$$

C_S and C_P are the Wilson coefficients for the Standard Model Higgs penguin, and the would-be neutral Goldstone boson penguin, respectively. However, these contributions to the amplitude are suppressed by a factor of m_b^2/M_W^2 relative to the main contribution and can be ignored (although it should be noted that C_S and C_P can become non-negligible for some extensions of the Standard Model). Thus, the Standard Model decay amplitude is given by the Wilson coefficient

$$C_{10} = -Y(x_t)/\sin^2\theta_W = -4.2 \ , \tag{8}$$

where

$$\begin{aligned} Y(x_t) &= \eta_\gamma \cdot Y_0(x_t) \\ Y_0(x_t) &= \frac{x}{8}\Big[\frac{x_t - 4}{x_t - 1} + \frac{3x_t}{(x_t-1)^2}\log x_t\Big] \\ x_t &= \frac{m_t^2}{M_W^2}. \end{aligned} \tag{9}$$

Here η_Y summarizes the next-to-leading-order correction with $\eta_Y = 1.012$. Evaluating the hadronic matrix element, the resulting branching ratio for $B_{q=s,d}$ is

$$\begin{aligned} \mathbf{B}(B_q \to l^+ l^-) = &\frac{G_F^2 \alpha^2 m_{B_q}^2 \tau_{B_q} f_{B_q}^2}{64\pi^3} \mid V_{tb}^* V_{tq} \mid^2 \sqrt{1 - \frac{4m_l^2}{m_{B_q}}} \\ &\times \left[\left(1 - \frac{4m_l^2}{m_{B_q}}\right) \mid \frac{m_{B_q}}{m_b + m_q} C_S \mid^2 + \mid \frac{2m_l}{m_{B_q}} C_{10} - \frac{m_{B_q}}{m_b + m_q} C_P \mid^2 \right], \end{aligned} \tag{10}$$

where τ_{B_q} signifies the B_q lifetime, and f_{B_q} is the B_q decay constant normalized according to $f_\pi = 132$ MeV. The Standard Model predictions are $BR(B_d^0 \to \mu^+\mu^-) = 1.02 \pm 0.09 \times 10^{-10}$, $BR(B_s^0 \to \mu^+\mu^-) = 3.37 \pm 0.31 \times 10^{-9}$ [2]. The 95% confidence level experimental limits by CDF are $B_d^0 \to \mu^+\mu^- < 3.0 \times 10^{-8}$ and $B_s^0 \to \mu^+\mu^- < 1.0 \times 10^{-7}$ [3].

4.2 Background

There are three main backgrounds to B_s production at the LHC [4]. Misidentified B-mesons provide the largest contribution, followed by combinatorics from di-muon events. The $B_c \to J/\Psi(\mu\mu)\mu\nu_\mu$ process (which passes the invariant mass cut because the B_c is slightly heavier than the B_s) is also significant. Provided the mass resolution of the detector is good enough, decays from other B-mesons can be safely ignored as background [4].

5 $H \to ZZ^* \to 4$ leptons

The search for the Higgs boson will be one of the primary tasks of the LHC and it has been established by many studies [5] that a Standard Model Higgs boson can be discovered with high significance at the LHC, over the full range of mass interest, from the lower limit of 114 GeV up to about 1 TeV.

The predominant Higgs production mechanism at the LHC will be gluon–gluon fusion, accounting for approximately 80% of all events (dependent on the Higgs mass). The second largest contribution comes from the fusion of vector bosons radiated from the initial-state quarks [5]. Production cross-sections as a function of Higgs mass are shown in Fig. 2 [6].

The $H \to \gamma\gamma$ channel looks to be a promising channel for Higgs masses less than 140 GeV, while for heavier Higgses the most promising searches involve decays to pairs of vector bosons (W^+W^-, ZZ). The only direct fermion decays with significant branching ratios are to $b\bar{b}$ and to two tau leptons. These are particularly important channels for a measurement of the Higgs boson coupling to fermions.

For $M_H > 125$ GeV, the four-lepton decay from $H \to ZZ^*$ provides a very clean signature over a wide mass range (up to 600 GeV) thanks to a combination of a narrow reconstructed mass peak and relatively low backgrounds. This is particularly true when $M_H > 180$ GeV, where the cross-section for two on-shell Z bosons opens up.

Furthermore, the $H \to ZZ^* \to 4l$ channel is also interesting because it allows for measurements of the spin of the Higgs to be made, through observations of the angle between pairs of leptons.

