Multivariable Advanced Calculus

Kenneth Kuttler

September 30, 2016

Contents

Chapter 1

Introduction

This book is directed to people who have a good understanding of the concepts of one variable calculus including the notions of limit of a sequence and completeness of $\mathbb{R}$. It develops multivariable advanced calculus.

In order to do multivariable calculus correctly, you must first understand some linear algebra. Therefore, a condensed course in linear algebra is presented first, emphasizing those topics in linear algebra which are useful in analysis, not those topics which are primarily dependent on row operations.

Many topics could be presented in greater generality than I have chosen to do. I have also attempted to feature calculus, not topology although there are many interesting topics from topology. This means I introduce the topology as it is needed rather than using the possibly more efficient practice of placing it right at the beginning in more generality than will be needed. I think it might make the topological concepts more memorable by linking them in this way to other concepts.

After the chapter on the n dimensional Lebesgue integral, you can make a choice between a very general treatment of integration of differential forms based on degree theory in chapters 10 and 11 or you can follow an independent path through a proof of a general version of Green's theorem in the plane leading to a very good version of Stoke's theorem for a two dimensional surface by following Chapters 12 and 13. This approach also leads naturally to contour integrals and complex analysis. I got this idea from reading Apostol's advanced calculus book. Finally, there is an introduction to Hausdorff measures and the area formula in the last chapter.

I have avoided many advanced topics like the Radon Nikodym theorem, representation theorems, function spaces, and differentiation theory. It seems to me these are topics for a more advanced course in real analysis. I chose to feature the Lebesgue integral because I have gone through the theory of the Riemann integral for a function of n variables and ended up thinking it was too fussy and that the extra abstraction of the Lebesgue integral was worthwhile in order to avoid this fussiness. Also, it seemed to me that this book should be in some sense "more advanced" than my calculus book which does contain in an appendix all this fussy theory.

Chapter 2

Some Fundamental Concepts

2.1 Set Theory

2.1.1 Basic Definitions

A set is a collection of things called elements of the set. For example, the set of integers, the collection of signed whole numbers such as 1,2,-4, etc. This set whose existence will be assumed is denoted by $\mathbb{Z}$. Other sets could be the set of people in a family or the set of donuts in a display case at the store. Sometimes parentheses, { } specify a set by listing the things which are in the set between the parentheses. For example the set of integers between -1 and 2, including these numbers could be denoted as $\{-1, 0, 1, 2\}$. The notation signifying x is an element of a set S, is written as $x \in S$. Thus, $1 \in \{-1, 0, 1, 2, 3\}$. Here are some axioms about sets. Axioms are statements which are accepted, not proved.

1. Two sets are equal if and only if they have the same elements.
2. To every set, A, and to every condition $S(x)$ there corresponds a set, B, whose elements are exactly those elements x of A for which $S(x)$ holds.
3. For every collection of sets there exists a set that contains all the elements that belong to at least one set of the given collection.
4. The Cartesian product of a nonempty family of nonempty sets is nonempty.
5. If A is a set there exists a set, $\mathcal{P}(A)$ such that $\mathcal{P}(A)$ is the set of all subsets of A. This is called the power set.

These axioms are referred to as the axiom of extension, axiom of specification, axiom of unions, axiom of choice, and axiom of powers respectively.

It seems fairly clear you should want to believe in the axiom of extension. It is merely saying, for example, that $\{1, 2, 3\} = \{2, 3, 1\}$ since these two sets have the same elements in them. Similarly, it would seem you should be able to specify a new set from a given set using some "condition" which can be used as a test to determine whether the element in question is in the set. For example, the set of all integers which are multiples of 2. This set could be specified as follows.

$$\{x \in \mathbb{Z} : x = 2y \text{ for some } y \in \mathbb{Z}\}.$$

In this notation, the colon is read as "such that" and in this case the condition is being a multiple of 2.

Another example of political interest, could be the set of all judges who are not judicial activists. I think you can see this last is not a very precise condition since there is no way to determine to everyone's satisfaction whether a given judge is an activist. Also, **just because something is grammatically correct does not mean it makes any sense.** For example consider the following nonsense.

$$S = \{x \in \text{set of dogs} \ : \ \text{it is colder in the mountains than in the winter}\}.$$

So what is a condition?

We will leave these sorts of considerations and assume our conditions make sense. The axiom of unions states that for any collection of sets, there is a set consisting of all the elements in each of the sets in the collection. Of course this is also open to further consideration. What is a collection? Maybe it would be better to say "set of sets" or, given a set whose elements are sets there exists a set whose elements consist of exactly those things which are elements of at least one of these sets. If $\mathcal{S}$ is such a set whose elements are sets,

$$\cup\{A : A \in \mathcal{S}\} \text{ or } \cup\mathcal{S}$$

signify this union.

Something is in the Cartesian product of a set or "family" of sets if it consists of a single thing taken from each set in the family. Thus $(1,2,3) \in \{1,4,.2\} \times \{1,2,7\} \times \{4,3,7,9\}$ because it consists of exactly one element from each of the sets which are separated by $\times$. Also, this is the notation for the Cartesian product of finitely many sets. If $\mathcal{S}$ is a set whose elements are sets,

$$\prod_{A\in\mathcal{S}} A$$

signifies the Cartesian product.

The Cartesian product is the set of choice functions, a choice function being a function which selects exactly one element of each set of $\mathcal{S}$. You may think the axiom of choice, stating that the Cartesian product of a nonempty family of nonempty sets is nonempty, is innocuous but there was a time when many mathematicians were ready to throw it out because it implies things which are very hard to believe, things which never happen without the axiom of choice.

A is a subset of B, written $A \subseteq B$, if every element of A is also an element of B. This can also be written as $B \supseteq A$. A is a proper subset of B, written $A \subset B$ or $B \supset A$ if A is a subset of B but A is not equal to $B, A \neq B$. $A \cap B$ denotes the intersection of the two sets A and B and it means the set of elements of A which are also elements of B. The axiom of specification shows this is a set. The empty set is the set which has no elements in it, denoted as $\emptyset$. $A \cup B$ denotes the union of the two sets A and B and it means the set of all elements which are in either of the sets. It is a set because of the axiom of unions.

The complement of a set, (the set of things which are not in the given set) must be taken with respect to a given set called the universal set which is a set which contains the one whose complement is being taken. Thus, the complement of A, denoted as A^C (or more precisely as $X \setminus A$) is a set obtained from using the axiom of specification to write

$$A^C \equiv \{x \in X : x \notin A\}$$

The symbol $\notin$ means: "is not an element of". Note the axiom of specification takes place relative to a given set. Without this universal set it makes no sense to use the axiom of specification to obtain the complement.

Words such as "all" or "there exists" are called quantifiers and they must be understood relative to some given set. For example, the set of all integers larger than 3. Or

there exists an integer larger than 7. Such statements have to do with a given set, in this case the integers. Failure to have a reference set when quantifiers are used turns out to be illogical even though such usage may be grammatically correct. Quantifiers are used often enough that there are symbols for them. The symbol $\forall$ is read as "for all" or "for every" and the symbol $\exists$ is read as "there exists". Thus $\forall\forall\exists\exists$ could mean for every upside down A there exists a backwards E.

DeMorgan's laws are very useful in mathematics. Let $\mathcal{S}$ be a set of sets each of which is contained in some universal set, U. Then

$$\cup\left\{A^C : A \in \mathcal{S}\right\} = \left(\cap\left\{A : A \in \mathcal{S}\right\}\right)^C$$

and

$$\cap\left\{A^C : A \in \mathcal{S}\right\} = \left(\cup\left\{A : A \in \mathcal{S}\right\}\right)^C.$$

These laws follow directly from the definitions. Also following directly from the definitions are:

Let $\mathcal{S}$ be a set of sets then

$$B \cup \cup\left\{A : A \in \mathcal{S}\right\} = \cup\left\{B \cup A : A \in \mathcal{S}\right\}.$$

and: Let $\mathcal{S}$ be a set of sets show

$$B \cap \cup\left\{A : A \in \mathcal{S}\right\} = \cup\left\{B \cap A : A \in \mathcal{S}\right\}.$$

Unfortunately, there is no single universal set which can be used for all sets. Here is why: Suppose there were. Call it S. Then you could consider A the set of all elements of S which are not elements of themselves, this from the axiom of specification. If A is an element of itself, then it fails to qualify for inclusion in A. Therefore, it must not be an element of itself. However, if this is so, it qualifies for inclusion in A so it is an element of itself and so this can't be true either. Thus the most basic of conditions you could imagine, that of being an element of, is meaningless and so allowing such a set causes the whole theory to be meaningless. The solution is to not allow a universal set. As mentioned by Halmos in Naive set theory, "Nothing contains everything". Always beware of statements involving quantifiers wherever they occur, even this one. This little observation described above is due to Bertrand Russell and is called Russell's paradox.

2.1.2 The Schroder Bernstein Theorem

It is very important to be able to compare the size of sets in a rational way. The most useful theorem in this context is the Schroder Bernstein theorem which is the main result to be presented in this section. The Cartesian product is discussed above. The next definition reviews this and defines the concept of a function.

Definition 2.1.1 *Let X and Y be sets.*

$$X \times Y \equiv \{(x, y) : x \in X \text{ and } y \in Y\}$$

A relation is defined to be a subset of $X \times Y$. A function, f, also called a mapping, is a relation which has the property that if (x, y) and (x, y_1) are both elements of the f, then $y = y_1$. The domain of f is defined as

$$D(f) \equiv \{x : (x, y) \in f\},$$

written as $f : D(f) \to Y$.

It is probably safe to say that most people do not think of functions as a type of relation which is a subset of the Cartesian product of two sets. A function is like a machine which takes inputs, x and makes them into a unique output, $f(x)$. Of course, that is what the above definition says with more precision. An ordered pair, (x,y) which is an element of the function or mapping has an input, x and a unique output, y,denoted as $f(x)$ while the name of the function is f. "mapping" is often a noun meaning function. However, it also is a verb as in "f is mapping A to B". That which a function is thought of as doing is also referred to using the word "maps" as in: f maps X to Y. However, a set of functions may be called a set of maps so this word might also be used as the plural of a noun. There is no help for it. You just have to suffer with this nonsense.

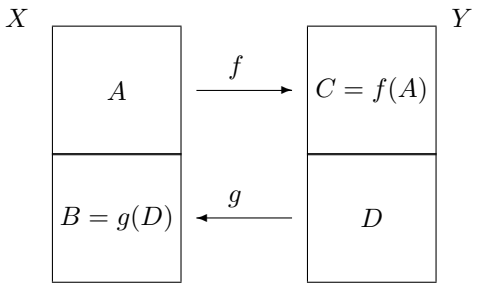
The following theorem which is interesting for its own sake will be used to prove the Schroder Bernstein theorem.

Theorem 2.1.2 *Let $f : X \to Y$ and $g : Y \to X$ be two functions. Then there exist sets A, B, C, D, such that*

$$A \cup B = X,\ C \cup D = Y,\ A \cap B = \emptyset,\ C \cap D = \emptyset,$$

$$f(A) = C,\ g(D) = B.$$

The following picture illustrates the conclusion of this theorem.



Proof: Consider the empty set, $\emptyset \subseteq X$. If $y \in Y \setminus f(\emptyset)$, then $g(y) \notin \emptyset$ because $\emptyset$ has no elements. Also, if A, B, C, and D are as described above, A also would have this same property that the empty set has. However, A is probably larger. Therefore, say $A_0 \subseteq X$ satisfies $\mathcal{P}$ if whenever $y \in Y \setminus f(A_0)$, $g(y) \notin A_0$.

$$\mathcal{A} \equiv \{A_0 \subseteq X : A_0 \text{ satisfies } \mathcal{P}\}.$$

Let $A = \cup\mathcal{A}$. If $y \in Y \setminus f(A)$, then for each $A_0 \in \mathcal{A}$, $y \in Y \setminus f(A_0)$ and so $g(y) \notin A_0$. Since $g(y) \notin A_0$ for all $A_0 \in \mathcal{A}$, it follows $g(y) \notin A$. Hence A satisfies $\mathcal{P}$ and is the largest subset of X which does so. Now define

$$C \equiv f(A),\ D \equiv Y \setminus C,\ B \equiv X \setminus A.$$

It only remains to verify that $g(D) = B$.

Suppose $x \in B = X \setminus A$. Then $A \cup \{x\}$ does not satisfy $\mathcal{P}$ and so there exists $y \in Y \setminus f(A \cup \{x\}) \subseteq D$ such that $g(y) \in A \cup \{x\}$. But $y \notin f(A)$ and so since A satisfies $\mathcal{P}$, it follows $g(y) \notin A$. Hence $g(y) = x$ and so $x \in g(D)$ and This proves the theorem. ■

Theorem 2.1.3 *(Schroder Bernstein) If $f : X \to Y$ and $g : Y \to X$ are one to one, then there exists $h : X \to Y$ which is one to one and onto.*

Proof: Let A, B, C, D be the sets of Theorem2.1.2 and define

$$h(x) \equiv \begin{cases} f(x) & \text{if } x \in A \\ g^{-1}(x) & \text{if } x \in B \end{cases}$$

Then h is the desired one to one and onto mapping.

Recall that the Cartesian product may be considered as the collection of choice functions.

Definition 2.1.4 *Let I be a set and let X_i be a set for each $i \in I$. f is a choice function written as*

$$f \in \prod_{i \in I} X_i$$

if $f(i) \in X_i$ for each $i \in I$.

The axiom of choice says that if $X_i \neq \emptyset$ for each $i \in I$, for I a set, then

$$\prod_{i \in I} X_i \neq \emptyset.$$

Sometimes the two functions, f and g are onto but not one to one. It turns out that with the axiom of choice, a similar conclusion to the above may be obtained.

Corollary 2.1.5 *If $f : X \to Y$ is onto and $g : Y \to X$ is onto, then there exists $h : X \to Y$ which is one to one and onto.*

Proof: For each $y \in Y$, $f^{-1}(y) \equiv \{x \in X : f(x) = y\} \neq \emptyset$. Therefore, by the axiom of choice, there exists $f_0^{-1} \in \prod_{y \in Y} f^{-1}(y)$ which is the same as saying that for each $y \in Y$, $f_0^{-1}(y) \in f^{-1}(y)$. Similarly, there exists $g_0^{-1}(x) \in g^{-1}(x)$ for all $x \in X$. Then f_0^{-1} is one to one because if $f_0^{-1}(y_1) = f_0^{-1}(y_2)$, then

$$y_1 = f\left(f_0^{-1}(y_1)\right) = f\left(f_0^{-1}(y_2)\right) = y_2.$$

Similarly g_0^{-1} is one to one. Therefore, by the Schroder Bernstein theorem, there exists $h : X \to Y$ which is one to one and onto.

Definition 2.1.6 *A set S, is finite if there exists a natural number n and a map θ which maps $\{1, \cdots, n\}$ one to one and onto S. S is infinite if it is not finite. A set S, is called countable if there exists a map θ mapping $\mathbb{N}$ one to one and onto S.(When θ maps a set A to a set B, this will be written as $\theta : A \to B$ in the future.) Here $\mathbb{N} \equiv \{1, 2, \cdots\}$, the natural numbers. S is at most countable if there exists a map $\theta : \mathbb{N} \to S$ which is onto.*

The property of being at most countable is often referred to as being countable because the question of interest is normally whether one can list all elements of the set, designating a first, second, third etc. in such a way as to give each element of the set a natural number. The possibility that a single element of the set may be counted more than once is often not important.

Theorem 2.1.7 *If X and Y are both at most countable, then $X \times Y$ is also at most countable. If either X or Y is countable, then $X \times Y$ is also countable.*

Proof: It is given that there exists a mapping $\eta : \mathbb{N} \to X$ which is onto. Define $\eta(i) \equiv x_i$ and consider X as the set $\{x_1, x_2, x_3, \cdots\}$. Similarly, consider Y as the set

$\{y_1, y_2, y_3, \cdots\}$. It follows the elements of $X \times Y$ are included in the following rectangular array.

$$\begin{array}{llll} (x_1,y_1) & (x_1,y_2) & (x_1,y_3) & \cdots \quad \leftarrow \text{Those which have } x_1 \text{ in first slot.} \\ (x_2,y_1) & (x_2,y_2) & (x_2,y_3) & \cdots \quad \leftarrow \text{Those which have } x_2 \text{ in first slot.} \\ (x_3,y_1) & (x_3,y_2) & (x_3,y_3) & \cdots \quad \leftarrow \text{Those which have } x_3 \text{ in first slot.} \\ \vdots & \vdots & \vdots & \qquad\qquad\qquad\qquad \vdots \end{array} \cdot$$

Follow a path through this array as follows.

$$\begin{array}{ccccccc} (x_1,y_1) & \rightarrow & (x_1,y_2) & & (x_1,y_3) & \rightarrow \\ & \swarrow & & \nearrow & \\ (x_2,y_1) & & (x_2,y_2) & & \\ \downarrow & \nearrow & & & \\ (x_3,y_1) & & & & \end{array}$$

Thus the first element of $X \times Y$ is (x_1, y_1), the second element of $X \times Y$ is (x_1, y_2), the third element of $X \times Y$ is (x_2, y_1) etc. This assigns a number from $\mathbb{N}$ to each element of $X \times Y$. Thus $X \times Y$ is at most countable.

It remains to show the last claim. Suppose without loss of generality that X is countable. Then there exists $\alpha : \mathbb{N} \to X$ which is one to one and onto. Let $\beta : X \times Y \to \mathbb{N}$ be defined by $\beta((x,y)) \equiv \alpha^{-1}(x)$. Thus β is onto $\mathbb{N}$. By the first part there exists a function from $\mathbb{N}$ onto $X \times Y$. Therefore, by Corollary 2.1.5, there exists a one to one and onto mapping from $X \times Y$ to $\mathbb{N}$. This proves the theorem. ■

Theorem 2.1.8 *If X and Y are at most countable, then $X \cup Y$ is at most countable. If either X or Y are countable, then $X \cup Y$ is countable.*

Proof: As in the preceding theorem,

$$X = \{x_1, x_2, x_3, \cdots\}$$

and

$$Y = \{y_1, y_2, y_3, \cdots\}.$$

Consider the following array consisting of $X \cup Y$ and path through it.

$$\begin{array}{ccccccc} x_1 & \rightarrow & x_2 & & x_3 & \rightarrow \\ & \swarrow & & \nearrow & \\ y_1 & \rightarrow & y_2 & & & \end{array}$$

Thus the first element of $X \cup Y$ is x_1, the second is x_2 the third is y_1 the fourth is y_2 etc.

Consider the second claim. By the first part, there is a map from $\mathbb{N}$ onto $X \times Y$. Suppose without loss of generality that X is countable and $\alpha : \mathbb{N} \to X$ is one to one and onto. Then define $\beta(y) \equiv 1$, for all $y \in Y$,and $\beta(x) \equiv \alpha^{-1}(x)$. Thus, β maps $X \times Y$ onto $\mathbb{N}$ and this shows there exist two onto maps, one mapping $X \cup Y$ onto $\mathbb{N}$ and the other mapping $\mathbb{N}$ onto $X \cup Y$. Then Corollary 2.1.5 yields the conclusion. This proves the theorem. ■

2.1.3 Equivalence Relations

There are many ways to compare elements of a set other than to say two elements are equal or the same. For example, in the set of people let two people be equivalent if they

have the same weight. This would not be saying they were the same person, just that they weighed the same. Often such relations involve considering one characteristic of the elements of a set and then saying the two elements are equivalent if they are the same as far as the given characteristic is concerned.

Definition 2.1.9 *Let S be a set. $\sim$ is an equivalence relation on S if it satisfies the following axioms.*

1. $x \sim x$ *for all* $x \in S$. *(Reflexive)*
2. *If* $x \sim y$ *then* $y \sim x$. *(Symmetric)*
3. *If* $x \sim y$ *and* $y \sim z$, *then* $x \sim z$. *(Transitive)*

Definition 2.1.10 $[x]$ *denotes the set of all elements of S which are equivalent to x and $[x]$ is called the equivalence class determined by x or just the equivalence class of x.*

With the above definition one can prove the following simple theorem.

Theorem 2.1.11 *Let $\sim$ be an equivalence class defined on a set, S and let $\mathcal{H}$ denote the set of equivalence classes. Then if $[x]$ and $[y]$ are two of these equivalence classes, either $x \sim y$ and $[x] = [y]$ or it is not true that $x \sim y$ and $[x] \cap [y] = \emptyset$.*

2.2 lim sup And lim inf

It is assumed in all that is done that $\mathbb{R}$ is complete. There are two ways to describe completeness of $\mathbb{R}$. One is to say that every bounded set has a least upper bound and a greatest lower bound. The other is to say that every Cauchy sequence converges. These two equivalent notions of completeness will be taken as given.

The symbol, $\mathbb{F}$ will mean either $\mathbb{R}$ or $\mathbb{C}$. The symbol $[-\infty, \infty]$ will mean all real numbers along with $+\infty$ and $-\infty$ which are points which we pretend are at the right and left ends of the real line respectively. The inclusion of these make believe points makes the statement of certain theorems less trouble.