The branching ratios for these main decays as a function of Higgs mass are shown in Fig. 3 [6].

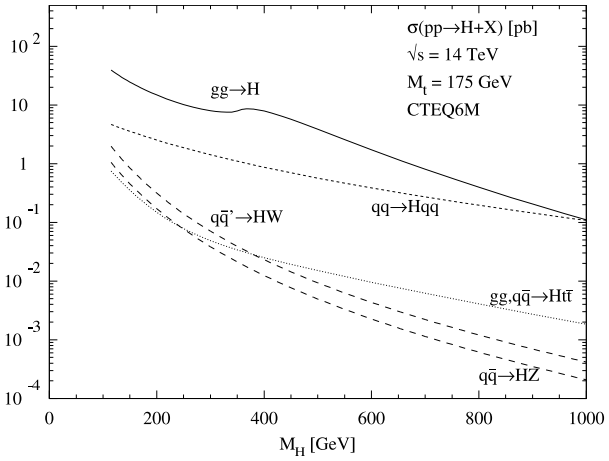


Fig. 2: Higgs production cross-section as a function of M_H

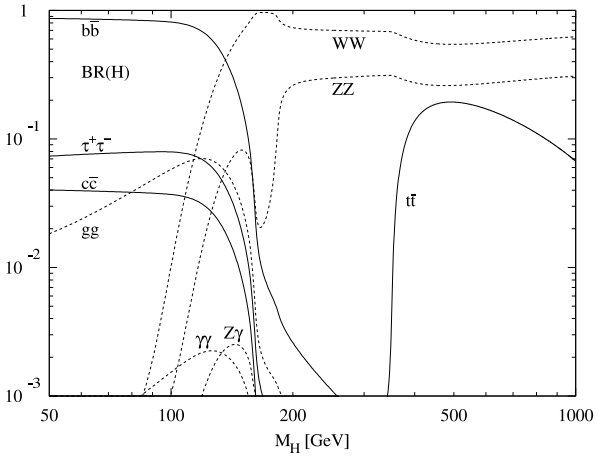


Fig. 3: Higgs decay branching ratios for various channels as functions of M_H

5.1 Signal signature

Although the Z bosons from the Higgs can decay to e^+e^-, $\mu^+\mu^-$, $\tau^+\tau^-$, $q\bar{q}$ or $\nu_{e/\mu/\tau}\bar{\nu}_{e/\mu/\tau}$, the prefered final state generally includes electrons and/or muons, since these provide a much cleaner signature. A Feynman representation of the $H \to ZZ \to 4l$ process is shown in Fig. 4. In principle, each flavour contributes to the loop, but as the Higgs couplings to fermions are proportional to the fermion masses, the top quark is responsible for the dominant contribution.

5.2 Theoretical framework

To compute the decay rate for the process it is necessary to use the corresponding Feynman rules, deduced from the Standard Model Lagrangian. Here, this corresponds to the right-hand side ($H \to ZZ$) vertex in Fig. 4. The complex amplitude is given by

$$i\mathcal{M} = 2i\frac{m_Z^2}{v} g^{\mu\nu} \varepsilon_\mu(k_1)\varepsilon_\nu(k_2) \,, \tag{11}$$

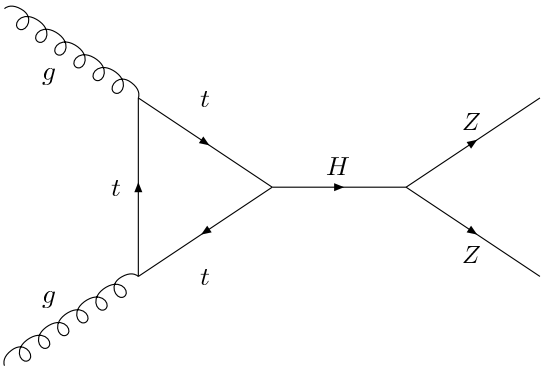


Fig. 4: A Feynman representation of Standard Model Higgs production via gluon fusion, and subsequent decay to two Z bosons

which implies

$$|\mathcal{M}|^2 = 4\frac{m_Z^4}{v^2}\varepsilon_\mu(k_1)\varepsilon_\nu^*(k_1)\varepsilon^\mu(k_2)\varepsilon^{\nu*}(k_2)\ , \tag{12}$$

where $\varepsilon_\mu(k_1)$ is the polarization vector of the outgoing particle with 4-momentum k_1, and v is the vacuum expectation value of the Higgs field.