Definition 2.2.1 *For $A \subseteq [-\infty, \infty]$, $A \neq \emptyset$ $\sup A$ is defined as the least upper bound in case A is bounded above by a real number and equals ∞ if A is not bounded above. Similarly* $\inf A$ *is defined to equal the greatest lower bound in case A is bounded below by a real number and equals $-\infty$ in case A is not bounded below.*

Lemma 2.2.2 *If $\{A_n\}$ is an increasing sequence in $[-\infty, \infty]$, then*

$$\sup \{A_n\} = \lim_{n \to \infty} A_n.$$

Similarly, if $\{A_n\}$ is decreasing, then

$$\inf \{A_n\} = \lim_{n \to \infty} A_n.$$

Proof: Let $\sup (\{A_n : n \in \mathbb{N}\}) = r$. In the first case, suppose $r < \infty$. Then letting $\varepsilon > 0$ be given, there exists n such that $A_n \in (r - \varepsilon, r]$. Since $\{A_n\}$ is increasing, it follows if $m > n$, then $r - \varepsilon < A_n \leq A_m \leq r$ and so $\lim_{n \to \infty} A_n = r$ as claimed. In the case where $r = \infty$, then if a is a real number, there exists n such that $A_n > a$. Since $\{A_k\}$ is increasing, it follows that if $m > n$, $A_m > a$. But this is what is meant by $\lim_{n \to \infty} A_n = \infty$. The other case is that $r = -\infty$. But in this case, $A_n = -\infty$ for all

n and so $\lim_{n\to\infty} A_n = -\infty$. The case where A_n is decreasing is entirely similar. This proves the lemma. ■

Sometimes the limit of a sequence does not exist. For example, if $a_n = (-1)^n$, then $\lim_{n\to\infty} a_n$ does not exist. This is because the terms of the sequence are a distance of 1 apart. Therefore there can't exist a single number such that all the terms of the sequence are ultimately within 1/4 of that number. The nice thing about lim sup and lim inf is that they **always** exist. First here is a simple lemma and definition.

Definition 2.2.3 *Denote by $[-\infty,\infty]$ the real line along with symbols ∞ and $-\infty$. It is understood that ∞ is larger than every real number and $-\infty$ is smaller than every real number. Then if $\{A_n\}$ is an increasing sequence of points of $[-\infty,\infty]$, $\lim_{n\to\infty} A_n$ equals ∞ if the only upper bound of the set $\{A_n\}$ is ∞. If $\{A_n\}$ is bounded above by a real number, then $\lim_{n\to\infty} A_n$ is defined in the usual way and equals the least upper bound of $\{A_n\}$. If $\{A_n\}$ is a decreasing sequence of points of $[-\infty,\infty]$, $\lim_{n\to\infty} A_n$ equals $-\infty$ if the only lower bound of the sequence $\{A_n\}$ is $-\infty$. If $\{A_n\}$ is bounded below by a real number, then $\lim_{n\to\infty} A_n$ is defined in the usual way and equals the greatest lower bound of $\{A_n\}$. More simply, if $\{A_n\}$ is increasing,*

$$\lim_{n\to\infty} A_n = \sup\{A_n\}$$

and if $\{A_n\}$ is decreasing then

$$\lim_{n\to\infty} A_n = \inf\{A_n\}.$$

Lemma 2.2.4 *Let $\{a_n\}$ be a sequence of real numbers and let $U_n \equiv \sup\{a_k : k \geq n\}$. Then $\{U_n\}$ is a decreasing sequence. Also if $L_n \equiv \inf\{a_k : k \geq n\}$, then $\{L_n\}$ is an increasing sequence. Therefore, $\lim_{n\to\infty} L_n$ and $\lim_{n\to\infty} U_n$ both exist.*

Proof: Let W_n be an upper bound for $\{a_k : k \geq n\}$. Then since these sets are getting smaller, it follows that for $m < n$, W_m is an upper bound for $\{a_k : k \geq n\}$. In particular if $W_m = U_m$, then U_m is an upper bound for $\{a_k : k \geq n\}$ and so U_m is at least as large as U_n, the least upper bound for $\{a_k : k \geq n\}$. The claim that $\{L_n\}$ is decreasing is similar. This proves the lemma. ■

From the lemma, the following definition makes sense.

Definition 2.2.5 *Let $\{a_n\}$ be any sequence of points of $[-\infty,\infty]$*

$$\limsup_{n\to\infty} a_n \equiv \lim_{n\to\infty} \sup\{a_k : k \geq n\}$$

$$\liminf_{n\to\infty} a_n \equiv \lim_{n\to\infty} \inf\{a_k : k \geq n\}.$$

Theorem 2.2.6 *Suppose $\{a_n\}$ is a sequence of real numbers and that*

$$\limsup_{n\to\infty} a_n$$

and

$$\liminf_{n\to\infty} a_n$$

are both real numbers. Then $\lim_{n\to\infty} a_n$ exists if and only if

$$\liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n$$

and in this case,

$$\lim_{n\to\infty} a_n = \liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n.$$

Proof: First note that

$$\sup\{a_k : k \geq n\} \geq \inf\{a_k : k \geq n\}$$

and so from Theorem 4.1.7,

$$\begin{aligned}\limsup_{n\to\infty} a_n &\equiv \lim_{n\to\infty} \sup\{a_k : k \geq n\} \\ &\geq \lim_{n\to\infty} \inf\{a_k : k \geq n\} \\ &\equiv \liminf_{n\to\infty} a_n.\end{aligned}$$

Suppose first that $\lim_{n\to\infty} a_n$ exists and is a real number. Then by Theorem 4.4.3 $\{a_n\}$ is a Cauchy sequence. Therefore, if $\varepsilon > 0$ is given, there exists N such that if $m, n \geq N$, then

$$|a_n - a_m| < \varepsilon/3.$$

From the definition of $\sup\{a_k : k \geq N\}$, there exists $n_1 \geq N$ such that

$$\sup\{a_k : k \geq N\} \leq a_{n_1} + \varepsilon/3.$$

Similarly, there exists $n_2 \geq N$ such that

$$\inf\{a_k : k \geq N\} \geq a_{n_2} - \varepsilon/3.$$

It follows that

$$\sup\{a_k : k \geq N\} - \inf\{a_k : k \geq N\} \leq |a_{n_1} - a_{n_2}| + \frac{2\varepsilon}{3} < \varepsilon.$$

Since the sequence, $\{\sup\{a_k : k \geq N\}\}_{N=1}^{\infty}$ is decreasing and $\{\inf\{a_k : k \geq N\}\}_{N=1}^{\infty}$ is increasing, it follows from Theorem 4.1.7

$$0 \leq \lim_{N\to\infty} \sup\{a_k : k \geq N\} - \lim_{N\to\infty} \inf\{a_k : k \geq N\} \leq \varepsilon$$

Since ε is arbitrary, this shows

$$\lim_{N\to\infty} \sup\{a_k : k \geq N\} = \lim_{N\to\infty} \inf\{a_k : k \geq N\} \tag{2.1}$$

Next suppose 2.1. Then

$$\lim_{N\to\infty} (\sup\{a_k : k \geq N\} - \inf\{a_k : k \geq N\}) = 0$$

Since $\sup\{a_k : k \geq N\} \geq \inf\{a_k : k \geq N\}$ it follows that for every $\varepsilon > 0$, there exists N such that

$$\sup\{a_k : k \geq N\} - \inf\{a_k : k \geq N\} < \varepsilon$$

Thus if $m, n > N$, then

$$|a_m - a_n| < \varepsilon$$

which means $\{a_n\}$ is a Cauchy sequence. Since $\mathbb{R}$ is complete, it follows that $\lim_{n\to\infty} a_n \equiv a$ exists. By the squeezing theorem, it follows

$$a = \liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n$$

and This proves the theorem. ■

With the above theorem, here is how to define the limit of a sequence of points in $[-\infty, \infty]$.

Definition 2.2.7 *Let $\{a_n\}$ be a sequence of points of $[-\infty,\infty]$. Then $\lim_{n\to\infty} a_n$ exists exactly when*

$$\lim_{n\to\infty}\inf a_n = \lim_{n\to\infty}\sup a_n$$

and in this case

$$\lim_{n\to\infty} a_n \equiv \lim_{n\to\infty}\inf a_n = \lim_{n\to\infty}\sup a_n.$$

The significance of lim sup and lim inf, in addition to what was just discussed, is contained in the following theorem which follows quickly from the definition.

Theorem 2.2.8 *Suppose $\{a_n\}$ is a sequence of points of $[-\infty,\infty]$. Let*

$$\lambda = \lim_{n\to\infty}\sup a_n.$$

Then if $b>\lambda$, it follows there exists N such that whenever $n\geq N$,

$$a_n \leq b.$$

If $c<\lambda$, then $a_n > c$ for infinitely many values of n. Let

$$\gamma = \lim_{n\to\infty}\inf a_n.$$

Then if $d<\gamma$, it follows there exists N such that whenever $n\geq N$,

$$a_n \geq d.$$

If $e>\gamma$, it follows $a_n < e$ for infinitely many values of n.

The proof of this theorem is left as an exercise for you. It follows directly from the definition and it is the sort of thing you must do yourself. Here is one other simple proposition.

Proposition 2.2.9 *Let $\lim_{n\to\infty} a_n = a > 0$. Then*

$$\lim_{n\to\infty}\sup a_n b_n = a\lim_{n\to\infty}\sup b_n.$$

Proof: This follows from the definition. Let $\lambda_n = \sup\{a_k b_k : k\geq n\}$. For all n large enough, $a_n > a-\varepsilon$ where ε is small enough that $a-\varepsilon>0$. Therefore,

$$\lambda_n \geq \sup\{b_k : k\geq n\}(a-\varepsilon)$$

for all n large enough. Then

$$\begin{aligned}\lim_{n\to\infty}\sup a_n b_n &= \lim_{n\to\infty}\lambda_n \equiv \lim_{n\to\infty}\sup a_n b_n\\ &\geq \lim_{n\to\infty}\left(\sup\{b_k : k\geq n\}(a-\varepsilon)\right)\\ &= (a-\varepsilon)\lim_{n\to\infty}\sup b_n\end{aligned}$$

Similar reasoning shows

$$\lim_{n\to\infty}\sup a_n b_n \leq (a+\varepsilon)\lim_{n\to\infty}\sup b_n$$

Now since $\varepsilon>0$ is arbitrary, the conclusion follows.

2.3 Double Series

Sometimes it is required to consider double series which are of the form

$$\sum_{k=m}^{\infty}\sum_{j=m}^{\infty}a_{jk} \equiv \sum_{k=m}^{\infty}\left(\sum_{j=m}^{\infty}a_{jk}\right).$$

In other words, first sum on j yielding something which depends on k and then sum these. The major consideration for these double series is the question of when

$$\sum_{k=m}^{\infty}\sum_{j=m}^{\infty}a_{jk} = \sum_{j=m}^{\infty}\sum_{k=m}^{\infty}a_{jk}.$$

In other words, when does it make no difference which subscript is summed over first? In the case of finite sums there is no issue here. You can always write

$$\sum_{k=m}^{M}\sum_{j=m}^{N}a_{jk} = \sum_{j=m}^{N}\sum_{k=m}^{M}a_{jk}$$

because addition is commutative. However, there are limits involved with infinite sums and the interchange in order of summation involves taking limits in a different order. Therefore, it is not always true that it is permissible to interchange the two sums. A general rule of thumb is this: If something involves changing the order in which two limits are taken, you may not do it without agonizing over the question. In general, limits foul up algebra and also introduce things which are counter intuitive. Here is an example. This example is a little technical. It is placed here just to prove conclusively there is a question which needs to be considered.

Example 2.3.1 *Consider the following picture which depicts some of the ordered pairs* (m,n) *where* m,n *are positive integers.*

The numbers next to the point are the values of a_{mn}. You see $a_{nn} = 0$ for all n, $a_{21} = a, a_{12} = b, a_{mn} = c$ for (m,n) on the line $y = 1 + x$ whenever $m > 1$, and $a_{mn} = -c$ for all (m,n) on the line $y = x - 1$ whenever $m > 2$.

Then $\sum_{m=1}^{\infty} a_{mn} = a$ if $n = 1$, $\sum_{m=1}^{\infty} a_{mn} = b - c$ if $n = 2$ and if $n > 2$, $\sum_{m=1}^{\infty} a_{mn} = 0$. Therefore,

$$\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} a_{mn} = a + b - c.$$

Next observe that $\sum_{n=1}^{\infty} a_{mn} = b$ if $m = 1$, $\sum_{n=1}^{\infty} a_{mn} = a + c$ if $m = 2$, and $\sum_{n=1}^{\infty} a_{mn} = 0$ if $m > 2$. Therefore,

$$\sum_{m=1}^{\infty}\sum_{n=1}^{\infty} a_{mn} = b + a + c$$

and so the two sums are different. Moreover, you can see that by assigning different values of a, b, and c, you can get an example for any two different numbers desired.

It turns out that if $a_{ij} \geq 0$ for all i, j, then you can always interchange the order of summation. This is shown next and is based on the following lemma. First, some notation should be discussed.

Definition 2.3.2 *Let $f(a,b) \in [-\infty, \infty]$ for $a \in A$ and $b \in B$ where A, B are sets which means that $f(a,b)$ is either a number, ∞, or $-\infty$. The symbol, $+\infty$ is interpreted as a point out at the end of the number line which is larger than every real number. Of course there is no such number. That is why it is called ∞. The symbol, $-\infty$ is interpreted similarly. Then $\sup_{a \in A} f(a,b)$ means $\sup(S_b)$ where $S_b \equiv \{f(a,b) : a \in A\}$.*

Unlike limits, you can take the sup in different orders.

Lemma 2.3.3 *Let $f(a,b) \in [-\infty,\infty]$ for $a \in A$ and $b \in B$ where A, B are sets. Then*

$$\sup_{a\in A}\sup_{b\in B} f(a,b) = \sup_{b\in B}\sup_{a\in A} f(a,b).$$

Proof: Note that for all a, b, $f(a,b) \leq \sup_{b\in B}\sup_{a\in A} f(a,b)$ and therefore, for all a, $\sup_{b\in B} f(a,b) \leq \sup_{b\in B}\sup_{a\in A} f(a,b)$. Therefore,

$$\sup_{a\in A}\sup_{b\in B} f(a,b) \leq \sup_{b\in B}\sup_{a\in A} f(a,b).$$

Repeat the same argument interchanging a and b, to get the conclusion of the lemma.

Theorem 2.3.4 *Let $a_{ij} \geq 0$. Then*

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij}.$$

Proof: First note there is no trouble in defining these sums because the a_{ij} are all nonnegative. If a sum diverges, it only diverges to ∞ and so ∞ is the value of the sum. Next note that

$$\sum_{j=r}^{\infty}\sum_{i=r}^{\infty} a_{ij} \geq \sup_n \sum_{j=r}^{\infty}\sum_{i=r}^{n} a_{ij}$$

because for all j,

$$\sum_{i=r}^{\infty} a_{ij} \geq \sum_{i=r}^{n} a_{ij}.$$

Therefore,

$$\sum_{j=r}^{\infty}\sum_{i=r}^{\infty} a_{ij} \geq \sup_n \sum_{j=r}^{\infty}\sum_{i=r}^{n} a_{ij} = \sup_n \lim_{m\to\infty} \sum_{j=r}^{m}\sum_{i=r}^{n} a_{ij}$$

$$\begin{aligned} &= \sup_n \lim_{m\to\infty} \sum_{i=r}^{n}\sum_{j=r}^{m} a_{ij} = \sup_n \sum_{i=r}^{n} \lim_{m\to\infty} \sum_{j=r}^{m} a_{ij} \\ &= \sup_n \sum_{i=r}^{n}\sum_{j=r}^{\infty} a_{ij} = \lim_{n\to\infty} \sum_{i=r}^{n}\sum_{j=r}^{\infty} a_{ij} = \sum_{i=r}^{\infty}\sum_{j=r}^{\infty} a_{ij} \end{aligned}$$

Interchanging the i and j in the above argument proves the theorem.

Chapter 3

Basic Linear Algebra

All the topics for calculus of one variable generalize to calculus of any number of variables in which the functions can have values in m dimensional space and there is more than one variable.

The notation, $\mathbb{C}^n$ refers to the collection of ordered lists of n complex numbers. Since every real number is also a complex number, this simply generalizes the usual notion of $\mathbb{R}^n$, the collection of all ordered lists of n real numbers. In order to avoid worrying about whether it is real or complex numbers which are being referred to, the symbol $\mathbb{F}$ will be used. If it is not clear, always pick $\mathbb{C}$.

Definition 3.0.5 *Define*

$$\mathbb{F}^n \equiv \{(x_1, \cdots, x_n) : x_j \in \mathbb{F} \text{ for } j = 1, \cdots, n\}.$$

$(x_1, \cdots, x_n) = (y_1, \cdots, y_n)$ *if and only if for all* $j = 1, \cdots, n$, $x_j = y_j$. *When*

$$(x_1, \cdots, x_n) \in \mathbb{F}^n,$$

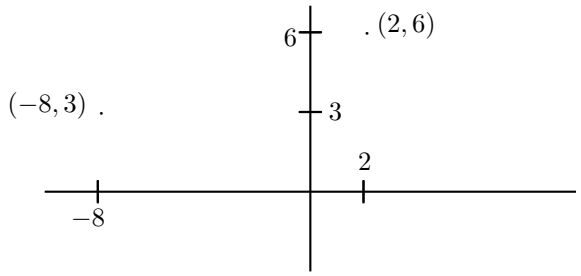
it is conventional to denote $(x_1, \cdots, x_n)$ *by the single bold face letter,* $\mathbf{x}$. *The numbers,* x_j *are called the coordinates. The set*

$$\{(0, \cdots, 0, t, 0, \cdots, 0) : t \in \mathbb{F}\}$$

for t *in the* i^{th} *slot is called the* i^{th} *coordinate axis. The point* $\mathbf{0} \equiv (0, \cdots, 0)$ *is called the origin.*

Thus $(1, 2, 4i) \in \mathbb{F}^3$ and $(2, 1, 4i) \in \mathbb{F}^3$ but $(1, 2, 4i) \neq (2, 1, 4i)$ because, even though the same numbers are involved, they don't match up. In particular, the first entries are not equal.

The geometric significance of $\mathbb{R}^n$ for $n \leq 3$ has been encountered already in calculus or in precalculus. Here is a short review. First consider the case when $n = 1$. Then from the definition, $\mathbb{R}^1 = \mathbb{R}$. Recall that $\mathbb{R}$ is identified with the points of a line. Look at the number line again. Observe that this amounts to identifying a point on this line with a real number. In other words a real number determines where you are on this line. Now suppose $n = 2$ and consider two lines which intersect each other at right angles as shown in the following picture.



Notice how you can identify a point shown in the plane with the ordered pair, $(2,6)$. You go to the right a distance of 2 and then up a distance of 6. Similarly, you can identify another point in the plane with the ordered pair $(-8,3)$. Go to the left a distance of 8 and then up a distance of 3. The reason you go to the left is that there is a $-$ sign on the eight. From this reasoning, every ordered pair determines a unique point in the plane. Conversely, taking a point in the plane, you could draw two lines through the point, one vertical and the other horizontal and determine unique points, x_1 on the horizontal line in the above picture and x_2 on the vertical line in the above picture, such that the point of interest is identified with the ordered pair, (x_1, x_2). In short, points in the plane can be identified with ordered pairs similar to the way that points on the real line are identified with real numbers. Now suppose $n = 3$. As just explained, the first two coordinates determine a point in a plane. Letting the third component determine how far up or down you go, depending on whether this number is positive or negative, this determines a point in space. Thus, $(1,4,-5)$ would mean to determine the point in the plane that goes with $(1,4)$ and then to go below this plane a distance of 5 to obtain a unique point in space. You see that the ordered triples correspond to points in space just as the ordered pairs correspond to points in a plane and single real numbers correspond to points on a line.

You can't stop here and say that you are only interested in $n \leq 3$. What if you were interested in the motion of two objects? You would need three coordinates to describe where the first object is and you would need another three coordinates to describe where the other object is located. Therefore, you would need to be considering $\mathbb{R}^6$. If the two objects moved around, you would need a time coordinate as well. As another example, consider a hot object which is cooling and suppose you want the temperature of this object. How many coordinates would be needed? You would need one for the temperature, three for the position of the point in the object and one more for the time. Thus you would need to be considering $\mathbb{R}^5$. Many other examples can be given. Sometimes n is very large. This is often the case in applications to business when they are trying to maximize profit subject to constraints. It also occurs in numerical analysis when people try to solve hard problems on a computer.

There are other ways to identify points in space with three numbers but the one presented is the most basic. In this case, the coordinates are known as Cartesian coordinates after Descartes[1] who invented this idea in the first half of the seventeenth century. I will often not bother to draw a distinction between the point in n dimensional space and its Cartesian coordinates.

The geometric significance of $\mathbb{C}^n$ for $n > 1$ is not available because each copy of $\mathbb{C}$ corresponds to the plane or $\mathbb{R}^2$.

[1]René Descartes 1596-1650 is often credited with inventing analytic geometry although it seems the ideas were actually known much earlier. He was interested in many different subjects, physiology, chemistry, and physics being some of them. He also wrote a large book in which he tried to explain the book of Genesis scientifically. Descartes ended up dying in Sweden.

3.1 Algebra in $\mathbb{F}^n$, Vector Spaces

There are two algebraic operations done with elements of $\mathbb{F}^n$. One is addition and the other is multiplication by numbers, called scalars. In the case of $\mathbb{C}^n$ the scalars are complex numbers while in the case of $\mathbb{R}^n$ the only allowed scalars are real numbers. Thus, the scalars always come from $\mathbb{F}$ in either case.

Definition 3.1.1 *If $\mathbf{x} \in \mathbb{F}^n$ and $a \in \mathbb{F}$, also called a scalar, then $a\mathbf{x} \in \mathbb{F}^n$ is defined by*

$$a\mathbf{x} = a(x_1, \cdots, x_n) \equiv (ax_1, \cdots, ax_n). \tag{3.1}$$

This is known as scalar multiplication. If $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$ then $\mathbf{x} + \mathbf{y} \in \mathbb{F}^n$ and is defined by

$$\begin{aligned} \mathbf{x} + \mathbf{y} &= (x_1, \cdots, x_n) + (y_1, \cdots, y_n) \\ &\equiv (x_1 + y_1, \cdots, x_n + y_n) \end{aligned} \tag{3.2}$$

the points in $\mathbb{F}^n$ are also referred to as vectors.

With this definition, the algebraic properties satisfy the conclusions of the following theorem. These conclusions are called the vector space axioms. Any time you have a set and a field of scalars satisfying the axioms of the following theorem, it is called a vector space.

Theorem 3.1.2 *For $\mathbf{v}, \mathbf{w} \in \mathbb{F}^n$ and α, β scalars, (real numbers), the following hold.*

$$\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}, \tag{3.3}$$

the commutative law of addition,

$$(\mathbf{v} + \mathbf{w}) + \mathbf{z} = \mathbf{v} + (\mathbf{w} + \mathbf{z}), \tag{3.4}$$

the associative law for addition,

$$\mathbf{v} + \mathbf{0} = \mathbf{v}, \tag{3.5}$$

the existence of an additive identity,

$$\mathbf{v} + (-\mathbf{v}) = \mathbf{0}, \tag{3.6}$$

the existence of an additive inverse, Also

$$\alpha(\mathbf{v} + \mathbf{w}) = \alpha\mathbf{v} + \alpha\mathbf{w}, \tag{3.7}$$

$$(\alpha + \beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}, \tag{3.8}$$

$$\alpha(\beta\mathbf{v}) = \alpha\beta(\mathbf{v}), \tag{3.9}$$

$$1\mathbf{v} = \mathbf{v}. \tag{3.10}$$

In the above $\mathbf{0} = (0, \cdots, 0)$.

You should verify these properties all hold. For example, consider 3.7

$$\begin{aligned} \alpha(\mathbf{v} + \mathbf{w}) &= \alpha(v_1 + w_1, \cdots, v_n + w_n) \\ &= (\alpha(v_1 + w_1), \cdots, \alpha(v_n + w_n)) \\ &= (\alpha v_1 + \alpha w_1, \cdots, \alpha v_n + \alpha w_n) \\ &= (\alpha v_1, \cdots, \alpha v_n) + (\alpha w_1, \cdots, \alpha w_n) \\ &= \alpha\mathbf{v} + \alpha\mathbf{w}. \end{aligned}$$

As usual subtraction is defined as $\mathbf{x} - \mathbf{y} \equiv \mathbf{x} + (-\mathbf{y})$.

3.2 Subspaces Spans And Bases

The concept of linear combination is fundamental in all of linear algebra.

Definition 3.2.1 *Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ be vectors in a vector space, Y having the field of scalars $\mathbb{F}$. A linear combination is any expression of the form*

$$\sum_{i=1}^{p} c_i \mathbf{x}_i$$

where the c_i are scalars. The set of all linear combinations of these vectors is called $\operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_n)$. *If $V \subseteq Y$, then V is called a subspace if whenever α, β are scalars and $\mathbf{u}$ and $\mathbf{v}$ are vectors of V, it follows $\alpha\mathbf{u} + \beta\mathbf{v} \in V$. That is, it is "closed under the algebraic operations of vector addition and scalar multiplication" and is therefore, a vector space. A linear combination of vectors is said to be trivial if all the scalars in the linear combination equal zero. A set of vectors is said to be linearly independent if the only linear combination of these vectors which equals the zero vector is the trivial linear combination. Thus $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ is called linearly independent if whenever*

$$\sum_{k=1}^{p} c_k \mathbf{x}_k = \mathbf{0}$$

it follows that all the scalars, c_k equal zero. A set of vectors, $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$, is called linearly dependent if it is not linearly independent. Thus the set of vectors is linearly dependent if there exist scalars, $c_i, i = 1, \cdots, n$, not all zero such that $\sum_{k=1}^{p} c_k \mathbf{x}_k = \mathbf{0}$.

Lemma 3.2.2 *A set of vectors $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ is linearly independent if and only if none of the vectors can be obtained as a linear combination of the others.*

Proof: Suppose first that $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ is linearly independent. If

$$\mathbf{x}_k = \sum_{j \neq k} c_j \mathbf{x}_j,$$

then

$$\mathbf{0} = 1\mathbf{x}_k + \sum_{j \neq k} (-c_j)\, \mathbf{x}_j,$$

a nontrivial linear combination, contrary to assumption. This shows that if the set is linearly independent, then none of the vectors is a linear combination of the others.

Now suppose no vector is a linear combination of the others. Is $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ linearly independent? If it is not, there exist scalars, c_i, not all zero such that

$$\sum_{i=1}^{p} c_i \mathbf{x}_i = \mathbf{0}.$$

Say $c_k \neq 0$. Then you can solve for $\mathbf{x}_k$ as

$$\mathbf{x}_k = \sum_{j \neq k} (-c_j)\, / c_k \mathbf{x}_j$$

contrary to assumption. This proves the lemma. ■

The following is called the exchange theorem.

Theorem 3.2.3 *Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ be a linearly independent set of vectors such that each $\mathbf{x}_i$ is in the* $\operatorname{span}\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$. *Then $r \leq s$.*

Proof: Define span $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\} \equiv V$, it follows there exist scalars, $c_1, \cdots, c_s$ such that

$$\mathbf{x}_1 = \sum_{i=1}^{s} c_i \mathbf{y}_i. \tag{3.11}$$

Not all of these scalars can equal zero because if this were the case, it would follow that $\mathbf{x}_1 = \mathbf{0}$ and so $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ would not be linearly independent. Indeed, if $\mathbf{x}_1 = \mathbf{0}$, $1\mathbf{x}_1 + \sum_{i=2}^{r} 0\mathbf{x}_i = \mathbf{x}_1 = \mathbf{0}$ and so there would exist a nontrivial linear combination of the vectors $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ which equals zero.

Say $c_k \neq 0$. Then solve (3.11) for $\mathbf{y}_k$ and obtain

$$\mathbf{y}_k \in \text{span}\left(\mathbf{x}_1, \overbrace{\mathbf{y}_1, \cdots, \mathbf{y}_{k-1}, \mathbf{y}_{k+1}, \cdots, \mathbf{y}_s}^{\text{s-1 vectors here}}\right).$$

Define $\{\mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\}$ by

$$\{\mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\} \equiv \{\mathbf{y}_1, \cdots, \mathbf{y}_{k-1}, \mathbf{y}_{k+1}, \cdots, \mathbf{y}_s\}$$

Therefore, span $\{\mathbf{x}_1, \mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\} = V$ because if $\mathbf{v} \in V$, there exist constants $c_1, \cdots, c_s$ such that

$$\mathbf{v} = \sum_{i=1}^{s-1} c_i \mathbf{z}_i + c_s \mathbf{y}_k.$$

Now replace the $\mathbf{y}_k$ in the above with a linear combination of the vectors, $\{\mathbf{x}_1, \mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\}$ to obtain $\mathbf{v} \in$ span $\{\mathbf{x}_1, \mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\}$. The vector $\mathbf{y}_k$, in the list $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$, has now been replaced with the vector $\mathbf{x}_1$ and the resulting modified list of vectors has the same span as the original list of vectors, $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$.

Now suppose that $r > s$ and that span $\{\mathbf{x}_1, \cdots, \mathbf{x}_l, \mathbf{z}_1, \cdots, \mathbf{z}_p\} = V$ where the vectors, $\mathbf{z}_1, \cdots, \mathbf{z}_p$ are each taken from the set, $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$ and $l + p = s$. This has now been done for $l = 1$ above. Then since $r > s$, it follows that $l \leq s < r$ and so $l + 1 \leq r$. Therefore, $\mathbf{x}_{l+1}$ is a vector not in the list, $\{\mathbf{x}_1, \cdots, \mathbf{x}_l\}$ and since span $\{\mathbf{x}_1, \cdots, \mathbf{x}_l, \mathbf{z}_1, \cdots, \mathbf{z}_p\} = V$ there exist scalars, c_i and d_j such that

$$\mathbf{x}_{l+1} = \sum_{i=1}^{l} c_i \mathbf{x}_i + \sum_{j=1}^{p} d_j \mathbf{z}_j. \tag{3.12}$$

Now not all the d_j can equal zero because if this were so, it would follow that $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ would be a linearly dependent set because one of the vectors would equal a linear combination of the others. Therefore, (3.12) can be solved for one of the $\mathbf{z}_i$, say $\mathbf{z}_k$, in terms of $\mathbf{x}_{l+1}$ and the other $\mathbf{z}_i$ and just as in the above argument, replace that $\mathbf{z}_i$ with $\mathbf{x}_{l+1}$ to obtain

$$\text{span}\left(\mathbf{x}_1, \cdots \mathbf{x}_l, \mathbf{x}_{l+1}, \overbrace{\mathbf{z}_1, \cdots \mathbf{z}_{k-1}, \mathbf{z}_{k+1}, \cdots, \mathbf{z}_p}^{\text{p-1 vectors here}}\right) = V.$$

Continue this way, eventually obtaining

$$\text{span}\,(\mathbf{x}_1, \cdots, \mathbf{x}_s) = V.$$

But then $\mathbf{x}_r \in$ span $\{\mathbf{x}_1, \cdots, \mathbf{x}_s\}$ contrary to the assumption that $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is linearly independent. Therefore, $r \leq s$ as claimed.

Here is another proof in case you didn't like the above proof.

Theorem 3.2.4 *If*

$$\mathrm{span}\left(\mathbf{u}_1,\cdots,\mathbf{u}_r\right)\subseteq \mathrm{span}\left(\mathbf{v}_1,\cdots,\mathbf{v}_s\right)\equiv V$$

and $\{\mathbf{u}_1,\cdots,\mathbf{u}_r\}$ *are linearly independent, then* $r\leq s$.

Proof: Suppose $r>s$. Let E_p denote a finite list of vectors of $\{\mathbf{v}_1,\cdots,\mathbf{v}_s\}$ and let $|E_p|$ denote the number of vectors in the list. Let F_p denote the first p vectors in $\{\mathbf{u}_1,\cdots,\mathbf{u}_r\}$. In case $p=0, F_p$ will denote the empty set. For $0\leq p\leq s$, let E_p have the property

$$\mathrm{span}\left(F_p,E_p\right)=V$$

and $|E_p|$ is as small as possible for this to happen. I claim $|E_p|\leq s-p$ if E_p is nonempty.

Here is why. For $p=0$, it is obvious. Suppose true for some $p<s$. Then

$$\mathbf{u}_{p+1}\in \mathrm{span}\left(F_p,E_p\right)$$

and so there are constants, $c_1,\cdots,c_p$ and $d_1,\cdots,d_m$ where $m\leq s-p$ such that

$$\mathbf{u}_{p+1}=\sum_{i=1}^{p}c_i\mathbf{u}_i+\sum_{j=1}^{m}d_i\mathbf{z}_j$$

for

$$\{\mathbf{z}_1,\cdots,\mathbf{z}_m\}\subseteq\{\mathbf{v}_1,\cdots,\mathbf{v}_s\}.$$

Then not all the d_i can equal zero because this would violate the linear independence of the $\{\mathbf{u}_1,\cdots,\mathbf{u}_r\}$. Therefore, you can solve for one of the $\mathbf{z}_k$ as a linear combination of $\{\mathbf{u}_1,\cdots,\mathbf{u}_{p+1}\}$ and the other $\mathbf{z}_j$. Thus you can change F_p to F_{p+1} and include one fewer vector in E_p. Thus $|E_{p+1}|\leq m-1\leq s-p-1$. This proves the claim.

Therefore, E_s is empty and $\mathrm{span}\left(\mathbf{u}_1,\cdots,\mathbf{u}_s\right)=V$. However, this gives a contradiction because it would require

$$\mathbf{u}_{s+1}\in \mathrm{span}\left(\mathbf{u}_1,\cdots,\mathbf{u}_s\right)$$

which violates the linear independence of these vectors. This proves the theorem. ■

Definition 3.2.5 *A finite set of vectors,* $\{\mathbf{x}_1,\cdots,\mathbf{x}_r\}$ *is a basis for a vector space* V *if*

$$\mathrm{span}\left(\mathbf{x}_1,\cdots,\mathbf{x}_r\right)=V$$

and $\{\mathbf{x}_1,\cdots,\mathbf{x}_r\}$ *is linearly independent. Thus if* $\mathbf{v}\in V$ *there exist unique scalars,* $v_1,\cdots,v_r$ *such that* $\mathbf{v}=\sum_{i=1}^{r}v_i\mathbf{x}_i$. *These scalars are called the components of* $\mathbf{v}$ *with respect to the basis* $\{\mathbf{x}_1,\cdots,\mathbf{x}_r\}$.

Corollary 3.2.6 *Let* $\{\mathbf{x}_1,\cdots,\mathbf{x}_r\}$ *and* $\{\mathbf{y}_1,\cdots,\mathbf{y}_s\}$ *be two bases*[2] *of* $\mathbb{F}^n$. *Then* $r=s=n$.

Proof: From the exchange theorem, $r\leq s$ and $s\leq r$. Now note the vectors,

$$\mathbf{e}_i=\overbrace{(0,\cdots,0,1,0\cdots,0)}^{1\text{ is in the }i^{th}\text{ slot}}$$

for $i=1,2,\cdots,n$ are a basis for $\mathbb{F}^n$. This proves the corollary. ■

[2]This is the plural form of basis. We could say basiss but it would involve an inordinate amount of hissing as in "The sixth shiek's sixth sheep is sick". This is the reason that bases is used instead of basiss.

Lemma 3.2.7 *Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ be a set of vectors. Then $V \equiv \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_r)$ is a subspace.*

Proof: Suppose α, β are two scalars and let $\sum_{k=1}^r c_k \mathbf{v}_k$ and $\sum_{k=1}^r d_k \mathbf{v}_k$ are two elements of V. What about

$$\alpha \sum_{k=1}^r c_k \mathbf{v}_k + \beta \sum_{k=1}^r d_k \mathbf{v}_k?$$

Is it also in V?

$$\alpha \sum_{k=1}^r c_k \mathbf{v}_k + \beta \sum_{k=1}^r d_k \mathbf{v}_k = \sum_{k=1}^r (\alpha c_k + \beta d_k) \mathbf{v}_k \in V$$

so the answer is yes. This proves the lemma. ■

Definition 3.2.8 *Let V be a vector space. Then $\dim(V)$ read as the dimension of V is the number of vectors in a basis.*

Of course you should wonder right now whether an arbitrary subspace of a finite dimensional vector space even has a basis. In fact it does and this is in the next theorem. First, here is an interesting lemma.

Lemma 3.2.9 *Suppose $\mathbf{v} \notin \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$ and $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ is linearly independent. Then $\{\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{v}\}$ is also linearly independent.*

Proof: Suppose $\sum_{i=1}^k c_i \mathbf{u}_i + d\mathbf{v} = \mathbf{0}$. It is required to verify that each $c_i = 0$ and that $d = 0$. But if $d \neq 0$, then you can solve for $\mathbf{v}$ as a linear combination of the vectors, $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$,

$$\mathbf{v} = -\sum_{i=1}^k \left(\frac{c_i}{d}\right) \mathbf{u}_i$$

contrary to assumption. Therefore, $d = 0$. But then $\sum_{i=1}^k c_i \mathbf{u}_i = 0$ and the linear independence of $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ implies each $c_i = 0$ also. This proves the lemma. ■

Theorem 3.2.10 *Let V be a nonzero subspace of Y a finite dimensional vector space having dimension n. Then V has a basis.*

Proof: Let $\mathbf{v}_1 \in V$ where $\mathbf{v}_1 \neq 0$. If $\text{span}\{\mathbf{v}_1\} = V$, stop. $\{\mathbf{v}_1\}$ is a basis for V. Otherwise, there exists $\mathbf{v}_2 \in V$ which is not in $\text{span}\{\mathbf{v}_1\}$. By Lemma 3.2.9 $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a linearly independent set of vectors. If $\text{span}\{\mathbf{v}_1, \mathbf{v}_2\} = V$ stop, $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a basis for V. If $\text{span}\{\mathbf{v}_1, \mathbf{v}_2\} \neq V$, then there exists $\mathbf{v}_3 \notin \text{span}\{\mathbf{v}_1, \mathbf{v}_2\}$ and $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is a larger linearly independent set of vectors. Continuing this way, the process must stop before $n+1$ steps because if not, it would be possible to obtain $n+1$ linearly independent vectors contrary to the exchange theorem and the assumed dimension of Y. This proves the theorem. ■

In words the following corollary states that any linearly independent set of vectors can be enlarged to form a basis.

Corollary 3.2.11 *Let V be a subspace of Y, a finite dimensional vector space of dimension n and let $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ be a linearly independent set of vectors in V. Then either it is a basis for V or there exist vectors, $\mathbf{v}_{r+1}, \cdots, \mathbf{v}_s$ such that $\{\mathbf{v}_1, \cdots, \mathbf{v}_r, \mathbf{v}_{r+1}, \cdots, \mathbf{v}_s\}$ is a basis for V.*

Proof: This follows immediately from the proof of Theorem 3.2.10. You do exactly the same argument except you start with $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ rather than $\{\mathbf{v}_1\}$.

It is also true that any spanning set of vectors can be restricted to obtain a basis.

Theorem 3.2.12 *Let V be a subspace of Y, a finite dimensional vector space of dimension n and suppose* $\operatorname{span}(\mathbf{u}_1 \cdots, \mathbf{u}_p) = V$ *where the $\mathbf{u}_i$ are nonzero vectors. Then there exist vectors, $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\}$ such that $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\} \subseteq \{\mathbf{u}_1 \cdots, \mathbf{u}_p\}$ and $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\}$ is a basis for V.*

Proof: Let r be the smallest positive integer with the property that for some set, $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\} \subseteq \{\mathbf{u}_1 \cdots, \mathbf{u}_p\}$,

$$\operatorname{span}(\mathbf{v}_1 \cdots, \mathbf{v}_r) = V.$$

Then $r \leq p$ and it must be the case that $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\}$ is linearly independent because if it were not so, one of the vectors, say $\mathbf{v}_k$ would be a linear combination of the others. But then you could delete this vector from $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\}$ and the resulting list of $r-1$ vectors would still span V contrary to the definition of r. This proves the theorem. ■

3.3 Linear Transformations

In calculus of many variables one studies functions of many variables and what is meant by their derivatives or integrals, etc. The simplest kind of function of many variables is a linear transformation. You have to begin with the simple things if you expect to make sense of the harder things. The following is the definition of a linear transformation.

Definition 3.3.1 *Let V and W be two finite dimensional vector spaces. A function, L which maps V to W is called a linear transformation and written as $L \in \mathcal{L}(V, W)$ if for all scalars α and β, and vectors $\mathbf{v}, \mathbf{w}$,*

$$L(\alpha\mathbf{v}+\beta\mathbf{w}) = \alpha L(\mathbf{v}) + \beta L(\mathbf{w}).$$

An example of a linear transformation is familiar matrix multiplication, familiar if you have had a linear algebra course. Let $A = (a_{ij})$ be an $m \times n$ matrix. Then an example of a linear transformation $L : \mathbb{F}^n \to \mathbb{F}^m$ is given by

$$(L\mathbf{v})_i \equiv \sum_{j=1}^{n} a_{ij} v_j.$$

Here

$$\mathbf{v} \equiv \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{F}^n.$$

In the general case, the space of linear transformations is itself a vector space. This will be discussed next.

Definition 3.3.2 *Given $L, M \in \mathcal{L}(V, W)$ define a new element of $\mathcal{L}(V, W)$, denoted by $L + M$ according to the rule*

$$(L + M)\mathbf{v} \equiv L\mathbf{v} + M\mathbf{v}.$$

For α a scalar and $L \in \mathcal{L}(V, W)$, define $\alpha L \in \mathcal{L}(V, W)$ by

$$\alpha L(\mathbf{v}) \equiv \alpha(L\mathbf{v}).$$

You should verify that all the axioms of a vector space hold for $\mathcal{L}(V,W)$ with the above definitions of vector addition and scalar multiplication. What about the dimension of $\mathcal{L}(V,W)$?

Before answering this question, here is a lemma.

Lemma 3.3.3 *Let V and W be vector spaces and suppose $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ is a basis for V. Then if $L: V \to W$ is given by $L\mathbf{v}_k = \mathbf{w}_k \in W$ and*

$$L\left(\sum_{k=1}^{n} a_k\mathbf{v}_k\right) \equiv \sum_{k=1}^{n} a_k L\mathbf{v}_k = \sum_{k=1}^{n} a_k\mathbf{w}_k$$

then L is well defined and is in $\mathcal{L}(V,W)$. Also, if L, M are two linear transformations such that $L\mathbf{v}_k = M\mathbf{v}_k$ for all k, then $M = L$.

Proof: L is well defined on V because, since $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ is a basis, there is exactly one way to write a given vector of V as a linear combination. Next, observe that L is obviously linear from the definition. If L, M are equal on the basis, then if $\sum_{k=1}^{n} a_k\mathbf{v}_k$ is an arbitrary vector of V,

$$\begin{aligned} L\left(\sum_{k=1}^{n} a_k\mathbf{v}_k\right) &= \sum_{k=1}^{n} a_k L\mathbf{v}_k \\ &= \sum_{k=1}^{n} a_k M\mathbf{v}_k = M\left(\sum_{k=1}^{n} a_k\mathbf{v}_k\right) \end{aligned}$$

and so $L = M$ because they give the same result for every vector in V.