When one is not interested in measuring the polarization of the outgoing Z's, it is useful to exploit the relationship (valid for massive gauge vectors)

$$\sum_{pol}\varepsilon_\alpha^*(k)\varepsilon_\beta(k) = -g_{\alpha\beta} + \frac{k_\alpha k_\beta}{m^2}$$

where m is the mass of the vector boson. After some algebra we obtain

$$\sum_{pol}|\mathcal{M}|^2 = 4\frac{m_Z^4}{v^2}\left[2 + \frac{(k_1\cdot k_2)^2}{m_Z^4}\right]. \tag{13}$$

Using the general considerations of Section 2 and working in the centre-of-mass frame one can show that

$$\sum_{pol}|\mathcal{M}|^2 = 4\frac{m_Z^4}{v^2}\left[3 - \frac{m_H^2}{m_Z^2} + \frac{m_H^4}{4m_Z^4}\right]. \tag{14}$$

Finally, taking into account that the Higgs is decaying into two identical particles, and setting $v^2 = m_H^2/2\lambda^2$), we arrive at

$$\Gamma = \frac{\lambda}{2\pi}\frac{m_Z^4}{m_H^2}\sqrt{m_H^2 - 4m_Z^2}\left[3 - \frac{m_H^2}{m_Z^2} + \frac{m_H^4}{4m_Z^4}\right]. \tag{15}$$

From this, it can be seen that $m_H \geq 2m_Z \sim 180$ GeV for the decay to occur. In these calculations, it was assumed that both Z bosons were on-shell—a justified simplification considering that the off-shell contribution for the process is heavily suppressed by the propagators for the virtual particles. Indeed, if we return to Fig. 3 we see a sharp increase in the $H \to ZZ$ branching ratio at around 180 GeV.

5.3 Background

The dominant background for the $H \to ZZ \to 4l$ process comes from the irreducible $ZZ \to 4l$ continuum over the full mass range. For smaller Higgs masses, where one of the Z bosons is off-shell, the leptons have a lower p_T [7]. In this region, backgrounds from $Zb\bar{b} \to 4l$ and $t\bar{t} \to W^+W^-b\bar{b} \to 4l$ are also significant, but reducible.

Both these backgrounds contain $b\bar{b}$ pairs which can decay to leptons, thus faking the signal. However, leptons from the signal should be isolated, whereas those from b-daughters are often accompanied by hadronic jets. Placing isolation requirements on the electrons and muons should help to reduce the number of b-daughters which are reconstructed as coming from the Z decay.

A veto on events with a significant amount of missing transverse energy can help to reduce the contribution from leptonic W decays (from top quark decays) since these are always accompanied by a neutrino.

6 Analysis cuts and potential pitfalls

To claim a discovery of rare decays like $B_s \to \mu^+\mu^-$ and $H \to ZZ$, a statistically significant peak in the mass distribution above the expected background must be identified. The reduction of background contributions over the full range of the mass interest is therefore crucial. Where the initial mass is known, the kinematic parameters of the decay are fully constrained and analysis cuts should not be based on the kinematic variables since this can further constrain the mass peak without necessarily improving the signal-to-background ratio. Instead, one should aim to base initial selection cuts on non-kinematic variables. For $B_s \to \mu^+\mu^-$ such cuts may be based upon

i) B_s impact parameter b or impact parameter significance (see Fig. 5).
ii) Angle between B_s momentum and the direction of primary vertex (PV) to secondary vertex (SV).
iii) To reduce combinatorial background, muons should come from the same SV, so that the mismatch x between the expected decay length of the B_s and the SV should be small (e.g., cut on secondary vertex χ^2).
iv) The angular distribution of the muons in the rest frame of the B_s should be isotropic. If Θ is the angle between the PV and SV direction and one-muon momentum this implies that the $\cos(\Theta)$ distribution is flat.

All cuts (direct or indirect) on the muon momentum and energy should be avoided as these will bias the mass distribution. A cut on momentum will remove background that falls outside of the mass peak, but not within. The ratio between the tails and the amplitude of the mass distribution would therefore appear to be improved, but any background that happens to be kinematically similar to the signal is not removed.