The message is that when you define a linear transformation, it suffices to tell what it does to a basis.

Definition 3.3.4 *The symbol, δ_{ij} is defined as 1 if $i = j$ and 0 if $i \neq j$.*

Theorem 3.3.5 *Let V and W be finite dimensional vector spaces of dimension n and m respectively Then* $\dim(\mathcal{L}(V,W)) = mn$.

Proof: Let two sets of bases be

$$\{\mathbf{v}_1,\cdots,\mathbf{v}_n\} \text{ and } \{\mathbf{w}_1,\cdots,\mathbf{w}_m\}$$

for V and W respectively. Using Lemma 3.3.3, let $\mathbf{w}_i\mathbf{v}_j \in \mathcal{L}(V,W)$ be the linear transformation defined on the basis, $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$, by

$$\mathbf{w}_i\mathbf{v}_k(\mathbf{v}_j) \equiv \mathbf{w}_i\delta_{jk}.$$

Note that to define these special linear transformations, sometimes called dyadics, it is necessary that $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ be a basis since their definition requires giving the values of the linear transformation on a basis.

Let $L \in \mathcal{L}(V,W)$. Since $\{\mathbf{w}_1,\cdots,\mathbf{w}_m\}$ is a basis, there exist constants d_{jr} such that

$$L\mathbf{v}_r = \sum_{j=1}^{m} d_{jr}\mathbf{w}_j$$

Then from the above,

$$L\mathbf{v}_r = \sum_{j=1}^{m} d_{jr}\mathbf{w}_j = \sum_{j=1}^{m}\sum_{k=1}^{n} d_{jr}\delta_{kr}\mathbf{w}_j = \sum_{j=1}^{m}\sum_{k=1}^{n} d_{jr}\mathbf{w}_j\mathbf{v}_k(\mathbf{v}_r)$$

which shows

$$L = \sum_{j=1}^{m} \sum_{k=1}^{n} d_{jk} \mathbf{w}_j \mathbf{v}_k$$

because the two linear transformations agree on a basis. Since L is arbitrary this shows

$$\{\mathbf{w}_i \mathbf{v}_k : i = 1, \cdots, m,\ k = 1, \cdots, n\}$$

spans $\mathcal{L}(V, W)$.

If

$$\sum_{i,k} d_{ik} \mathbf{w}_i \mathbf{v}_k = \mathbf{0},$$

then

$$\mathbf{0} = \sum_{i,k} d_{ik} \mathbf{w}_i \mathbf{v}_k (\mathbf{v}_l) = \sum_{i=1}^{m} d_{il} \mathbf{w}_i$$

and so, since $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ is a basis, $d_{il} = 0$ for each $i = 1, \cdots, m$. Since l is arbitrary, this shows $d_{il} = 0$ for all i and l. Thus these linear transformations form a basis and this shows the dimension of $\mathcal{L}(V, W)$ is mn as claimed.

Definition 3.3.6 *Let V, W be finite dimensional vector spaces such that a basis for V is*

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$$

and a basis for W is

$$\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}.$$

Then as explained in Theorem 3.3.5, for $L \in \mathcal{L}(V, W)$, there exist scalars l_{ij} such that

$$L = \sum_{ij} l_{ij} \mathbf{w}_i \mathbf{v}_j$$

Consider a rectangular array of scalars such that the entry in the i^{th} row and the j^{th} column is l_{ij},

$$\begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1n} \\ l_{21} & l_{22} & \cdots & l_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ l_{m1} & l_{m2} & \cdots & l_{mn} \end{pmatrix}$$

This is called the matrix of the linear transformation with respect to the two bases. This will typically be denoted by (l_{ij}). It is called a matrix and in this case the matrix is $m \times n$ because it has m rows and n columns.

Theorem 3.3.7 *Let $L \in \mathcal{L}(V, W)$ and let (l_{ij}) be the matrix of L with respect to the two bases,*

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_n\} \text{ and } \{\mathbf{w}_1, \cdots, \mathbf{w}_m\}.$$

of V and W respectively. Then for $\mathbf{v} \in V$ having components $(x_1, \cdots, x_n)$ with respect to the basis $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$, the components of $L\mathbf{v}$ with respect to the basis $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ are

$$\left(\sum_j l_{1j} x_j, \cdots, \sum_j l_{mj} x_j \right)$$

Proof: From the definition of l_{ij},

$$\begin{aligned} L\mathbf{v} &= \sum_{ij} l_{ij}\mathbf{w}_i\mathbf{v}_j(\mathbf{v}) = \sum_{ij} l_{ij}\mathbf{w}_i\mathbf{v}_j\left(\sum_k x_k\mathbf{v}_k\right) \\ &= \sum_{ijk} l_{ij}\mathbf{w}_i\mathbf{v}_j(\mathbf{v}_k)x_k = \sum_{ijk} l_{ij}\mathbf{w}_i\delta_{jk}x_k = \sum_i\left(\sum_j l_{ij}x_j\right)\mathbf{w}_i \end{aligned}$$

and This proves the theorem. ■

Theorem 3.3.8 *Let $(V,\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}),(U,\{\mathbf{u}_1,\cdots,\mathbf{u}_m\}),(W,\{\mathbf{w}_1,\cdots,\mathbf{w}_p\})$ be three vector spaces along with bases for each one. Let $L\in\mathcal{L}(V,U)$ and $M\in\mathcal{L}(U,W)$. Then $ML\in\mathcal{L}(V,W)$ and if (c_{ij}) is the matrix of ML with respect to $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ and $\{\mathbf{w}_1,\cdots,\mathbf{w}_p\}$ and (l_{ij}) and (m_{ij}) are the matrices of L and M respectively with respect to the given bases, then*

$$c_{rj} = \sum_{s=1}^{m} m_{rs}l_{sj}.$$

Proof: First note that from the definition,

$$(\mathbf{w}_i\mathbf{u}_j)(\mathbf{u}_k\mathbf{v}_l)(\mathbf{v}_r) = (\mathbf{w}_i\mathbf{u}_j)\mathbf{u}_k\delta_{lr} = \mathbf{w}_i\delta_{jk}\delta_{lr}$$

and

$$\mathbf{w}_i\mathbf{v}_l\delta_{jk}(\mathbf{v}_r) = \mathbf{w}_i\delta_{jk}\delta_{lr}$$

which shows

$$(\mathbf{w}_i\mathbf{u}_j)(\mathbf{u}_k\mathbf{v}_l) = \mathbf{w}_i\mathbf{v}_l\delta_{jk} \tag{3.13}$$

Therefore,

$$\begin{aligned} ML &= \left(\sum_{rs} m_{rs}\mathbf{w}_r\mathbf{u}_s\right)\left(\sum_{ij} l_{ij}\mathbf{u}_i\mathbf{v}_j\right) \\ &= \sum_{rsij} m_{rs}l_{ij}(\mathbf{w}_r\mathbf{u}_s)(\mathbf{u}_i\mathbf{v}_j) = \sum_{rsij} m_{rs}l_{ij}\mathbf{w}_r\mathbf{v}_j\delta_{is} \\ &= \sum_{rsj} m_{rs}l_{sj}\mathbf{w}_r\mathbf{v}_j = \sum_{rj}\left(\sum_s m_{rs}l_{sj}\right)\mathbf{w}_r\mathbf{v}_j \end{aligned}$$

and This proves the theorem. ■

The relation 3.13 is a very important cancellation property which is used later as well as in this theorem.

Theorem 3.3.9 *Suppose $(V,\{\mathbf{v}_1,\cdots,\mathbf{v}_n\})$ is a vector space and a basis and $(V,\{\mathbf{v}'_1,\cdots,\mathbf{v}'_n\})$ is the same vector space with a different basis. Suppose $L\in\mathcal{L}(V,V)$. Let (l_{ij}) be the matrix of L taken with respect to $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ and let $\left(l'_{ij}\right)$ be the $n\times n$ matrix of L taken with respect to $\{\mathbf{v}'_1,\cdots,\mathbf{v}'_n\}$ That is,*

$$L = \sum_{ij} l_{ij}\mathbf{v}_i\mathbf{v}_j,\ L = \sum_{rs} l'_{rs}\mathbf{v}'_r\mathbf{v}'_s.$$

Then there exist $n\times n$ matrices (d_{ij}) and $\left(d'_{ij}\right)$ satisfying

$$\sum_j d_{ij}d'_{jk} = \delta_{ik}$$

such that

$$l_{ij} = \sum_{rs} d_{ir} l'_{rs} d'_{sj}$$

Proof: First consider the identity map, id defined by $\mathrm{id}(\mathbf{v}) = \mathbf{v}$ with respect to the two bases, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and $\{\mathbf{v}'_1, \cdots, \mathbf{v}'_n\}$.

$$\mathrm{id} = \sum_{tu} d'_{tu} \mathbf{v}'_t \mathbf{v}_u, \ \mathrm{id} = \sum_{ij} d_{ij} \mathbf{v}_i \mathbf{v}'_j \tag{3.14}$$

Now it follows from 3.13

$$\begin{aligned} \mathrm{id} &= \mathrm{id} \circ \mathrm{id} = \sum_{tuij} d'_{tu} d_{ij} \left(\mathbf{v}'_t \mathbf{v}_u\right) \left(\mathbf{v}_i \mathbf{v}'_j\right) = \sum_{tuij} d'_{tu} d_{ij} \delta_{iu} \mathbf{v}'_t \mathbf{v}'_j \\ &= \sum_{tij} d'_{ti} d_{ij} \mathbf{v}'_t \mathbf{v}'_j = \sum_{tj} \left(\sum_i d'_{ti} d_{ij} \right) \mathbf{v}'_t \mathbf{v}'_j \end{aligned}$$

On the other hand,

$$\mathrm{id} = \sum_{tj} \delta_{tj} \mathbf{v}'_t \mathbf{v}'_j$$

because $\mathrm{id}(\mathbf{v}'_k) = \mathbf{v}'_k$ and

$$\sum_{tj} \delta_{tj} \mathbf{v}'_t \mathbf{v}'_j \left(\mathbf{v}'_k\right) = \sum_{tj} \delta_{tj} \mathbf{v}'_t \delta_{jk} = \sum_t \delta_{tk} \mathbf{v}'_t = \mathbf{v}'_k.$$

Therefore,

$$\left(\sum_i d'_{ti} d_{ij} \right) = \delta_{tj}.$$

Switching the order of the above products shows

$$\left(\sum_i d_{ti} d'_{ij} \right) = \delta_{tj}$$

In terms of matrices, this says $\left(d'_{ij}\right)$ is the inverse matrix of (d_{ij}).

Now using 3.14 and the cancellation property 3.13,

$$\begin{aligned} L &= \sum_{iu} l_{iu} \mathbf{v}_i \mathbf{v}_u = \sum_{rs} l'_{rs} \mathbf{v}'_r \mathbf{v}'_s = \mathrm{id} \sum_{rs} l'_{rs} \mathbf{v}'_r \mathbf{v}'_s \, \mathrm{id} \\ &= \sum_{ij} d_{ij} \mathbf{v}_i \mathbf{v}'_j \sum_{rs} l'_{rs} \mathbf{v}'_r \mathbf{v}'_s \sum_{tu} d'_{tu} \mathbf{v}'_t \mathbf{v}_u \\ &= \sum_{ijturs} d_{ij} l'_{rs} d'_{tu} \left(\mathbf{v}_i \mathbf{v}'_j\right) \left(\mathbf{v}'_r \mathbf{v}'_s\right) \left(\mathbf{v}'_t \mathbf{v}_u\right) \\ &= \sum_{ijturs} d_{ij} l'_{rs} d'_{tu} \mathbf{v}_i \mathbf{v}_u \delta_{jr} \delta_{st} = \sum_{iu} \left(\sum_{js} d_{ij} l'_{js} d'_{su} \right) \mathbf{v}_i \mathbf{v}_u \end{aligned}$$

and since the linear transformations, $\{\mathbf{v}_i \mathbf{v}_u\}$ are linearly independent, this shows

$$l_{iu} = \sum_{js} d_{ij} l'_{js} d'_{su}$$

as claimed. This proves the theorem. ■

Recall the following definition which is a review of important terminology about matrices.

Definition 3.3.10 *If A is an $m \times n$ matrix and B is an $n \times p$ matrix, $A = (A_{ij}), B = (B_{ij})$, then if $(AB)_{ij}$ is the ij^{th} entry of the product, then*

$$(AB)_{ij} = \sum_k A_{ik} B_{kj}$$

An $n \times n$ matrix, A is said to be invertible if there exists another $n \times n$ matrix, denoted by A^{-1} such that $AA^{-1} = A^{-1}A = I$ where the ij^{th} entry of I is δ_{ij}. Recall also that $\left(A^T\right)_{ij} \equiv A_{ji}$. This is called the transpose of A.

Theorem 3.3.11 *The following are important properties of matrices.*

1. $IA = AI = A$
2. $(AB)C = A(BC)$
3. A^{-1} *is unique if it exists.*
4. *When the inverses exist,* $(AB)^{-1} = B^{-1}A^{-1}$
5. $(AB)^T = B^T A^T$

Proof: I will prove these things directly from the above definition but there are more elegant ways to see these things in terms of composition of linear transformations which is really what matrix multiplication corresponds to.

First, $(IA)_{ij} \equiv \sum_k \delta_{ik} A_{kj} = A_{ij}$. The other order is similar.

Next consider the associative law of multiplication.

$$\begin{aligned} ((AB)C)_{ij} &\equiv \sum_k (AB)_{ik} C_{kj} = \sum_k \sum_r A_{ir} B_{rk} C_{kj} \\ &= \sum_r A_{ir} \sum_k B_{rk} C_{kj} = \sum_r A_{ir} (BC)_{rj} = (A(BC))_{ij} \end{aligned}$$

Since the ij^{th} entries are equal, the two matrices are equal.

Next consider the uniqueness of the inverse. If $AB = BA = I$, then using the associative law,

$$B = IB = \left(A^{-1}A\right)B = A^{-1}(AB) = A^{-1}I = A^{-1}$$

Thus if it acts like the inverse, it is the inverse.

Consider now the inverse of a product.

$$AB\left(B^{-1}A^{-1}\right) = A\left(BB^{-1}\right)A^{-1} = AIA^{-1} = I$$

Similarly, $\left(B^{-1}A^{-1}\right)AB = I$. Hence from what was just shown, $(AB)^{-1}$ exists and equals $B^{-1}A^{-1}$.

Finally consider the statement about transposes.

$$\left((AB)^T\right)_{ij} \equiv (AB)_{ji} \equiv \sum_k A_{jk} B_{ki} \equiv \sum_k \left(B^T\right)_{ik} \left(A^T\right)_{kj} \equiv \left(B^T A^T\right)_{ij}$$

Since the ij^{th} entries are the same, the two matrices are equal. This proves the theorem. ■

In terms of matrix multiplication, Theorem 3.3.9 says that if M_1 and M_2 are matrices for the same linear transformation relative to two different bases, it follows there exists an invertible matrix, S such that

$$M_1 = S^{-1} M_2 S$$

This is called a similarity transformation and is important in linear algebra but this is as far as the theory will be developed here.

3.4 Block Multiplication Of Matrices

Consider the following problem

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix}$$

You know how to do this from the above definition of matrix mutiplication. You get

$$\begin{pmatrix} AE+BG & AF+BH \\ CE+DG & CF+DH \end{pmatrix}.$$

Now what if instead of numbers, the entries, A, B, C, D, E, F, G are matrices of a size such that the multiplications and additions needed in the above formula all make sense. Would the formula be true in this case? I will show below that this is true.

Suppose A is a matrix of the form

$$\begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{r1} & \cdots & A_{rm} \end{pmatrix} \tag{3.15}$$

where A_{ij} is a $s_i \times p_j$ matrix where s_i does not depend on j and p_j does not depend on i. Such a matrix is called a **block matrix,** also a **partitioned matrix**. Let $n = \sum_j p_j$ and $k = \sum_i s_i$ so A is an $k \times n$ matrix. What is $A\mathbf{x}$ where $\mathbf{x} \in \mathbb{F}^n$? From the process of multiplying a matrix times a vector, the following lemma follows.

Lemma 3.4.1 *Let A be an $m \times n$ block matrix as in 3.15 and let $\mathbf{x} \in \mathbb{F}^n$. Then $A\mathbf{x}$ is of the form*

$$A\mathbf{x} = \begin{pmatrix} \sum_j A_{1j}\mathbf{x}_j \\ \vdots \\ \sum_j A_{rj}\mathbf{x}_j \end{pmatrix}$$

where $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_m)^T$ and $\mathbf{x}_i \in \mathbb{F}^{p_i}$.

Suppose also that B is a block matrix of the form

$$\begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{r1} & \cdots & B_{rp} \end{pmatrix} \tag{3.16}$$

and A is a block matrix of the form

$$\begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{p1} & \cdots & A_{pm} \end{pmatrix} \tag{3.17}$$

and that for all i, j, it makes sense to multiply $B_{is}A_{sj}$ for all $s \in \{1, \cdots, m\}$ and that for each $s, B_{is}A_{sj}$ is the same size so that it makes sense to write $\sum_s B_{is}A_{sj}$.

Theorem 3.4.2 *Let B be a block matrix as in 3.16 and let A be a block matrix as in 3.17 such that B_{is} is conformable with A_{sj} and each product, $B_{is}A_{sj}$ is of the same size so they can be added. Then BA is a block matrix such that the ij^{th} block is of the form*

$$\sum_s B_{is}A_{sj}. \tag{3.18}$$

Proof: Let B_{is} be a $q_i \times p_s$ matrix and A_{sj} be a $p_s \times r_j$ matrix. Also let $\mathbf{x} \in \mathbb{F}^n$ and let $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_m)^T$ and $\mathbf{x}_i \in \mathbb{F}^{r_i}$ so it makes sense to multiply $A_{sj}\mathbf{x}_j$. Then from the associative law of matrix multiplication and Lemma 3.4.1 applied twice,

$$\left(\begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{r1} & \cdots & B_{rp} \end{pmatrix}\begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{p1} & \cdots & A_{pm} \end{pmatrix}\right)\begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \end{pmatrix}$$

$$= \begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{r1} & \cdots & B_{rp} \end{pmatrix}\begin{pmatrix} \sum_j A_{1j}\mathbf{x}_j \\ \vdots \\ \sum_j A_{rj}\mathbf{x}_j \end{pmatrix}$$

$$= \begin{pmatrix} \sum_s \sum_j B_{1s}A_{sj}\mathbf{x}_j \\ \vdots \\ \sum_s \sum_j B_{rs}A_{sj}\mathbf{x}_j \end{pmatrix} = \begin{pmatrix} \sum_j \left(\sum_s B_{1s}A_{sj}\right)\mathbf{x}_j \\ \vdots \\ \sum_j \left(\sum_s B_{rs}A_{sj}\right)\mathbf{x}_j \end{pmatrix}$$

$$= \begin{pmatrix} \sum_s B_{1s}A_{s1} & \cdots & \sum_s B_{1s}A_{sm} \\ \vdots & \ddots & \vdots \\ \sum_s B_{rs}A_{s1} & \cdots & \sum_s B_{rs}A_{sm} \end{pmatrix}\begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \end{pmatrix}$$

By Lemma 3.4.1, this shows that $(BA)\mathbf{x}$ equals the block matrix whose ij^{th} entry is given by 3.18 times $\mathbf{x}$. Since $\mathbf{x}$ is an arbitrary vector in $\mathbb{F}^n$, This proves the theorem. ■

The message of this theorem is that you can formally multiply block matrices as though the blocks were numbers. You just have to pay attention to the preservation of order.

3.5 Determinants

3.5.1 The Determinant Of A Matrix

The following Lemma will be essential in the definition of the determinant.