Cuts on the opening angle of the muons in the rest frame of the B_s will affect the signal in the same way. This will remove background that is not decaying with an opening angle of 180 degrees, but one must bear in mind that all two-body decays will behave in the same manner. Again, we observe that backgrounds kinematically similar to the signal (e.g., B_s to K^+K^-) are not removed, thus artificially enhancing the peak in the invariant mass distribution. The same reasoning can be applied to $H \to ZZ$ analysis cuts.

7 Conclusions

We have worked out the general form of two-body decays, and applied it to the study of two important processes expected to be observed at the LHC. Owing to energy–momentum conservation, the kinematical magnitudes of the final states are fully fixed, depending exclusively on the mass of the particles and the energy of the initial particle.

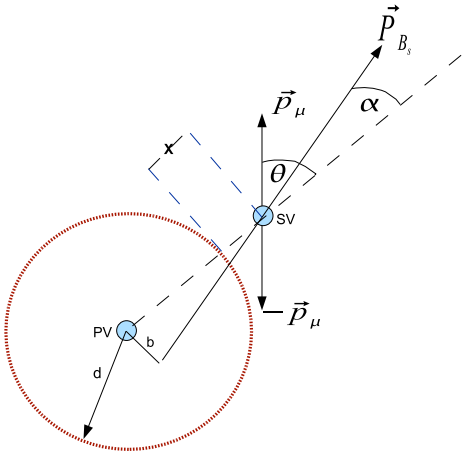


Fig. 5: Kinematics in the rest frame: where impact parameter b = distance of B_s momentum to primary vertex (PV), α = angle between B_s momentum and the direction of PV to secondary vertex (SV), d = decay length of B_s, Θ = angle between the PV and SV direction and one muon momentum, and x = mismatch between SV and d (B_s decay length)

One of the most important characteristics of these processes is their angular dependence when observed in the centre-of-mass frame. In this scenario—when the emission of decay particles occurs back-to-back—severe constraints are imposed on the potential for cleanly separating the signal from background. In particular, one should note that any cut that depends on either the energies or the 3-momenta of the final-state particles has the potential to bias results. Furthermore, even when secondary decays take place, the detected decay products must be isotropically distributed in the centre-of-mass frame.

Not taking into account these simple considerations when imposing cuts may mean that the wrong conclusions can be drawn, due to possible enhancement of background noise in the relevant region of observation.

References

[1] M.E. Peskin and D.V. Schroeder, An Introduction to Quantum Field Theory (Addison-Wesley, Reading, USA, 1995), p. 842.

[2] M. E. Albrecht, W. Altmannshofer, A. J. Buras, D. Guadagnoli, and D. M. Straub, Challenging SO(10) SUSY GUTs with family symmetries through FCNC processes, arXiv:0707.3954v3 [hep-ph].

[3] http://www-cdf.fnal.gov/physics/new/bottom/060316.blessed-bsmumu3/ and CDF public note 8176.

[4] D. Martinez, J. A. Hernando, and F. Teubert, LHCb potential to measure/exclude the branching ratio of the decay $B_s \to \mu^+\mu^-$, CERN-LHCB-2007-033 (2007).

[5] ATLAS Collaboration, ATLAS Detector and Physics Performance Technical Design Report, CERN/LHCC/99-14 (1999).

[6] ATLAS Collaboration, Expected Performance of the ATLAS Experiment, Detector, Trigger and Physics, CERN-OPEN-2008-020 (2008).

[7] A. D'Orazio on behalf of the ATLAS Collaboration, Standard Model Higgs search in the 4-lepton final state with ATLAS, Proceedings of 'Physics at the LHC 2008', CERN-ATL-PHYS-PROC-2009-019 (2009).

The accelerating Universe

K. Alwyn[1]*, A. Austregesilo*[2,3]*, R. Benavides Palacios*[4]*, J. Brochero Cifuentes*[5]*, L. Caminada*[6]*, G. Fiorentini*[7]*, P. H. Flose Reimberg*[8] *, V. I. Giraldo Rivera, F. A. Gomez Albarracin*[10]*, M. L. Gonzalez Silva*[11]*, H. J. Hortua Orjuela*[12]*, J. Imong*[13]*, C. Martinez*[14]*, D. A. Martinez Caicedo*[7]*, M. Nowakowski*[5]*, F. Quinonez Granados*[14]