Lemma 3.5.1 *There exists a unique function,* sgn_n *which maps each list of numbers from* $\{1, \cdots, n\}$ *to one of the three numbers,* $0, 1,$ *or* -1 *which also has the following properties.*

$$\operatorname{sgn}_n(1, \cdots, n) = 1 \tag{3.19}$$

$$\operatorname{sgn}_n(i_1, \cdots, p, \cdots, q, \cdots, i_n) = -\operatorname{sgn}_n(i_1, \cdots, q, \cdots, p, \cdots, i_n) \tag{3.20}$$

In words, the second property states that if two of the numbers are switched, the value of the function is multiplied by -1*. Also, in the case where* $n > 1$ *and* $\{i_1, \cdots, i_n\} = \{1, \cdots, n\}$ *so that every number from* $\{1, \cdots, n\}$ *appears in the ordered list,* $(i_1, \cdots, i_n)$*,*

$$\operatorname{sgn}_n(i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n) \equiv$$

$$(-1)^{n-\theta}\operatorname{sgn}_{n-1}(i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n) \tag{3.21}$$

where $n = i_\theta$ *in the ordered list,* $(i_1, \cdots, i_n)$*.*

Proof: To begin with, it is necessary to show the existence of such a function. This is clearly true if $n = 1$. Define $\operatorname{sgn}_1(1) \equiv 1$ and observe that it works. No switching is possible. In the case where $n = 2$, it is also clearly true. Let $\operatorname{sgn}_2(1, 2) = 1$ and $\operatorname{sgn}_2(2, 1) = -1$ while $\operatorname{sgn}_2(2, 2) = \operatorname{sgn}_2(1, 1) = 0$ and verify it works. Assuming such a function exists for n, sgn_{n+1} will be defined in terms of sgn_n. If there are any repeated

numbers in $(i_1, \cdots, i_{n+1})$, $\operatorname{sgn}_{n+1}(i_1, \cdots, i_{n+1}) \equiv 0$. If there are no repeats, then $n+1$ appears somewhere in the ordered list. Let θ be the position of the number $n+1$ in the list. Thus, the list is of the form $(i_1, \cdots, i_{\theta-1}, n+1, i_{\theta+1}, \cdots, i_{n+1})$. From 3.21 it must be that

$$\operatorname{sgn}_{n+1}(i_1, \cdots, i_{\theta-1}, n+1, i_{\theta+1}, \cdots, i_{n+1}) \equiv$$
$$(-1)^{n+1-\theta} \operatorname{sgn}_n(i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_{n+1}).$$

It is necessary to verify this satisfies 3.19 and 3.20 with n replaced with $n+1$. The first of these is obviously true because

$$\operatorname{sgn}_{n+1}(1, \cdots, n, n+1) \equiv (-1)^{n+1-(n+1)} \operatorname{sgn}_n(1, \cdots, n) = 1.$$

If there are repeated numbers in $(i_1, \cdots, i_{n+1})$, then it is obvious 3.20 holds because both sides would equal zero from the above definition. It remains to verify 3.20 in the case where there are no numbers repeated in $(i_1, \cdots, i_{n+1})$. Consider

$$\operatorname{sgn}_{n+1}\left(i_1, \cdots, \overset{r}{p}, \cdots, \overset{s}{q}, \cdots, i_{n+1}\right),$$

where the r above the p indicates the number, p is in the r^{th} position and the s above the q indicates that the number, q is in the s^{th} position. Suppose first that $r < \theta < s$. Then

$$\operatorname{sgn}_{n+1}\left(i_1, \cdots, \overset{r}{p}, \cdots, \overset{\theta}{n+1}, \cdots, \overset{s}{q}, \cdots, i_{n+1}\right) \equiv$$
$$(-1)^{n+1-\theta} \operatorname{sgn}_n\left(i_1, \cdots, \overset{r}{p}, \cdots, \overset{s-1}{q}, \cdots, i_{n+1}\right)$$

while

$$\operatorname{sgn}_{n+1}\left(i_1, \cdots, \overset{r}{q}, \cdots, \overset{\theta}{n+1}, \cdots, \overset{s}{p}, \cdots, i_{n+1}\right) =$$
$$(-1)^{n+1-\theta} \operatorname{sgn}_n\left(i_1, \cdots, \overset{r}{q}, \cdots, \overset{s-1}{p}, \cdots, i_{n+1}\right)$$

and so, by induction, a switch of p and q introduces a minus sign in the result. Similarly, if $\theta > s$ or if $\theta < r$ it also follows that 3.20 holds. The interesting case is when $\theta = r$ or $\theta = s$. Consider the case where $\theta = r$ and note the other case is entirely similar.

$$\operatorname{sgn}_{n+1}\left(i_1, \cdots, \overset{r}{n+1}, \cdots, \overset{s}{q}, \cdots, i_{n+1}\right) =$$
$$(-1)^{n+1-r} \operatorname{sgn}_n\left(i_1, \cdots, \overset{s-1}{q}, \cdots, i_{n+1}\right) \tag{3.22}$$

while

$$\operatorname{sgn}_{n+1}\left(i_1, \cdots, \overset{r}{q}, \cdots, \overset{s}{n+1}, \cdots, i_{n+1}\right) =$$
$$(-1)^{n+1-s} \operatorname{sgn}_n\left(i_1, \cdots, \overset{r}{q}, \cdots, i_{n+1}\right). \tag{3.23}$$

By making $s-1-r$ switches, move the q which is in the $s-1^{th}$ position in 3.22 to the r^{th} position in 3.23. By induction, each of these switches introduces a factor of -1 and so

$$\operatorname{sgn}_n\left(i_1, \cdots, \overset{s-1}{q}, \cdots, i_{n+1}\right) = (-1)^{s-1-r} \operatorname{sgn}_n\left(i_1, \cdots, \overset{r}{q}, \cdots, i_{n+1}\right).$$

Therefore,

$$\operatorname{sgn}_{n+1}\left(i_1, \cdots, \overset{r}{n+1}, \cdots, \overset{s}{q}, \cdots, i_{n+1}\right) = (-1)^{n+1-r} \operatorname{sgn}_n\left(i_1, \cdots, \overset{s-1}{q}, \cdots, i_{n+1}\right)$$

$$= (-1)^{n+1-r}(-1)^{s-1-r}\operatorname{sgn}_n\left(i_1,\cdots,\overset{r}{q},\cdots,i_{n+1}\right)$$

$$= (-1)^{n+s}\operatorname{sgn}_n\left(i_1,\cdots,\overset{r}{q},\cdots,i_{n+1}\right) = (-1)^{2s-1}(-1)^{n+1-s}\operatorname{sgn}_n\left(i_1,\cdots,\overset{r}{q},\cdots,i_{n+1}\right)$$

$$= -\operatorname{sgn}_{n+1}\left(i_1,\cdots,\overset{r}{q},\cdots,\overset{s}{n+1},\cdots,i_{n+1}\right).$$

This proves the existence of the desired function.

To see this function is unique, note that you can obtain any ordered list of distinct numbers from a sequence of switches. If there exist two functions, f and g both satisfying 3.19 and 3.20, you could start with $f(1,\cdots,n) = g(1,\cdots,n)$ and applying the same sequence of switches, eventually arrive at $f(i_1,\cdots,i_n) = g(i_1,\cdots,i_n)$. If any numbers are repeated, then 3.20 gives both functions are equal to zero for that ordered list. This proves the lemma. ■

In what follows sgn will often be used rather than sgn_n because the context supplies the appropriate n.

Definition 3.5.2 *Let f be a real valued function which has the set of ordered lists of numbers from $\{1,\cdots,n\}$ as its domain. Define*

$$\sum_{(k_1,\cdots,k_n)} f(k_1\cdots k_n)$$

to be the sum of all the $f(k_1\cdots k_n)$ for all possible choices of ordered lists $(k_1,\cdots,k_n)$ of numbers of $\{1,\cdots,n\}$. For example,

$$\sum_{(k_1,k_2)} f(k_1,k_2) = f(1,2) + f(2,1) + f(1,1) + f(2,2).$$

Definition 3.5.3 *Let $(a_{ij}) = A$ denote an $n\times n$ matrix. The determinant of A, denoted by $\det(A)$ is defined by*

$$\det(A) \equiv \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{1k_1}\cdots a_{nk_n}$$

where the sum is taken over all ordered lists of numbers from $\{1,\cdots,n\}$. Note it suffices to take the sum over only those ordered lists in which there are no repeats because if there are, $\operatorname{sgn}(k_1,\cdots,k_n) = 0$ and so that term contributes 0 to the sum.

Let A be an $n\times n$ matrix, $A = (a_{ij})$ and let $(r_1,\cdots,r_n)$ denote an ordered list of n numbers from $\{1,\cdots,n\}$. Let $A(r_1,\cdots,r_n)$ denote the matrix whose k^{th} row is the r_k row of the matrix, A. Thus

$$\det(A(r_1,\cdots,r_n)) = \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n} \tag{3.24}$$

and

$$A(1,\cdots,n) = A.$$

Proposition 3.5.4 *Let*

$$(r_1,\cdots,r_n)$$

be an ordered list of numbers from $\{1,\cdots,n\}$. Then

$$\operatorname{sgn}(r_1,\cdots,r_n)\det(A)$$

$$= \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n} \tag{3.25}$$

$$= \det(A(r_1,\cdots,r_n)). \tag{3.26}$$

Proof: Let $(1,\cdots,n) = (1,\cdots,r,\cdots s,\cdots,n)$ so $r<s$.

$$\det(A(1,\cdots,r,\cdots,s,\cdots,n)) = \tag{3.27}$$

$$\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_r,\cdots,k_s,\cdots,k_n)\, a_{1k_1}\cdots a_{rk_r}\cdots a_{sk_s}\cdots a_{nk_n},$$

and renaming the variables, calling k_s, k_r and k_r, k_s, this equals

$$= \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_s,\cdots,k_r,\cdots,k_n)\, a_{1k_1}\cdots a_{rk_s}\cdots a_{sk_r}\cdots a_{nk_n}$$

$$= \sum_{(k_1,\cdots,k_n)} -\operatorname{sgn}\left(k_1,\cdots,\overbrace{k_r,\cdots,k_s}^{\text{These got switched}},\cdots,k_n\right) a_{1k_1}\cdots a_{sk_r}\cdots a_{rk_s}\cdots a_{nk_n}$$

$$= -\det(A(1,\cdots,s,\cdots,r,\cdots,n)). \tag{3.28}$$

Consequently,

$$\det(A(1,\cdots,s,\cdots,r,\cdots,n)) =$$

$$-\det(A(1,\cdots,r,\cdots,s,\cdots,n)) = -\det(A)$$

Now letting $A(1,\cdots,s,\cdots,r,\cdots,n)$ play the role of A, and continuing in this way, switching pairs of numbers,

$$\det(A(r_1,\cdots,r_n)) = (-1)^p \det(A)$$

where it took p switches to obtain$(r_1,\cdots,r_n)$ from $(1,\cdots,n)$. By Lemma 3.5.1, this implies

$$\det(A(r_1,\cdots,r_n)) = (-1)^p \det(A) = \operatorname{sgn}(r_1,\cdots,r_n)\det(A)$$

and proves the proposition in the case when there are no repeated numbers in the ordered list, $(r_1,\cdots,r_n)$. However, if there is a repeat, say the r^{th} row equals the s^{th} row, then the reasoning of 3.27 -3.28 shows that $A(r_1,\cdots,r_n) = 0$ and also $\operatorname{sgn}(r_1,\cdots,r_n) = 0$ so the formula holds in this case also.

Observation 3.5.5 *There are $n!$ ordered lists of distinct numbers from $\{1,\cdots,n\}$.*

With the above, it is possible to give a more symmetric description of the determinant from which it will follow that $\det(A) = \det(A^T)$.

Corollary 3.5.6 *The following formula for $\det(A)$ is valid.*

$$\det(A) = \frac{1}{n!}\cdot$$

$$\sum_{(r_1,\cdots,r_n)} \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n}. \tag{3.29}$$

And also $\det(A^T) = \det(A)$ where A^T is the transpose of A. (Recall that for $A^T = (a_{ij}^T)$, $a_{ij}^T = a_{ji}$.)

Proof: From Proposition 3.5.4, if the r_i are distinct,

$$\det(A) = \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n}.$$

Summing over all ordered lists, $(r_1, \cdots, r_n)$ where the r_i are distinct, (If the r_i are not distinct, $\operatorname{sgn}(r_1, \cdots, r_n) = 0$ and so there is no contribution to the sum.)

$$n! \det(A) =$$

$$\sum_{(r_1,\cdots,r_n)} \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(r_1, \cdots, r_n) \operatorname{sgn}(k_1, \cdots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n}.$$

This proves the corollary. ■ since the formula gives the same number for A as it does for A^T.

Corollary 3.5.7 *If two rows or two columns in an $n \times n$ matrix, A, are switched, the determinant of the resulting matrix equals (-1) times the determinant of the original matrix. If A is an $n \times n$ matrix in which two rows are equal or two columns are equal then* $\det(A) = 0$. *Suppose the i^{th} row of A equals $(xa_1 + yb_1, \cdots, xa_n + yb_n)$. Then*

$$\det(A) = x \det(A_1) + y \det(A_2)$$

where the i^{th} row of A_1 is $(a_1, \cdots, a_n)$ and the i^{th} row of A_2 is $(b_1, \cdots, b_n)$, all other rows of A_1 and A_2 coinciding with those of A. In other words, det *is a linear function of each row A. The same is true with the word "row" replaced with the word "column".*

Proof: By Proposition 3.5.4 when two rows are switched, the determinant of the resulting matrix is (-1) times the determinant of the original matrix. By Corollary 3.5.6 the same holds for columns because the columns of the matrix equal the rows of the transposed matrix. Thus if A_1 is the matrix obtained from A by switching two columns,

$$\det(A) = \det\left(A^T\right) = -\det\left(A_1^T\right) = -\det(A_1).$$

If A has two equal columns or two equal rows, then switching them results in the same matrix. Therefore, $\det(A) = -\det(A)$ and so $\det(A) = 0$.

It remains to verify the last assertion.

$$\begin{aligned}
\det(A) &\equiv \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1, \cdots, k_n) a_{1k_1} \cdots (xa_{k_i} + yb_{k_i}) \cdots a_{nk_n} \\
&= x \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1, \cdots, k_n) a_{1k_1} \cdots a_{k_i} \cdots a_{nk_n} \\
&\quad + y \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1, \cdots, k_n) a_{1k_1} \cdots b_{k_i} \cdots a_{nk_n} \\
&\equiv x \det(A_1) + y \det(A_2).
\end{aligned}$$

The same is true of columns because $\det\left(A^T\right) = \det(A)$ and the rows of A^T are the columns of A.

The following corollary is also of great use.

Corollary 3.5.8 *Suppose A is an $n \times n$ matrix and some column (row) is a linear combination of r other columns (rows). Then* $\det(A) = 0$.

Proof: Let $A = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix}$ be the columns of A and suppose the condition that one column is a linear combination of r of the others is satisfied. Then by using Corollary 3.5.7 you may rearrange the columns to have the n^{th} column a linear combination of the first r columns. Thus $\mathbf{a}_n = \sum_{k=1}^{r} c_k \mathbf{a}_k$ and so

$$\det(A) = \det\begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_r & \cdots & \mathbf{a}_{n-1} & \sum_{k=1}^{r} c_k \mathbf{a}_k \end{pmatrix}.$$

By Corollary 3.5.7

$$\det(A) = \sum_{k=1}^{r} c_k \det\begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_r & \cdots & \mathbf{a}_{n-1} & \mathbf{a}_k \end{pmatrix} = 0.$$

The case for rows follows from the fact that $\det(A) = \det\left(A^T\right)$. This proves the corollary. ■

Recall the following definition of matrix multiplication.

Definition 3.5.9 *If A and B are $n \times n$ matrices, $A = (a_{ij})$ and $B = (b_{ij})$, $AB = (c_{ij})$ where*

$$c_{ij} \equiv \sum_{k=1}^{n} a_{ik} b_{kj}.$$

One of the most important rules about determinants is that the determinant of a product equals the product of the determinants.

Theorem 3.5.10 *Let A and B be $n \times n$ matrices. Then*

$$\det(AB) = \det(A)\det(B).$$

Proof: Let c_{ij} be the ij^{th} entry of AB. Then by Proposition 3.5.4,

$$\det(AB) =$$

$$\begin{aligned}
&\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, c_{1k_1}\cdots c_{nk_n} \\
= &\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n) \left(\sum_{r_1} a_{1r_1} b_{r_1k_1}\right)\cdots\left(\sum_{r_n} a_{nr_n} b_{r_nk_n}\right) \\
= &\sum_{(r_1\cdots,r_n)} \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, b_{r_1k_1}\cdots b_{r_nk_n}\left(a_{1r_1}\cdots a_{nr_n}\right) \\
= &\sum_{(r_1\cdots,r_n)} \operatorname{sgn}(r_1\cdots r_n)\, a_{1r_1}\cdots a_{nr_n} \det(B) = \det(A)\det(B).
\end{aligned}$$

This proves the theorem. ■

Lemma 3.5.11 *Suppose a matrix is of the form*

$$M = \begin{pmatrix} A & * \\ \mathbf{0} & a \end{pmatrix} \tag{3.30}$$

or

$$M = \begin{pmatrix} A & \mathbf{0} \\ * & a \end{pmatrix} \tag{3.31}$$

where a is a number and A is an $(n-1)\times(n-1)$ matrix and $*$ denotes either a column or a row having length $n-1$ and the $\mathbf{0}$ denotes either a column or a row of length $n-1$ consisting entirely of zeros. Then

$$\det(M) = a\det(A).$$

Proof: Denote M by (m_{ij}). Thus in the first case, $m_{nn} = a$ and $m_{ni} = 0$ if $i \neq n$ while in the second case, $m_{nn} = a$ and $m_{in} = 0$ if $i \neq n$. From the definition of the determinant,

$$\det(M) \equiv \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}_n(k_1,\cdots,k_n)\, m_{1k_1}\cdots m_{nk_n}$$

Letting θ denote the position of n in the ordered list, $(k_1,\cdots,k_n)$ then using the earlier conventions used to prove Lemma 3.5.1, $\det(M)$ equals

$$\sum_{(k_1,\cdots,k_n)} (-1)^{n-\theta} \operatorname{sgn}_{n-1}\left(k_1,\cdots,k_{\theta-1},\overset{\theta}{k_{\theta+1}},\cdots,\overset{n-1}{k_n}\right) m_{1k_1}\cdots m_{nk_n}$$

Now suppose 3.31. Then if $k_n \neq n$, the term involving m_{nk_n} in the above expression equals zero. Therefore, the only terms which survive are those for which $\theta = n$ or in other words, those for which $k_n = n$. Therefore, the above expression reduces to

$$a \sum_{(k_1,\cdots,k_{n-1})} \operatorname{sgn}_{n-1}(k_1,\cdots k_{n-1})\, m_{1k_1}\cdots m_{(n-1)k_{n-1}} = a\det(A).$$

To get the assertion in the situation of 3.30 use Corollary 3.5.6 and 3.31 to write

$$\det(M) = \det\left(M^T\right) = \det\left(\begin{pmatrix} A^T & \mathbf{0} \\ * & a \end{pmatrix}\right) = a\det\left(A^T\right) = a\det(A).$$

This proves the lemma. ■

In terms of the theory of determinants, arguably the most important idea is that of Laplace expansion along a row or a column. This will follow from the above definition of a determinant.

Definition 3.5.12 *Let $A = (a_{ij})$ be an $n \times n$ matrix. Then a new matrix called the cofactor matrix,* $\operatorname{cof}(A)$ *is defined by* $\operatorname{cof}(A) = (c_{ij})$ *where to obtain c_{ij} delete the i^{th} row and the j^{th} column of A, take the determinant of the $(n-1)\times(n-1)$ matrix which results, (This is called the ij^{th} minor of A.) and then multiply this number by $(-1)^{i+j}$. To make the formulas easier to remember,* $\operatorname{cof}(A)_{ij}$ *will denote the ij^{th} entry of the cofactor matrix.*

The following is the main result.

Theorem 3.5.13 *Let A be an $n \times n$ matrix where $n \geq 2$. Then*

$$\det(A) = \sum_{j=1}^{n} a_{ij}\operatorname{cof}(A)_{ij} = \sum_{i=1}^{n} a_{ij}\operatorname{cof}(A)_{ij}. \tag{3.32}$$

The first formula consists of expanding the determinant along the i^{th} row and the second expands the determinant along the j^{th} column.

Proof: Let $(a_{i1},\cdots,a_{in})$ be the i^{th} row of A. Let B_j be the matrix obtained from A by leaving every row the same except the i^{th} row which in B_j equals $(0,\cdots,0,a_{ij},0,\cdots,0)$. Then by Corollary 3.5.7,

$$\det(A) = \sum_{j=1}^{n} \det(B_j)$$

Denote by A^{ij} the $(n-1)\times(n-1)$ matrix obtained by deleting the i^{th} row and the j^{th} column of A. Thus $\operatorname{cof}(A)_{ij} \equiv (-1)^{i+j}\det\left(A^{ij}\right)$. At this point, recall that from

Proposition 3.5.4, when two rows or two columns in a matrix, M, are switched, this results in multiplying the determinant of the old matrix by -1 to get the determinant of the new matrix. Therefore, by Lemma 3.5.11,

$$\begin{aligned}\det(B_j) &= (-1)^{n-j}(-1)^{n-i}\det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij}\end{pmatrix}\right) \\ &= (-1)^{i+j}\det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij}\end{pmatrix}\right) = a_{ij}\operatorname{cof}(A)_{ij}.\end{aligned}$$

Therefore,

$$\det(A) = \sum_{j=1}^{n} a_{ij}\operatorname{cof}(A)_{ij}$$

which is the formula for expanding $\det(A)$ along the i^{th} row. Also,

$$\begin{aligned}\det(A) &= \det\left(A^T\right) = \sum_{j=1}^{n} a_{ij}^T \operatorname{cof}\left(A^T\right)_{ij} \\ &= \sum_{j=1}^{n} a_{ji}\operatorname{cof}(A)_{ji}\end{aligned}$$

which is the formula for expanding $\det(A)$ along the i^{th} column. This proves the theorem. ■

Note that this gives an easy way to write a formula for the inverse of an $n\times n$ matrix.

Theorem 3.5.14 *A^{-1} exists if and only if $\det(A)\neq 0$. If $\det(A)\neq 0$, then $A^{-1} = \left(a_{ij}^{-1}\right)$ where*

$$a_{ij}^{-1} = \det(A)^{-1}\operatorname{cof}(A)_{ji}$$

for $\operatorname{cof}(A)_{ij}$ the ij^{th} cofactor of A.