[1] University of Manchester, Manchester, United Kingdom
[2] TU Muenchen, Munich, Germany
[3] CERN, Geneva, Switzerland
[4] Universidad de Antioquia, Medellíin, Colombia
[5] Universidad de los Andes, Bogotá, Columbia
[6] ETH, Zurich, Switzerland
[7] Centro Brasileiro de Pesquisas Fìsicas, Brazil
[8] Universidade de São Paulo, São Paulo, Brazil
[9] University of Liverpool, Liverpool, United Kingdom
[10] Universidad Nacional de La Plata, Argentina
[11] Universidad de Buenos Aires, Buenos Aires, Argentina
[12] Universidad Nacional de Colombia, Bogotá, Colombia
[13] University of Bristol, Bristol, United Kingdom
[14] Pontificia Universidad Católica de Chile, Santiago, Chile

No written report is available.
A copy of the slides presented during the oral report at the school can be found at the URL below
`http://cern.ch/PhysicSchool/LatAmSchool/2009/Presentations/pDG4.pdf`

International Scientific Committee

Alvaro De Rújula (CERN, Geneva, Switzerland)
Carlos García Canal (Universidad Nacional de La Plata, Argentina)
John Ellis (CERN, Geneva, Switzerland)
Nick Ellis (CERN, Geneva, Switzerland)
Robert Fleischer (CERN, Geneva, Switzerland)
Egil Lillestøl (University of Bergen and CERN, Geneva, Switzerland), CERN Schools Director
Danielle Métral (CERN, Geneva, Switzerland)
Ron Shellard (Centro Brasileiro de Pesquisas Fisicas, Rio de Janeiro, Brazil)
Arnulfo Zepeda (CINVESTAV, Mexico City, Mexico)

Local Organizing Committee

Marta Losada (University Antonio Nariño, Bogotá), Local Director
Enrico Nardi (University de Antioquia, Medellin and INFN–LNF, Frascati, Italy)
Carlos Quimbay (University Nacional de Colombia, Bogotá)
Juan Carlos Sanabria (University de los Andes, Bogotá)

Lecturers

Luis Álvarez-Gaumé (CERN, Geneva, Switzerland)
John Ellis (CERN, Geneva, Switzerland)
Nick Ellis (CERN, Geneva, Switzerland)
Pilar Hernández (University of Valencia, Spain)
Gerardo Herrera Corral (CINVESTAV, Mexico City, Mexico)
Andreas Hoecker (CERN, Geneva, Switzerland)
Jordan Nash (CERN, Geneva, Switzerland)
Yosef Nir (Weizmann Institute of Science, Rehovot, Israel)
Antonio Riotto (CERN, Geneva, Switzerland)
Dmitri Semikoz (APC, Paris, France)
Michael Seymour (CERN, Geneva, Switzerland)
Special lecture: Ricardo Callejas Posada (Instituto de Biología, Universidad de Antioquia, Medellín, Colombia)
'Botanical exploration of the state of Antioquia during the XIX century'

Discussion Leaders

Alberto Gago Medina (Pontificia Universidad Católica del Perú, Lima, Peru)
Marek Nowakowski (Departamento de Fisica, Universidad de los Andes, Bogotá, Colombia)
Diego Restrepo (Instituto de Física, Universidad de Antioquia, Medellín, Colombia)
John Swain (Northeastern University, Boston, USA)

Students

Kim ALWYN
Carolina ARBELAEZ
Cesar ARIAS
Ivan ARRAUT
Enrique ARRIETA DIAZ
Alexander AUSTREGESILO
Richard BENAVIDES PALACIOS
Federico BENITEZ
Javier BROCHERO CIFUENTES
Angela BUECHLER
Mauricio BUSTAMANTE
Lea CAMINADA
Blanca Cecilia CANHAS
German David CARRILLO MONTOYA
Leandro CIERI
Wayne DE PAULA
Mary DIAZ
Javier Alberto DUARTE CHAVEZ
Guillermo FIORENTINI
Carlos FLOREZ BUSTOS
Paulo Henrique FLOSE REIMBERG
Estela Alejandra GARCES-GARCIA
Victor Ivan GIRALDO RIVERA
Yithsbey GIRALDO
Flavia Alejandra GOMEZ ALBARRACIN
Bruno GONÇALVES
Maria Laura GONZALEZ SILVA
Hayk HAKOBYAN
Daniel Fernando HIGUITA BORJA
Hector Javier HORTUA ORJUELA
Jonathan IMONG
Alejandro JARAMILLO MORENO
Pavel JEZ
Folkert KOETSVELD
Katharine LENEY
Hugo Raymundo MARQUEZ FALCON
David Alejandro MARTINEZ CAICEDO
Cristian MARTINEZ
Hector MARTINEZ
Jordan MARTINS
Carlos MEDINA HERNANDEZ
Jhovanny Andres MEJIA GUISAO
Miguel MONCADA
Jose Andres MONROY MONTAÑEZ
Jorge Luis NISPERUZA TOLEDO
Antonio ORTIZ VELASQUEZ
Boris OSORNO TORRES
Fatima PADILLA CABAL
Miguel PINO ROZAS
Fernando QUINONEZ GRANADOS
Pedro QUINTERO
Patricia REBELLO TELES
Diana Osorno RIVERA
Diego Julian RODRIGUEZ PATARROYO
Gaston Leonardo ROMEO
Diego Alonso ROMERO MALTRANA
Elias RON
José Alejandro ROSABAL RODRIGUEZ
Jose David RUIZ ÁLVAREZ
Ulises Jesus SALDAÑA-SALAZAR
Christophe SALZMANN
Marta Liliana SÁNCHEZ PELÁEZ
Geraldo Magela SEVERINO VASCONCELOS
Kate SHAW
Javier TIFFENBERG
Mauricio VELASQUEZ
Fabián Darío VILLALBA-PARDO
Marina VON STEINKIRCH
Bruce YEE RENDON
Pía ZURITA

Posters

Author	Poster title
Kim ALWYN	Searches for second-class currents in τ decays at BaBar
Ivan ARRAUT	Generalized uncertainty principle with a cosmological constant. (Consequences on Hawking radiation).
Alexander AUSTREGESILO	The PixelGEM — a high-rate beam tracker for COMPASS
Javier BROCHERO CIFUENTES	Diffractive production of W bosons in the $\mu\nu$ channel at 1.96 TeV energy
Angela BUECHLER	Tracker Turicensis
Mauricio BUSTAMANTE	Effects of energy-independent new physics on the high-energy astrophysical neutrino flavour ratios
Lea CAMINADA	The CMS pixel barrel detector — from construction to commissioning
German David CARRILLO MONTOYA	Higgs searches in the di-boson decay channel in the ATLAS detector
Leandro CIERI	Transverse-momentum resummation in Drell–Yan processes
Wayne DE PAULA	Linear Regge trajectories from a dynamical AdS/QCD model
Mary DIAZ	Anisotropy studies in ultra high-energy cosmic rays
Paulo Henrique FLOSE REIMBERG	Polarization of the cosmic microwave background
Flavia Alejandra GOMEZ ALBARRACIN	What are cosmic rays made of? Mass composition at the Pierre Auger observatory
Bruno GONÇALVES	Exact Foldy–Wouthuysen transformation for Dirac particles interacting with magnetic field and gravitational wave backgrounds
Maria Laura GONZALEZ SILVA	ATLAS experiment: commissioning with cosmic rays
Hayk HAKOBYAN	Quark propagation and hadron formation
Pavel JEZ	τ trigger at the ATLAS experiment
Folkert KOETSVELD	Searching for supersymmetry using a data-driven background fit
Katharine LENEY	Data quality from the detector control system at the ATLAS experiment
Hugo Raymundo MARQUEZ FALCON	Anisotropy studies in the ultra high-energy cosmic rays Auger observatory
Cristian MARTINEZ	Analytic QCD and narrow width approximation
Carlos MEDINA HERNANDEZ	M5/M7 studies on the ATLAS intermediate tile calorimeter with cosmic data

Author	Poster title
Jose Andres MONROY MONTAÑEZ	Study of diffractive production of the $Z \to e^+e^-$ at 1.96 TeV energy in the D0 experiment
Antonio ORTIZ VELASQUEZ	Topological studies of high multiplicity p–p collisions with ALICE at the LHC
Fatima PADILLA CABAL	Drift velocity from the SDD detector of the ITS, ALICE experiment
Patricia REBELLO TELES	Searching for new physics at the LHC: anomalous gauge-boson couplings
Gaston Leonardo ROMEO	ATLAS experiment: commissioning with cosmic rays
Geraldo Magela SEVERINO VASCONCELOS	Study of multi (strange) particle production in relativistic heavy-ion collisions
Kate SHAW	The ATLAS experiment
Fabián Darío VILLALBA-PARDO	Neutrino dispersion relation in the left–right symmetric model at finite temperature