Proof: By Theorem 3.5.13 and letting $(a_{ir}) = A$, if $\det(A)\neq 0$,

$$\sum_{i=1}^{n} a_{ir}\operatorname{cof}(A)_{ir}\det(A)^{-1} = \det(A)\det(A)^{-1} = 1.$$

Now consider

$$\sum_{i=1}^{n} a_{ir}\operatorname{cof}(A)_{ik}\det(A)^{-1}$$

when $k\neq r$. Replace the k^{th} column with the r^{th} column to obtain a matrix, B_k whose determinant equals zero by Corollary 3.5.7. However, expanding this matrix along the k^{th} column yields

$$0 = \det(B_k)\det(A)^{-1} = \sum_{i=1}^{n} a_{ir}\operatorname{cof}(A)_{ik}\det(A)^{-1}$$

Summarizing,

$$\sum_{i=1}^{n} a_{ir}\operatorname{cof}(A)_{ik}\det(A)^{-1} = \delta_{rk}.$$

Using the other formula in Theorem 3.5.13, and similar reasoning,

$$\sum_{j=1}^{n} a_{rj}\operatorname{cof}(A)_{kj}\det(A)^{-1} = \delta_{rk}$$

This proves that if $\det(A) \neq 0$, then A^{-1} exists with $A^{-1} = \left(a_{ij}^{-1}\right)$, where

$$a_{ij}^{-1} = \operatorname{cof}(A)_{ji} \det(A)^{-1}.$$

Now suppose A^{-1} exists. Then by Theorem 3.5.10,

$$1 = \det(I) = \det\left(AA^{-1}\right) = \det(A) \det\left(A^{-1}\right)$$

so $\det(A) \neq 0$. This proves the theorem. ■

The next corollary points out that if an $n \times n$ matrix, A has a right or a left inverse, then it has an inverse.

Corollary 3.5.15 *Let A be an $n \times n$ matrix and suppose there exists an $n \times n$ matrix, B such that $BA = I$. Then A^{-1} exists and $A^{-1} = B$. Also, if there exists C an $n \times n$ matrix such that $AC = I$, then A^{-1} exists and $A^{-1} = C$.*

Proof: Since $BA = I$, Theorem 3.5.10 implies

$$\det B \det A = 1$$

and so $\det A \neq 0$. Therefore from Theorem 3.5.14, A^{-1} exists. Therefore,

$$A^{-1} = (BA) A^{-1} = B\left(AA^{-1}\right) = BI = B.$$

The case where $CA = I$ is handled similarly.

The conclusion of this corollary is that left inverses, right inverses and inverses are all the same in the context of $n \times n$ matrices.

Theorem 3.5.14 says that to find the inverse, take the transpose of the cofactor matrix and divide by the determinant. The transpose of the cofactor matrix is called the adjugate or sometimes the classical adjoint of the matrix A. It is an abomination to call it the adjoint although you do sometimes see it referred to in this way. In words, A^{-1} is equal to one over the determinant of A times the adjugate matrix of A.

In case you are solving a system of equations, $A\mathbf{x} = \mathbf{y}$ for $\mathbf{x}$, it follows that if A^{-1} exists,

$$\mathbf{x} = \left(A^{-1}A\right)\mathbf{x} = A^{-1}(A\mathbf{x}) = A^{-1}\mathbf{y}$$

thus solving the system. Now in the case that A^{-1} exists, there is a formula for A^{-1} given above. Using this formula,

$$x_i = \sum_{j=1}^{n} a_{ij}^{-1} y_j = \sum_{j=1}^{n} \frac{1}{\det(A)} \operatorname{cof}(A)_{ji} y_j.$$

By the formula for the expansion of a determinant along a column,

$$x_i = \frac{1}{\det(A)} \det \begin{pmatrix} * & \cdots & y_1 & \cdots & * \\ \vdots & & \vdots & & \vdots \\ * & \cdots & y_n & \cdots & * \end{pmatrix},$$

where here the i^{th} column of A is replaced with the column vector, $(y_1 \cdots, y_n)^T$, and the determinant of this modified matrix is taken and divided by $\det(A)$. This formula is known as Cramer's rule.

Definition 3.5.16 *A matrix M, is upper triangular if $M_{ij} = 0$ whenever $i > j$. Thus such a matrix equals zero below the main diagonal, the entries of the form M_{ii} as shown.*

$$\begin{pmatrix} * & * & \cdots & * \\ 0 & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & * \end{pmatrix}$$

A lower triangular matrix is defined similarly as a matrix for which all entries above the main diagonal are equal to zero.

With this definition, here is a simple corollary of Theorem 3.5.13.

Corollary 3.5.17 *Let M be an upper (lower) triangular matrix. Then $\det(M)$ is obtained by taking the product of the entries on the main diagonal.*

Definition 3.5.18 *A submatrix of a matrix A is the rectangular array of numbers obtained by deleting some rows and columns of A. Let A be an $m \times n$ matrix. The **determinant rank** of the matrix equals r where r is the largest number such that some $r \times r$ submatrix of A has a non zero determinant. The **row rank** is defined to be the dimension of the span of the rows. The **column rank** is defined to be the dimension of the span of the columns.*

Theorem 3.5.19 *If A has determinant rank, r, then there exist r rows of the matrix such that every other row is a linear combination of these r rows.*

Proof: Suppose the determinant rank of $A = (a_{ij})$ equals r. If rows and columns are interchanged, the determinant rank of the modified matrix is unchanged. Thus rows and columns can be interchanged to produce an $r \times r$ matrix in the upper left corner of the matrix which has non zero determinant. Now consider the $r+1 \times r+1$ matrix, M,

$$\begin{pmatrix} a_{11} & \cdots & a_{1r} & a_{1p} \\ \vdots & & \vdots & \vdots \\ a_{r1} & \cdots & a_{rr} & a_{rp} \\ a_{l1} & \cdots & a_{lr} & a_{lp} \end{pmatrix}$$

where C will denote the $r \times r$ matrix in the upper left corner which has non zero determinant. I claim $\det(M) = 0$.

There are two cases to consider in verifying this claim. First, suppose $p > r$. Then the claim follows from the assumption that A has determinant rank r. On the other hand, if $p < r$, then the determinant is zero because there are two identical columns. Expand the determinant along the last column and divide by $\det(C)$ to obtain

$$a_{lp} = -\sum_{i=1}^{r} \frac{\operatorname{cof}(M)_{ip}}{\det(C)} a_{ip}.$$

Now note that $\operatorname{cof}(M)_{ip}$ does not depend on p. Therefore the above sum is of the form

$$a_{lp} = \sum_{i=1}^{r} m_i a_{ip}$$

which shows the l^{th} row is a linear combination of the first r rows of A. Since l is arbitrary, This proves the theorem. ■

Corollary 3.5.20 *The determinant rank equals the row rank.*

Proof: From Theorem 3.5.19, the row rank is no larger than the determinant rank. Could the row rank be smaller than the determinant rank? If so, there exist p rows for $p < r$ such that the span of these p rows equals the row space. But this implies that the $r \times r$ submatrix whose determinant is nonzero also has row rank no larger than p which is impossible if its determinant is to be nonzero because at least one row is a linear combination of the others.

Corollary 3.5.21 *If A has determinant rank, r, then there exist r columns of the matrix such that every other column is a linear combination of these r columns. Also the column rank equals the determinant rank.*

Proof: This follows from the above by considering A^T. The rows of A^T are the columns of A and the determinant rank of A^T and A are the same. Therefore, from Corollary 3.5.20, column rank of A = row rank of A^T = determinant rank of A^T = determinant rank of A.

The following theorem is of fundamental importance and ties together many of the ideas presented above.

Theorem 3.5.22 *Let A be an $n \times n$ matrix. Then the following are equivalent.*

1. $\det(A) = 0$.
2. *A, A^T are not one to one.*
3. *A is not onto.*

Proof: Suppose $\det(A) = 0$. Then the determinant rank of $A = r < n$. Therefore, there exist r columns such that every other column is a linear combination of these columns by Theorem 3.5.19. In particular, it follows that for some m, the m^{th} column is a linear combination of all the others. Thus letting $A = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_m & \cdots & \mathbf{a}_n \end{pmatrix}$ where the columns are denoted by $\mathbf{a}_i$, there exists scalars, α_i such that

$$\mathbf{a}_m = \sum_{k \neq m} \alpha_k \mathbf{a}_k.$$

Now consider the column vector, $\mathbf{x} \equiv \begin{pmatrix} \alpha_1 & \cdots & -1 & \cdots & \boldsymbol{\alpha}_n \end{pmatrix}^T$. Then

$$A\mathbf{x} = -\mathbf{a}_m + \sum_{k \neq m} \alpha_k \mathbf{a}_k = \mathbf{0}.$$

Since also $A\mathbf{0} = \mathbf{0}$, it follows A is not one to one. Similarly, A^T is not one to one by the same argument applied to A^T. This verifies that 1.) implies 2.).

Now suppose 2.). Then since A^T is not one to one, it follows there exists $\mathbf{x} \neq \mathbf{0}$ such that

$$A^T\mathbf{x} = \mathbf{0}.$$

Taking the transpose of both sides yields

$$\mathbf{x}^T A = \mathbf{0}$$

where the $\mathbf{0}$ is a $1 \times n$ matrix or row vector. Now if $A\mathbf{y} = \mathbf{x}$, then

$$|\mathbf{x}|^2 = \mathbf{x}^T (A\mathbf{y}) = \left(\mathbf{x}^T A\right) \mathbf{y} = \mathbf{0}\mathbf{y} = 0$$

contrary to $\mathbf{x} \neq \mathbf{0}$. Consequently there can be no $\mathbf{y}$ such that $A\mathbf{y} = \mathbf{x}$ and so A is not onto. This shows that 2.) implies 3.).

Finally, suppose 3.). If 1.) does not hold, then $\det(A) \neq 0$ but then from Theorem 3.5.14 A^{-1} exists and so for every $\mathbf{y} \in \mathbb{F}^n$ there exists a unique $\mathbf{x} \in \mathbb{F}^n$ such that $A\mathbf{x} = \mathbf{y}$. In fact $\mathbf{x} = A^{-1}\mathbf{y}$. Thus A would be onto contrary to 3.). This shows 3.) implies 1.) and proves the theorem.

Corollary 3.5.23 *Let A be an $n \times n$ matrix. Then the following are equivalent.*

1. *$det(A) \neq 0$.*
2. *A and A^T are one to one.*
3. *A is onto.*

Proof: This follows immediately from the above theorem.

3.5.2 The Determinant Of A Linear Transformation

One can also define the determinant of a linear transformation.

Definition 3.5.24 *Let $L \in \mathcal{L}(V,V)$ and let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis for V. Thus the matrix of L with respect to this basis is $(l_{ij}) \equiv M_L$ where*

$$L = \sum_{ij} l_{ij} \mathbf{v}_i \mathbf{v}_j$$

Then define

$$\det(L) \equiv \det((l_{ij})).$$

Proposition 3.5.25 *The above definition is well defined.*

Proof: Suppose $\{\mathbf{v}'_1, \cdots, \mathbf{v}'_n\}$ is another basis for V and $\left(l'_{ij}\right) \equiv M'_L$ is the matrix of L with respect to this basis. Then by Theorem 3.3.9,

$$M'_L = S^{-1} M_L S$$

for some matrix, S. Then by Theorem 3.5.10,

$$\begin{aligned} \det(M'_L) &= \det\left(S^{-1}\right) \det(M_L) \det(S) \\ &= \det\left(S^{-1}S\right) \det(M_L) = \det(M_L) \end{aligned}$$

because $S^{-1}S = I$ and $\det(I) = 1$. This shows the definition is well defined.

Also there is an equivalence just as in the case of matrices between various properties of L and the nonvanishing of the determinant.

Theorem 3.5.26 *Let $L \in \mathcal{L}(V,V)$ for V a finite dimensional vector space. Then the following are equivalent.*

1. *$\det(L) = 0$.*
2. *L is not one to one.*
3. *L is not onto.*

Proof: Suppose 1.). Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis for V and let (l_{ij}) be the matrix of L with respect to this basis. By definition, $\det((l_{ij})) = 0$ and so (l_{ij}) is not one to one. Thus there is a nonzero vector $\mathbf{x} \in \mathbb{F}^n$ such that $\sum_j l_{ij}x_j = 0$ for each i. Then letting $\mathbf{v} \equiv \sum_{j=1}^n x_j\mathbf{v}_j$,

$$\begin{aligned} L\mathbf{v} &= \sum_{rs} l_{rs}\mathbf{v}_r\mathbf{v}_s \left(\sum_{j=1}^n x_j\mathbf{v}_j\right) = \sum_j \sum_{rs} l_{rs}\mathbf{v}_r\delta_{sj}x_j \\ &= \sum_r \left(\sum_j l_{rj}x_j\right)\mathbf{v}_r = \mathbf{0} \end{aligned}$$

Thus L is not one to one because $L\mathbf{0} = \mathbf{0}$ and $L\mathbf{v} = \mathbf{0}$.

Suppose 2.). Thus there exists $\mathbf{v} \neq \mathbf{0}$ such that $L\mathbf{v} = \mathbf{0}$. Say

$$\mathbf{v} = \sum_i x_i\mathbf{v}_i.$$

Then if $\{L\mathbf{v}_i\}_{i=1}^n$ were linearly independent, it would follow that

$$\mathbf{0} = L\mathbf{v} = \sum_i x_i L\mathbf{v}_i$$

and so all the x_i would equal zero which is not the case. Hence these vectors cannot be linearly independent so they do not span V. Hence there exists

$$\mathbf{w} \in V \setminus \operatorname{span}(L\mathbf{v}_1, \cdots, L\mathbf{v}_n)$$

and therefore, there is no $\mathbf{u} \in V$ such that $L\mathbf{u} = \mathbf{w}$ because if there were such a $\mathbf{u}$, then

$$\mathbf{u} = \sum_i x_i\mathbf{v}_i$$

and so $L\mathbf{u} = \sum_i x_i L\mathbf{v}_i \in \operatorname{span}(L\mathbf{v}_1, \cdots, L\mathbf{v}_n)$.

Finally suppose L is not onto. Then (l_{ij}) also cannot be onto $\mathbb{F}^n$. Therefore, $\det((l_{ij})) \equiv \det(L) = 0$. Why can't (l_{ij}) be onto? If it were, then for any $\mathbf{y} \in \mathbb{F}^n$, there exists $\mathbf{x} \in \mathbb{F}^n$ such that $y_i = \sum_j l_{ij}x_j$. Thus

$$\sum_k y_k\mathbf{v}_k = \sum_{rs} l_{rs}\mathbf{v}_r\mathbf{v}_s \left(\sum_j x_j\mathbf{v}_j\right) = L\left(\sum_j x_j\mathbf{v}_j\right)$$

but the expression on the left in the above formula is that of a general element of V and so L would be onto. This proves the theorem. ■

3.6 Eigenvalues And Eigenvectors Of Linear Transformations

Let V be a finite dimensional vector space. For example, it could be a subspace of $\mathbb{C}^n$or $\mathbb{R}^n$. Also suppose $A \in \mathcal{L}(V, V)$.

Definition 3.6.1 *The characteristic polynomial of A is defined as $q(\lambda) \equiv \det(\lambda\,\mathrm{id} - A)$ where* id *is the identity map which takes every vector in V to itself. The zeros of $q(\lambda)$ in $\mathbb{C}$ are called the eigenvalues of A.*

Lemma 3.6.2 *When λ is an eigenvalue of A which is also in $\mathbb{F}$, the field of scalars, then there exists $\mathbf{v} \neq \mathbf{0}$ such that $A\mathbf{v} = \lambda\mathbf{v}$.*

Proof: This follows from Theorem 3.5.26. Since $\lambda \in \mathbb{F}$,

$$\lambda \,\mathrm{id} - A \in \mathcal{L}(V, V)$$

and since it has zero determinant, it is not one to one so there exists $\mathbf{v} \neq \mathbf{0}$ such that $(\lambda \,\mathrm{id} - A)\,\mathbf{v} = \mathbf{0}$.

The following lemma gives the existence of something called the minimal polynomial. It is an interesting application of the notion of the dimension of $\mathcal{L}(V, V)$.

Lemma 3.6.3 *Let $A \in \mathcal{L}(V, V)$ where V is either a real or a complex finite dimensional vector space of dimension n. Then there exists a polynomial of the form*

$$p(\lambda) = \lambda^m + c_{m-1}\lambda^{m-1} + \cdots + c_1\lambda + c_0$$

such that $p(A) = 0$ and m is as small as possible for this to occur.

Proof: Consider the linear transformations, $I, A, A^2, \cdots, A^{n^2}$. There are $n^2 + 1$ of these transformations and so by Theorem 3.3.5 the set is linearly dependent. Thus there exist constants, $c_i \in \mathbb{F}$ (either $\mathbb{R}$ or $\mathbb{C}$) such that

$$c_0 I + \sum_{k=1}^{n^2} c_k A^k = 0.$$

This implies there exists a polynomial, $q(\lambda)$ which has the property that $q(A) = 0$. In fact, one example is $q(\lambda) \equiv c_0 + \sum_{k=1}^{n^2} c_k \lambda^k$. Dividing by the leading term, it can be assumed this polynomial is of the form $\lambda^m + c_{m-1}\lambda^{m-1} + \cdots + c_1\lambda + c_0$, a monic polynomial. Now consider all such monic polynomials, q such that $q(A) = 0$ and pick the one which has the smallest degree. This is called the minimial polynomial and will be denoted here by $p(\lambda)$. This proves the lemma. ■

Theorem 3.6.4 *Let V be a nonzero finite dimensional vector space of dimension n with the field of scalars equal to $\mathbb{F}$ which is either $\mathbb{R}$ or $\mathbb{C}$. Suppose $A \in \mathcal{L}(V, V)$ and for $p(\lambda)$ the minimal polynomial defined above, let $\mu \in \mathbb{F}$ be a zero of this polynomial. Then there exists $\mathbf{v} \neq \mathbf{0}, \mathbf{v} \in V$ such that*

$$A\mathbf{v} = \mu\mathbf{v}.$$

If $\mathbb{F} = \mathbb{C}$, then A always has an eigenvector and eigenvalue. Furthermore, if $\{\lambda_1, \cdots, \lambda_m\}$ are the zeros of $p(\lambda)$ in $\mathbb{F}$, these are exactly the eigenvalues of A for which there exists an eigenvector in V.

Proof: Suppose first μ is a zero of $p(\lambda)$. Since $p(\mu) = 0$, it follows

$$p(\lambda) = (\lambda - \mu)\, k(\lambda)$$

where $k(\lambda)$ is a polynomial having coefficients in $\mathbb{F}$. Since p has minimal degree, $k(A) \neq 0$ and so there exists a vector, $\mathbf{u} \neq \mathbf{0}$ such that $k(A)\,\mathbf{u} \equiv \mathbf{v} \neq \mathbf{0}$. But then

$$(A - \mu I)\,\mathbf{v} = (A - \mu I)\, k(A)\,(\mathbf{u}) = \mathbf{0}.$$

The next claim about the existence of an eigenvalue follows from the fundamental theorem of algebra and what was just shown.

It has been shown that every zero of $p(\lambda)$ is an eigenvalue which has an eigenvector in V. Now suppose μ is an eigenvalue which has an eigenvector in V so that $A\mathbf{v} = \mu\mathbf{v}$ for some $\mathbf{v} \in V, \mathbf{v} \neq \mathbf{0}$. Does it follow μ is a zero of $p(\lambda)$?

$$\mathbf{0} = p(A)\mathbf{v} = p(\mu)\mathbf{v}$$

and so μ is indeed a zero of $p(\lambda)$. This proves the theorem. ■

In summary, the theorem says the eigenvalues which have eigenvectors in V are exactly the zeros of the minimal polynomial which are in the field of scalars, $\mathbb{F}$.

The idea of block multiplication turns out to be very useful later. For now here is an interesting and significant application which has to do with characteristic polynomials. In this theorem, $p_M(t)$ denotes the polynomial, $\det(tI - M)$. Thus the zeros of this polynomial are the eigenvalues of the matrix, M.

Theorem 3.6.5 *Let A be an $m \times n$ matrix and let B be an $n \times m$ matrix for $m \leq n$. Then*

$$p_{BA}(t) = t^{n-m} p_{AB}(t),$$

so the eigenvalues of BA and AB are the same including multiplicities except that BA has $n - m$ extra zero eigenvalues.

Proof: Use block multiplication to write

$$\begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} AB & ABA \\ B & BA \end{pmatrix}$$

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix} \begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix} = \begin{pmatrix} AB & ABA \\ B & BA \end{pmatrix}.$$

Therefore,

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix}$$

Since the two matrices above are similar it follows that $\begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix}$ and $\begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix}$ have the same characteristic polynomials. Therefore, noting that BA is an $n \times n$ matrix and AB is an $m \times m$ matrix,

$$t^m \det(tI - BA) = t^n \det(tI - AB)$$

and so $\det(tI - BA) = p_{BA}(t) = t^{n-m} \det(tI - AB) = t^{n-m} p_{AB}(t)$. This proves the theorem. ■

3.7 Exercises

1. Let M be an $n \times n$ matrix. Thus letting $M\mathbf{x}$ be defined by ordinary matrix multiplication, it follows $M \in \mathcal{L}(\mathbb{C}^n, \mathbb{C}^n)$. Show that all the zeros of the minimal polynomial are also zeros of the characteristic polynomial. Explain why this requires the minimal polynomial to divide the characteristic polynomial. Thus $q(\lambda) = p(\lambda) k(\lambda)$ for some polynomial $k(\lambda)$ where $q(\lambda)$ is the characteristic polynomial. Now explain why $q(M) = 0$. That every $n \times n$ matrix satisfies its characteristic polynomial is the Cayley Hamilton theorem. Can you extend this to a result about $L \in \mathcal{L}(V, V)$ for V an n dimensional real or complex vector space?

2. Give examples of subspaces of $\mathbb{R}^n$ and examples of subsets of $\mathbb{R}^n$ which are not subspaces.

3. Let $L \in \mathcal{L}(V,W)$. Define $\ker L \equiv \{\mathbf{v} \in V : L\mathbf{v} = \mathbf{0}\}$. Determine whether $\ker L$ is a subspace.

4. Let $L \in \mathcal{L}(V,W)$. Then $L(V)$ denotes those vectors in W such that for some $\mathbf{v}$,$L\mathbf{v} = \mathbf{w}$. Show $L(V)$ is a subspace.

5. Let $L \in \mathcal{L}(V,W)$ and suppose $\{\mathbf{w}_1, \cdots, \mathbf{w}_k\}$ are linearly independent and that $L\mathbf{z}_i = \mathbf{w}_i$. Show $\{\mathbf{z}_1, \cdots, \mathbf{z}_k\}$ is also linearly independent.

6. If $L \in \mathcal{L}(V,W)$ and $\{\mathbf{z}_1, \cdots, \mathbf{z}_k\}$ is linearly independent, what is needed in order that $\{L\mathbf{z}_1, \cdots, L\mathbf{z}_k\}$ be linearly independent? Explain your answer.

7. Let $L \in \mathcal{L}(V,W)$. The rank of L is defined as the dimension of $L(V)$. The nullity of L is the dimension of $\ker(L)$. Show
$$\dim(V) = \text{rank} + \text{nullity}.$$

8. Let $L \in \mathcal{L}(V,W)$ and let $M \in \mathcal{L}(W,Y)$. Show
$$\text{rank}(ML) \leq \min(\text{rank}(L), \text{rank}(M)).$$

9. Let $M(t) = (\mathbf{b}_1(t), \cdots, \mathbf{b}_n(t))$ where each $\mathbf{b}_k(t)$ is a column vector whose component functions are differentiable functions. For such a column vector,
$$\mathbf{b}(t) = (b_1(t), \cdots, b_n(t))^T,$$
define
$$\mathbf{b}'(t) \equiv (b_1'(t), \cdots, b_n'(t))^T$$
Show
$$\det(M(t))' = \sum_{i=1}^{n} \det M_i(t)$$
where $M_i(t)$ has all the same columns as $M(t)$ except the i^{th} column is replaced with $\mathbf{b}_i'(t)$.

10. Let $A = (a_{ij})$ be an $n \times n$ matrix. Consider this as a linear transformation using ordinary matrix multiplication. Show
$$A = \sum_{ij} a_{ij}\mathbf{e}_i\mathbf{e}_j$$
where $\mathbf{e}_i$ is the vector which has a 1 in the i^{th} place and zeros elsewhere.

11. Let $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ be a basis for the vector space, V. Show id, the identity map is given by
$$\text{id} = \sum_{ij} \delta_{ij}\mathbf{w}_i\mathbf{w}_j$$

3.8 Inner Product And Normed Linear Spaces

3.8.1 The Inner Product In $\mathbb{F}^n$

To do calculus, you must understand what you mean by distance. For functions of one variable, the distance was provided by the absolute value of the difference of two numbers. This must be generalized to $\mathbb{F}^n$ and to more general situations.

Definition 3.8.1 *Let $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$. Thus $\mathbf{x} = (x_1, \cdots, x_n)$ where each $x_k \in \mathbb{F}$ and a similar formula holding for $\mathbf{y}$. Then the dot product of these two vectors is defined to be*

$$\mathbf{x} \cdot \mathbf{y} \equiv \sum_j x_j \overline{y_j} \equiv x_1 \overline{y_1} + \cdots + x_n \overline{y_n}.$$

This is also often denoted by $(\mathbf{x}, \mathbf{y})$ and is called an inner product. I will use either notation.

Notice how you put the conjugate on the entries of the vector, $\mathbf{y}$. It makes no difference if the vectors happen to be real vectors but with complex vectors you must do it this way. The reason for this is that when you take the dot product of a vector with itself, you want to get the square of the length of the vector, a positive number. Placing the conjugate on the components of $\mathbf{y}$ in the above definition assures this will take place. Thus

$$\mathbf{x} \cdot \mathbf{x} = \sum_j x_j \overline{x_j} = \sum_j |x_j|^2 \geq 0.$$

If you didn't place a conjugate as in the above definition, things wouldn't work out correctly. For example,

$$(1+i)^2 + 2^2 = 4 + 2i$$

and this is not a positive number.

The following properties of the dot product follow immediately from the definition and you should verify each of them.

Properties of the dot product:

1. $\mathbf{u} \cdot \mathbf{v} = \overline{\mathbf{v} \cdot \mathbf{u}}$.
2. If a, b are numbers and $\mathbf{u}, \mathbf{v}, \mathbf{z}$ are vectors then $(a\mathbf{u} + b\mathbf{v}) \cdot \mathbf{z} = a(\mathbf{u} \cdot \mathbf{z}) + b(\mathbf{v} \cdot \mathbf{z})$.
3. $\mathbf{u} \cdot \mathbf{u} \geq 0$ and it equals 0 if and only if $\mathbf{u} = \mathbf{0}$.

Note this implies $(\mathbf{x} \cdot \alpha\mathbf{y}) = \overline{\alpha}(\mathbf{x} \cdot \mathbf{y})$ because

$$(\mathbf{x} \cdot \alpha\mathbf{y}) = \overline{(\alpha\mathbf{y} \cdot \mathbf{x})} = \overline{\alpha(\mathbf{y} \cdot \mathbf{x})} = \overline{\alpha}(\mathbf{x} \cdot \mathbf{y})$$

The norm is defined as follows.

Definition 3.8.2 *For $\mathbf{x} \in \mathbb{F}^n$,*

$$|\mathbf{x}| \equiv \left(\sum_{k=1}^n |x_k|^2 \right)^{1/2} = (\mathbf{x} \cdot \mathbf{x})^{1/2}$$

3.8.2 General Inner Product Spaces

Any time you have a vector space which possesses an inner product, something satisfying the properties 1 - 3 above, it is called an inner product space.

Here is a fundamental inequality called the **Cauchy Schwarz inequality** which holds in any inner product space. First here is a simple lemma.

Lemma 3.8.3 *If $z \in \mathbb{F}$ there exists $\theta \in \mathbb{F}$ such that $\theta z = |z|$ and $|\theta| = 1$.*

Proof: Let $\theta = 1$ if $z = 0$ and otherwise, let $\theta = \dfrac{\overline{z}}{|z|}$. Recall that for $z = x + iy$, $\overline{z} = x - iy$ and $\overline{z}z = |z|^2$. In case z is real, there is no change in the above.

Theorem 3.8.4 *(Cauchy Schwarz)Let H be an inner product space. The following inequality holds for $\mathbf{x}$ and $\mathbf{y} \in H$.*

$$|(\mathbf{x}\cdot\mathbf{y})| \leq (\mathbf{x}\cdot\mathbf{x})^{1/2}(\mathbf{y}\cdot\mathbf{y})^{1/2} \tag{3.33}$$

Equality holds in this inequality if and only if one vector is a multiple of the other.

Proof: Let $\theta \in \mathbb{F}$ such that $|\theta| = 1$ and

$$\theta(\mathbf{x}\cdot\mathbf{y}) = |(\mathbf{x}\cdot\mathbf{y})|$$

Consider $p(t) \equiv \left(\mathbf{x}+\overline{\theta}t\mathbf{y}\cdot\mathbf{x}+t\overline{\theta}\mathbf{y}\right)$ where $t \in \mathbb{R}$. Then from the above list of properties of the dot product,

$$\begin{aligned} 0 &\leq p(t) = (\mathbf{x}\cdot\mathbf{x}) + t\theta(\mathbf{x}\cdot\mathbf{y}) + t\overline{\theta}(\mathbf{y}\cdot\mathbf{x}) + t^2(\mathbf{y}\cdot\mathbf{y}) \\ &= (\mathbf{x}\cdot\mathbf{x}) + t\theta(\mathbf{x}\cdot\mathbf{y}) + t\overline{\theta}\overline{(\mathbf{x}\cdot\mathbf{y})} + t^2(\mathbf{y}\cdot\mathbf{y}) \\ &= (\mathbf{x}\cdot\mathbf{x}) + 2t\,\mathrm{Re}\,(\theta(\mathbf{x}\cdot\mathbf{y})) + t^2(\mathbf{y}\cdot\mathbf{y}) \\ &= (\mathbf{x}\cdot\mathbf{x}) + 2t|(\mathbf{x}\cdot\mathbf{y})| + t^2(\mathbf{y}\cdot\mathbf{y}) \end{aligned} \tag{3.34}$$

and this must hold for all $t \in \mathbb{R}$. Therefore, if $(\mathbf{y}\cdot\mathbf{y}) = 0$ it must be the case that $|(\mathbf{x}\cdot\mathbf{y})| = 0$ also since otherwise the above inequality would be violated. Therefore, in this case,

$$|(\mathbf{x}\cdot\mathbf{y})| \leq (\mathbf{x}\cdot\mathbf{x})^{1/2}(\mathbf{y}\cdot\mathbf{y})^{1/2}.$$

On the other hand, if $(\mathbf{y}\cdot\mathbf{y}) \neq 0$, then $p(t) \geq 0$ for all t means the graph of $y = p(t)$ is a parabola which opens up and it either has exactly one real zero in the case its vertex touches the t axis or it has no real zeros. From the quadratic formula this happens exactly when

$$4|(\mathbf{x}\cdot\mathbf{y})|^2 - 4(\mathbf{x}\cdot\mathbf{x})(\mathbf{y}\cdot\mathbf{y}) \leq 0$$

which is equivalent to 3.33.

It is clear from a computation that if one vector is a scalar multiple of the other that equality holds in 3.33. Conversely, suppose equality does hold. Then this is equivalent to saying $4|(\mathbf{x}\cdot\mathbf{y})|^2 - 4(\mathbf{x}\cdot\mathbf{x})(\mathbf{y}\cdot\mathbf{y}) = 0$ and so from the quadratic formula, there exists one real zero to $p(t) = 0$. Call it t_0. Then

$$p(t_0) \equiv \left(\mathbf{x}+\overline{\theta}t_0\mathbf{y}, \mathbf{x}+t_0\overline{\theta}\mathbf{y}\right) = \left|\mathbf{x}+\overline{\theta}t\mathbf{y}\right|^2 = 0$$

and so $\mathbf{x} = -\overline{\theta}t_0\mathbf{y}$. This proves the theorem. ■

Note that in establishing the inequality, I only used part of the above properties of the dot product. It was not necessary to use the one which says that if $(\mathbf{x}\cdot\mathbf{x}) = 0$ then $\mathbf{x} = \mathbf{0}$.

Now the length of a vector can be defined.

Definition 3.8.5 *Let $\mathbf{z} \in H$. Then $|\mathbf{z}| \equiv (\mathbf{z}\cdot\mathbf{z})^{1/2}$.*

Theorem 3.8.6 *For length defined in Definition 3.8.5, the following hold.*

$$|\mathbf{z}| \geq 0 \text{ and } |\mathbf{z}| = 0 \text{ if and only if } \mathbf{z} = \mathbf{0} \tag{3.35}$$

$$\text{If } \alpha \text{ is a scalar, } |\alpha\mathbf{z}| = |\alpha||\mathbf{z}| \tag{3.36}$$

$$|\mathbf{z}+\mathbf{w}| \leq |\mathbf{z}| + |\mathbf{w}|. \tag{3.37}$$

Proof: The first two claims are left as exercises. To establish the third,

$$\begin{aligned}
|\mathbf{z}+\mathbf{w}|^2 &\equiv (\mathbf{z}+\mathbf{w},\mathbf{z}+\mathbf{w}) \\
&= \mathbf{z}\cdot\mathbf{z}+\mathbf{w}\cdot\mathbf{w}+\mathbf{w}\cdot\mathbf{z}+\mathbf{z}\cdot\mathbf{w} \\
&= |\mathbf{z}|^2+|\mathbf{w}|^2+2\,\mathrm{Re}\,\mathbf{w}\cdot\mathbf{z} \\
&\leq |\mathbf{z}|^2+|\mathbf{w}|^2+2\,|\mathbf{w}\cdot\mathbf{z}| \\
&\leq |\mathbf{z}|^2+|\mathbf{w}|^2+2\,|\mathbf{w}|\,|\mathbf{z}| = (|\mathbf{z}|+|\mathbf{w}|)^2 .
\end{aligned}$$

3.8.3 Normed Vector Spaces

The best sort of a norm is one which comes from an inner product. However, any vector space, V which has a function, $||\cdot||$ which maps V to $[0,\infty)$ is called a normed vector space if $||\cdot||$ satisfies 3.35 - 3.37. That is

$$||\mathbf{z}|| \geq 0 \text{ and } ||\mathbf{z}|| = 0 \text{ if and only if } \mathbf{z}=\mathbf{0} \tag{3.38}$$

$$\text{If } \alpha \text{ is a scalar, } ||\alpha\mathbf{z}|| = |\alpha|\,||\mathbf{z}|| \tag{3.39}$$

$$||\mathbf{z}+\mathbf{w}|| \leq ||\mathbf{z}|| + ||\mathbf{w}|| . \tag{3.40}$$

The last inequality above is called the triangle inequality. Another version of this is

$$|\,||\mathbf{z}|| - ||\mathbf{w}||\,| \leq ||\mathbf{z}-\mathbf{w}|| \tag{3.41}$$

To see that 3.41 holds, note

$$||\mathbf{z}|| = ||\mathbf{z}-\mathbf{w}+\mathbf{w}|| \leq ||\mathbf{z}-\mathbf{w}|| + ||\mathbf{w}||$$

which implies

$$||\mathbf{z}|| - ||\mathbf{w}|| \leq ||\mathbf{z}-\mathbf{w}||$$

and now switching $\mathbf{z}$ and $\mathbf{w}$, yields

$$||\mathbf{w}|| - ||\mathbf{z}|| \leq ||\mathbf{z}-\mathbf{w}||$$

which implies 3.41.

3.8.4 The p Norms

Examples of norms are the p norms on $\mathbb{C}^n$.

Definition 3.8.7 *Let $\mathbf{x}\in\mathbb{C}^n$. Then define for $p\geq 1$,*

$$||\mathbf{x}||_p \equiv \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$$

The following inequality is called Holder's inequality.

Proposition 3.8.8 *For $\mathbf{x},\mathbf{y}\in\mathbb{C}^n$,*

$$\sum_{i=1}^n |x_i|\,|y_i| \leq \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}\left(\sum_{i=1}^n |y_i|^{p'}\right)^{1/p'}$$

The proof will depend on the following lemma.

Lemma 3.8.9 *If $a, b \geq 0$ and p' is defined by $\frac{1}{p} + \frac{1}{p'} = 1$, then*

$$ab \leq \frac{a^p}{p} + \frac{b^{p'}}{p'}.$$

Proof of the Proposition: If $\mathbf{x}$ or $\mathbf{y}$ equals the zero vector there is nothing to prove. Therefore, assume they are both nonzero. Let $A = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ and $B = \left(\sum_{i=1}^n |y_i|^{p'}\right)^{1/p'}$. Then using Lemma 3.8.9,

$$\begin{aligned}
\sum_{i=1}^n \frac{|x_i|}{A}\frac{|y_i|}{B} &\leq \sum_{i=1}^n \left[\frac{1}{p}\left(\frac{|x_i|}{A}\right)^p + \frac{1}{p'}\left(\frac{|y_i|}{B}\right)^{p'}\right] \\
&= \frac{1}{p}\frac{1}{A^p}\sum_{i=1}^n |x_i|^p + \frac{1}{p'}\frac{1}{B^p}\sum_{i=1}^n |y_i|^{p'} \\
&= \frac{1}{p} + \frac{1}{p'} = 1
\end{aligned}$$

and so

$$\sum_{i=1}^n |x_i|\,|y_i| \leq AB = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}\left(\sum_{i=1}^n |y_i|^{p'}\right)^{1/p'}.$$

This proves the proposition. ■

Theorem 3.8.10 *The p norms do indeed satisfy the axioms of a norm.*

Proof: It is obvious that $||\cdot||_p$ does indeed satisfy most of the norm axioms. The only one that is not clear is the triangle inequality. To save notation write $||\cdot||$ in place of $||\cdot||_p$ in what follows. Note also that $\frac{p}{p'} = p - 1$. Then using the Holder inequality,

$$\begin{aligned}
||\mathbf{x}+\mathbf{y}||^p &= \sum_{i=1}^n |x_i + y_i|^p \\
&\leq \sum_{i=1}^n |x_i + y_i|^{p-1}|x_i| + \sum_{i=1}^n |x_i + y_i|^{p-1}|y_i| \\
&= \sum_{i=1}^n |x_i + y_i|^{\frac{p}{p'}}|x_i| + \sum_{i=1}^n |x_i + y_i|^{\frac{p}{p'}}|y_i| \\
&\leq \left(\sum_{i=1}^n |x_i + y_i|^p\right)^{1/p'}\left[\left(\sum_{i=1}^n |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^n |y_i|^p\right)^{1/p}\right] \\
&= ||\mathbf{x}+\mathbf{y}||^{p/p'}\left(||\mathbf{x}||_p + ||\mathbf{y}||_p\right)
\end{aligned}$$

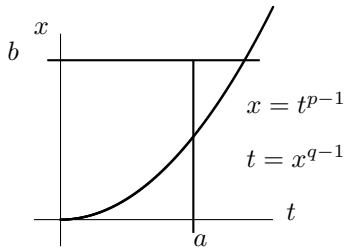
so dividing by $||\mathbf{x}+\mathbf{y}||^{p/p'}$, it follows

$$||\mathbf{x}+\mathbf{y}||^p\,||\mathbf{x}+\mathbf{y}||^{-p/p'} = ||\mathbf{x}+\mathbf{y}|| \leq ||\mathbf{x}||_p + ||\mathbf{y}||_p$$

$\left(p - \frac{p}{p'} = p\left(1 - \frac{1}{p'}\right) = p\frac{1}{p} = 1.\right)$. This proves the theorem. ■

It only remains to prove Lemma 3.8.9.

Proof of the lemma: Let $p' = q$ to save on notation and consider the following picture:



$$ab \leq \int_0^a t^{p-1}dt + \int_0^b x^{q-1}dx = \frac{a^p}{p} + \frac{b^q}{q}.$$

Note equality occurs when $a^p = b^q$.

Alternate proof of the lemma: Let

$$f(t) \equiv \frac{1}{p}(at)^p + \frac{1}{q}\left(\frac{b}{t}\right)^q,\ t > 0$$

You see right away it is decreasing for a while, having an assymptote at $t = 0$ and then reaches a minimum and increases from then on. Take its derivative.

$$f'(t) = (at)^{p-1}a + \left(\frac{b}{t}\right)^{q-1}\left(\frac{-b}{t^2}\right)$$

Set it equal to 0. This happens when

$$t^{p+q} = \frac{b^q}{a^p}. \tag{3.42}$$

Thus

$$t = \frac{b^{q/(p+q)}}{a^{p/(p+q)}}$$

and so at this value of t,

$$at = (ab)^{q/(p+q)},\ \left(\frac{b}{t}\right) = (ab)^{p/(p+q)}.$$

Thus the minimum of f is

$$\frac{1}{p}\left((ab)^{q/(p+q)}\right)^p + \frac{1}{q}\left((ab)^{p/(p+q)}\right)^q = (ab)^{pq/(p+q)}$$

but recall $1/p + 1/q = 1$ and so $pq/(p+q) = 1$. Thus the minimum value of f is ab. Letting $t = 1$, this shows

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Note that equality occurs when the minimum value happens for $t = 1$ and this indicates from 3.42 that $a^p = b^q$. This proves the lemma. ■

3.8.5 Orthonormal Bases

Not all bases for an inner product space H are created equal. The best bases are orthonormal.

Definition 3.8.11 *Suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is a set of vectors in an inner product space H. It is an orthonormal set if*

$$\mathbf{v}_i \cdot \mathbf{v}_j = \delta_{ij} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

Every orthonormal set of vectors is automatically linearly independent.

Proposition 3.8.12 *Suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is an orthonormal set of vectors. Then it is linearly independent.*

Proof: Suppose $\sum_{i=1}^k c_i \mathbf{v}_i = \mathbf{0}$. Then taking dot products with $\mathbf{v}_j$,

$$0 = \mathbf{0} \cdot \mathbf{v}_j = \sum_i c_i \mathbf{v}_i \cdot \mathbf{v}_j = \sum_i c_i \delta_{ij} = c_j.$$

Since j is arbitrary, this shows the set is linearly independent as claimed.

It turns out that if X is any subspace of H, then there exists an orthonormal basis for X.

Lemma 3.8.13 *Let X be a subspace of dimension n whose basis is $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$. Then there exists an orthonormal basis for X, $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ which has the property that for each $k \leq n$, span$(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$.*

Proof: Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be a basis for X. Let $\mathbf{u}_1 \equiv \mathbf{x}_1 / |\mathbf{x}_1|$. Thus for $k = 1$, $\text{span}(\mathbf{u}_1) = \text{span}(\mathbf{x}_1)$ and $\{\mathbf{u}_1\}$ is an orthonormal set. Now suppose for some $k < n$, $\mathbf{u}_1, \cdots, \mathbf{u}_k$ have been chosen such that $(\mathbf{u}_j, \mathbf{u}_l) = \delta_{jl}$ and $\text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$. Then define

$$\mathbf{u}_{k+1} \equiv \frac{\mathbf{x}_{k+1} - \sum_{j=1}^k (\mathbf{x}_{k+1} \cdot \mathbf{u}_j) \mathbf{u}_j}{\left| \mathbf{x}_{k+1} - \sum_{j=1}^k (\mathbf{x}_{k+1} \cdot \mathbf{u}_j) \mathbf{u}_j \right|}, \tag{3.43}$$

where the denominator is not equal to zero because the $\mathbf{x}_j$ form a basis and so

$$\mathbf{x}_{k+1} \notin \text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$$

Thus by induction,

$$\mathbf{u}_{k+1} \in \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{x}_{k+1}) = \text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k, \mathbf{x}_{k+1}).$$

Also, $\mathbf{x}_{k+1} \in \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1})$ which is seen easily by solving 3.43 for $\mathbf{x}_{k+1}$ and it follows

$$\text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k, \mathbf{x}_{k+1}) = \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1}).$$

If $l \leq k$, then denoting by C the scalar $\left| \mathbf{x}_{k+1} - \sum_{j=1}^k (\mathbf{x}_{k+1} \cdot \mathbf{u}_j) \mathbf{u}_j \right|^{-1}$,

$$\begin{aligned} (\mathbf{u}_{k+1} \cdot \mathbf{u}_l) &= C \left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - \sum_{j=1}^k (\mathbf{x}_{k+1} \cdot \mathbf{u}_j) (\mathbf{u}_j \cdot \mathbf{u}_l) \right) \\ &= C \left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - \sum_{j=1}^k (\mathbf{x}_{k+1} \cdot \mathbf{u}_j) \delta_{lj} \right) \\ &= C \left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - (\mathbf{x}_{k+1} \cdot \mathbf{u}_l) \right) = 0. \end{aligned}$$

The vectors, $\{\mathbf{u}_j\}_{j=1}^n$, generated in this way are therefore an orthonormal basis because each vector has unit length.

The process by which these vectors were generated is called the Gram Schmidt process.

3.8.6 The Adjoint Of A Linear Transformation

There is a very important collection of ideas which relates a linear transformation to the inner product in an inner product space. In order to discuss these ideas, it is necessary to prove a simple and very interesting lemma about linear transformations which map an inner product space H to the field of scalars, $\mathbb{F}$. This is sometimes called the Riesz representation theorem.

Theorem 3.8.14 *Let H be a finite dimensional inner product space and let $L \in \mathcal{L}(H, \mathbb{F})$. Then there exists a unique $\mathbf{z} \in H$ such that for all $\mathbf{x} \in H$,*

$$L\mathbf{x} = (\mathbf{x} \cdot \mathbf{z}).$$

Proof: By the Gram Schmidt process, there exists an orthonormal basis for H,

$$\{\mathbf{e}_1, \cdots, \mathbf{e}_n\}.$$

First note that if $\mathbf{x}$ is arbitrary, there exist unique scalars, x_i such that

$$\mathbf{x} = \sum_{i=1}^n x_i \mathbf{e}_i$$

Taking the dot product of both sides with $\mathbf{e}_k$ yields

$$(\mathbf{x} \cdot \mathbf{e}_k) = \left(\sum_{i=1}^n x_i \mathbf{e}_i \cdot \mathbf{e}_k \right) = \sum_{i=1}^n x_i (\mathbf{e}_i \cdot \mathbf{e}_k) = \sum_{i=1}^n x_i \delta_{ik} = x_k$$

which shows that

$$\mathbf{x} = \sum_{i=1}^n (\mathbf{x} \cdot \mathbf{e}_i) \mathbf{e}_i$$

and so by the properties of the dot product,

$$L\mathbf{x} = \sum_{i=1}^n (\mathbf{x} \cdot \mathbf{e}_i) L\mathbf{e}_i = \left(\mathbf{x} \cdot \sum_{i=1}^n \mathbf{e}_i \overline{L\mathbf{e}_i} \right)$$

so let $\mathbf{z} = \sum_{i=1}^n \mathbf{e}_i \overline{L\mathbf{e}_i}$. It only remains to verify $\mathbf{z}$ is unique. However, this is obvious because if $(\mathbf{x} \cdot \mathbf{z}_1) = (\mathbf{x} \cdot \mathbf{z}_2) = L\mathbf{x}$ for all $\mathbf{x}$, then

$$(\mathbf{x} \cdot \mathbf{z}_1 - \mathbf{z}_2) = 0$$

for all $\mathbf{x}$ and in particular for $\mathbf{x} = \mathbf{z}_1 - \mathbf{z}_2$ which requires $\mathbf{z}_1 = \mathbf{z}_2$. This proves the theorem. ■

Now with this theorem, it becomes easy to define something called the adjoint of a linear operator. Let $L \in \mathcal{L}(H_1, H_2)$ where H_1 and H_2 are finite dimensional inner product spaces. Then letting $(\cdot)_i$ denote the inner product in H_i,

$$\mathbf{x} \to (L\mathbf{x} \cdot \mathbf{y})_2$$

is in $\mathcal{L}(H_1, \mathbb{F})$ and so from Theorem 3.8.14 there exists a unique element of H_1, denoted by $L^*\mathbf{y}$ such that for all $\mathbf{x} \in H_1$,

$$(L\mathbf{x} \cdot \mathbf{y})_2 = (\mathbf{x} \cdot L^*\mathbf{y})_1$$

Thus $L^*\mathbf{y} \in H_1$ when $\mathbf{y} \in H_2$. Also L^* is linear. This is because by the properties of the dot product,

$$\begin{aligned}(\mathbf{x} \cdot L^*(\alpha\mathbf{y} + \beta\mathbf{z}))_1 &\equiv (L\mathbf{x} \cdot \alpha\mathbf{y} + \beta\mathbf{z})_2 \\ &= \overline{\alpha}(L\mathbf{x} \cdot \mathbf{y})_2 + \overline{\beta}(L\mathbf{x} \cdot \mathbf{z})_2 \\ &= \overline{\alpha}(\mathbf{x} \cdot L^*\mathbf{y})_1 + \overline{\beta}(\mathbf{x} \cdot L^*\mathbf{z})_1 \\ &= \overline{\alpha}(\mathbf{x} \cdot L^*\mathbf{y})_1 + \overline{\beta}(\mathbf{x} \cdot L^*\mathbf{z})_1\end{aligned}$$

and

$$(\mathbf{x} \cdot \alpha L^*\mathbf{y} + \beta L^*\mathbf{z})_1 = \overline{\alpha}(\mathbf{x} \cdot L^*\mathbf{y})_1 + \overline{\beta}(\mathbf{x} \cdot L^*\mathbf{z})_1$$

Since

$$(\mathbf{x} \cdot L^*(\alpha\mathbf{y} + \beta\mathbf{z}))_1 = (\mathbf{x} \cdot \alpha L^*\mathbf{y} + \beta L^*\mathbf{z})_1$$

for all $\mathbf{x}$, this requires

$$L^*(\alpha\mathbf{y} + \beta\mathbf{z}) = \alpha L^*\mathbf{y} + \beta L^*\mathbf{z}.$$

In simple words, when you take it across the dot, you put a star on it. More precisely, here is the definition.

Definition 3.8.15 *Let H_1 and H_2 be finite dimensional inner product spaces and let $L \in \mathcal{L}(H_1, H_2)$. Then $L^* \in \mathcal{L}(H_2, H_1)$ is defined by the formula*

$$(L\mathbf{x} \cdot \mathbf{y})_2 = (\mathbf{x} \cdot L^*\mathbf{y})_1.$$

In the case where $H_1 = H_2 = H$, an operator $L \in \mathcal{L}(H, H)$ is said to be self adjoint if $L = L^*$. This is also called Hermitian.

The following diagram might help.

$$\begin{array}{ccc} H_1 & \overset{L^*}{\leftarrow} & H_2 \\ H_1 & \overset{L}{\rightarrow} & H_2 \end{array}$$

I will not bother to place subscripts on the symbol for the dot product in the future. I will be clear from context which inner product is meant.

Proposition 3.8.16 *The adjoint has the following properties.*

1. $(\mathbf{x} \cdot L\mathbf{y}) = (L^*\mathbf{x} \cdot \mathbf{y}),\ (L\mathbf{x} \cdot \mathbf{y}) = (\mathbf{x} \cdot L^*\mathbf{y})$
2. $(L^*)^* = L$
3. $(aL + bM)^* = \overline{a}L^* + \overline{b}M^*$
4. $(ML)^* = L^*M^*$

Proof: Consider the first claim.

$$(\mathbf{x} \cdot L\mathbf{y}) = \overline{(L\mathbf{y} \cdot \mathbf{x})} = \overline{(\mathbf{y} \cdot L^*\mathbf{x})} = (L^*\mathbf{x} \cdot \mathbf{y})$$

This does the first claim. The second part was discussed earlier when the adjoint was defined.

Consider the second claim. From the first claim,

$$(L\mathbf{x}\cdot\mathbf{y}) = (\mathbf{x}\cdot L^*\mathbf{y}) = \left(\left(L^*\right)^*\mathbf{x}\cdot\mathbf{y}\right)$$

and since this holds for all $\mathbf{y}$, it follows $L\mathbf{x} = \left(L^*\right)^*\mathbf{x}$.

Consider the third claim.

$$\left(\mathbf{x}\cdot\left(aL+bM\right)^*\mathbf{y}\right) = \left(\left(aL+bM\right)\mathbf{x}\cdot\mathbf{y}\right) = a\left(L\mathbf{x}\cdot\mathbf{y}\right) + b\left(M\mathbf{x}\cdot\mathbf{y}\right)$$

and

$$\left(\mathbf{x}\cdot\left(\overline{a}L^*+\overline{b}M^*\right)\mathbf{y}\right) = a\left(\mathbf{x}\cdot L^*\mathbf{y}\right) + b\left(\mathbf{x}\cdot M^*\mathbf{y}\right) = a\left(L\mathbf{x}\cdot\mathbf{y}\right) + b\left(M\mathbf{x}\cdot\mathbf{y}\right)$$

and since $\left(\mathbf{x}\cdot\left(aL+bM\right)^*\mathbf{y}\right) = \left(\mathbf{x}\cdot\left(\overline{a}L^*+\overline{b}M^*\right)\mathbf{y}\right)$ for all $\mathbf{x}$, it must be that

$$\left(aL+bM\right)^*\mathbf{y} = \left(\overline{a}L^*+\overline{b}M^*\right)\mathbf{y}$$

for all $\mathbf{y}$ which yields the third claim.

Consider the fourth.

$$\left(\mathbf{x}\cdot\left(ML\right)^*\mathbf{y}\right) = \left(\left(ML\right)\mathbf{x}\cdot\mathbf{y}\right) = \left(L\mathbf{x}\cdot M^*\mathbf{y}\right) = \left(\mathbf{x}\cdot L^*M^*\mathbf{y}\right)$$

Since this holds for all $\mathbf{x},\mathbf{y}$ the conclusion follows as above. This proves the theorem. ■

Here is a very important example.

Example 3.8.17 *Suppose $F \in \mathcal{L}\left(H_1,H_2\right)$. Then $FF^* \in \mathcal{L}\left(H_2,H_2\right)$ and is self adjoint.*

To see this is so, note it is the composition of linear transformations and is therefore linear as stated. To see it is self adjoint, Proposition 3.8.16 implies

$$\left(FF^*\right)^* = \left(F^*\right)^*F^* = FF^*$$

In the case where $A \in \mathcal{L}\left(\mathbb{F}^n,\mathbb{F}^m\right)$, considering the matrix of A with respect to the usual bases, there is no loss of generality in considering A to be an $m\times n$ matrix,

$$\left(A\mathbf{x}\right)_i = \sum_j A_{ij}x_j.$$

Then in terms of components of the matrix, A,

$$\left(A^*\right)_{ij} = \overline{A_{ji}}.$$

You should verify this is so from the definition of the usual inner product on $\mathbb{F}^k$. The following little proposition is useful.

Proposition 3.8.18 *Suppose A is an $m\times n$ matrix where $m\leq n$. Also suppose*

$$\det\left(AA^*\right) \neq 0.$$

Then A has m linearly independent rows and m independent columns.

Proof: Since $\det\left(AA^*\right)\neq 0$, it follows the $m\times m$ matrix AA^* has m independent rows. If this is not true of A, then there exists $\mathbf{x}$ a $1\times m$ matrix such that

$$\mathbf{x}A = \mathbf{0}.$$

Hence

$$\mathbf{x}AA^* = \mathbf{0}$$

and this contradicts the independence of the rows of AA^*. Thus the row rank of A equals m and by Corollary 3.5.20 this implies the column rank of A also equals m. This proves the proposition. ■

3.8.7 Schur's Theorem

Recall that for a linear transformation, $L \in \mathcal{L}(V,V)$, it could be represented in the form

$$L = \sum_{ij} l_{ij} \mathbf{v}_i \mathbf{v}_j$$

where $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis. Of course different bases will yield different matrices, (l_{ij}). Schur's theorem gives the existence of a basis in an inner product space such that (l_{ij}) is particularly simple.

Definition 3.8.19 *Let $L \in \mathcal{L}(V,V)$ where V is vector space. Then a subspace U of V is L invariant if $L(U) \subseteq U$.*

Theorem 3.8.20 *Let $L \in \mathcal{L}(H,H)$ for H a finite dimensional inner product space such that the restriction of L^* to every L invariant subspace has its eigenvalues in $\mathbb{F}$. Then there exist constants, c_{ij} for $i \leq j$ and an orthonormal basis, $\{\mathbf{w}_i\}_{i=1}^n$ such that*

$$L = \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \mathbf{w}_j$$

The constants, c_{ii} are the eigenvalues of L.

Proof: If $\dim(H) = 1$ let $H = \operatorname{span}(\mathbf{w})$ where $|\mathbf{w}| = 1$. Then $L\mathbf{w} = k\mathbf{w}$ for some k. Then

$$L = k\mathbf{w}\mathbf{w}$$

because by definition, $\mathbf{w}\mathbf{w}(\mathbf{w}) = \mathbf{w}$. Therefore, the theorem holds if H is 1 dimensional.

Now suppose the theorem holds for $n-1 = \dim(H)$. By Theorem 3.6.4 and the assumption, there exists $\mathbf{w}_n$, an eigenvector for L^*. Dividing by its length, it can be assumed $|\mathbf{w}_n| = 1$. Say $L^*\mathbf{w}_n = \mu\mathbf{w}_n$. Using the Gram Schmidt process, there exists an orthonormal basis for H of the form $\{\mathbf{v}_1, \cdots, \mathbf{v}_{n-1}, \mathbf{w}_n\}$. Then

$$(L\mathbf{v}_k \cdot \mathbf{w}_n) = (\mathbf{v}_k \cdot L^*\mathbf{w}_n) = (\mathbf{v}_k \cdot \mu\mathbf{w}_n) = 0,$$

which shows

$$L : H_1 \equiv \operatorname{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{n-1}) \to \operatorname{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{n-1}).$$

Denote by L_1 the restriction of L to H_1. Since H_1 has dimension $n-1$, the induction hypothesis yields an orthonormal basis, $\{\mathbf{w}_1, \cdots, \mathbf{w}_{n-1}\}$ for H_1 such that

$$L_1 = \sum_{j=1}^{n-1} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \mathbf{w}_j. \tag{3.44}$$

Then $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ is an orthonormal basis for H because every vector in

$$\operatorname{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{n-1})$$

has the property that its dot product with $\mathbf{w}_n$ is 0 so in particular, this is true for the vectors $\{\mathbf{w}_1, \cdots, \mathbf{w}_{n-1}\}$. Now define c_{in} to be the scalars satisfying

$$L\mathbf{w}_n \equiv \sum_{i=1}^{n} c_{in} \mathbf{w}_i \tag{3.45}$$

and let

$$B \equiv \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \mathbf{w}_j.$$

Then by 3.45,

$$B\mathbf{w}_n = \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \delta_{nj} = \sum_{j=1}^{n} c_{in} \mathbf{w}_i = L\mathbf{w}_n.$$

If $1 \leq k \leq n-1$,

$$B\mathbf{w}_k = \sum_{j=1}^{n} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \delta_{kj} = \sum_{i=1}^{k} c_{ik} \mathbf{w}_i$$

while from 3.44,

$$L\mathbf{w}_k = L_1 \mathbf{w}_k = \sum_{j=1}^{n-1} \sum_{i=1}^{j} c_{ij} \mathbf{w}_i \delta_{jk} = \sum_{i=1}^{k} c_{ik} \mathbf{w}_i.$$

Since $L = B$ on the basis $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$, it follows $L = B$.

It remains to verify the constants, c_{kk} are the eigenvalues of L, solutions of the equation, $\det(\lambda I - L) = 0$. However, the definition of $\det(\lambda I - L)$ is the same as

$$\det(\lambda I - C)$$

where C is the upper triangular matrix which has c_{ij} for $i \leq j$ and zeros elsewhere. This equals 0 if and only if λ is one of the diagonal entries, one of the c_{kk}. This proves the theorem. ■

There is a technical assumption in the above theorem about the eigenvalues of restrictions of L^* being in $\mathbb{F}$, the field of scalars. If $\mathbb{F} = \mathbb{C}$ this is no restriction. There is also another situation in which $\mathbb{F} = \mathbb{R}$ for which this will hold.

Lemma 3.8.21 *Suppose H is a finite dimensional inner product space and*

$$\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$$

is an orthonormal basis for H. Then

$$(\mathbf{w}_i \mathbf{w}_j)^* = \mathbf{w}_j \mathbf{w}_i$$

Proof: It suffices to verify the two linear transformations are equal on $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$. Then

$$\begin{aligned} \left(\mathbf{w}_p \cdot (\mathbf{w}_i \mathbf{w}_j)^* \mathbf{w}_k\right) &\equiv ((\mathbf{w}_i \mathbf{w}_j) \mathbf{w}_p \cdot \mathbf{w}_k) = (\mathbf{w}_i \delta_{jp} \cdot \mathbf{w}_k) = \delta_{jp}\delta_{ik} \\ (\mathbf{w}_p \cdot (\mathbf{w}_j \mathbf{w}_i) \mathbf{w}_k) &= (\mathbf{w}_p \cdot \mathbf{w}_j \delta_{ik}) = \delta_{ik}\delta_{jp} \end{aligned}$$

Since $\mathbf{w}_p$ is arbitrary, it follows from the properties of the inner product that

$$\left(\mathbf{x} \cdot (\mathbf{w}_i \mathbf{w}_j)^* \mathbf{w}_k\right) = (\mathbf{x} \cdot (\mathbf{w}_j \mathbf{w}_i) \mathbf{w}_k)$$

for all $\mathbf{x} \in H$ and hence $(\mathbf{w}_i \mathbf{w}_j)^* \mathbf{w}_k = (\mathbf{w}_j \mathbf{w}_i) \mathbf{w}_k$. Since $\mathbf{w}_k$ is arbitrary, This proves the lemma. ■

Lemma 3.8.22 *Let $L \in \mathcal{L}(H, H)$ for H an inner product space. Then if $L = L^*$ so L is self adjoint, it follows all the eigenvalues of L are real.*

Proof: Let $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ be an orthonormal basis for H and let (l_{ij}) be the matrix of L with respect to this orthonormal basis. Thus

$$L = \sum_{ij} l_{ij}\mathbf{w}_i\mathbf{w}_j, \ \mathrm{id} = \sum_{ij} \delta_{ij}\mathbf{w}_i\mathbf{w}_j$$

Denote by M_L the matrix whose ij^{th} entry is l_{ij}. Then by definition of what is meant by the determinant of a linear transformation,

$$\det(\lambda\,\mathrm{id} - L) = \det(\lambda I - M_L)$$

and so the eigenvalues of L are the same as the eigenvalues of M_L. However, $M_L \in \mathcal{L}(\mathbb{C}^n, \mathbb{C}^n)$ with $M_L\mathbf{x}$ determined by ordinary matrix multiplication. Therefore, by the fundamental theorem of algebra and Theorem 3.6.4, if λ is an eigenvalue of L it follows there exists a nonzero $\mathbf{x} \in \mathbb{C}^n$ such that $M_L\mathbf{x} = \lambda\mathbf{x}$. Since L is self adjoint, it follows from Lemma 3.8.21

$$L = \sum_{ij} l_{ij}\mathbf{w}_i\mathbf{w}_j = L^* = \sum_{ij} \overline{l_{ij}}\mathbf{w}_j\mathbf{w}_i = \sum_{ij} \overline{l_{ji}}\mathbf{w}_i\mathbf{w}_j$$

which shows $l_{ij} = \overline{l_{ji}}$.

Then

$$\begin{aligned} \lambda|\mathbf{x}|^2 &= \lambda(\mathbf{x}\cdot\mathbf{x}) = (\lambda\mathbf{x}\cdot\mathbf{x}) = (M_L\mathbf{x}\cdot\mathbf{x}) = \sum_{ij} l_{ij}x_j\overline{x_i} \\ &= \overline{\sum_{ij} \overline{l_{ij}}\,\overline{x_j}x_i} = \overline{\sum_{ij} l_{ji}\overline{x_j}x_i} = \overline{(M_L\mathbf{x}\cdot\mathbf{x})} = (\mathbf{x}\cdot M_L\mathbf{x}) = \overline{\lambda}|\mathbf{x}|^2 \end{aligned}$$

showing $\lambda = \overline{\lambda}$. This proves the lemma. ■

If L is a self adjoint operator on H, either a real or complex inner product space, it follows the condition about the eigenvalues of the restrictions of L^* to L invariant subspaces of H must hold because these restrictions are self adjoint. Here is why. Let $\mathbf{x}, \mathbf{y}$ be in one of those invariant subspaces. Then since $L^* = L$,

$$(L^*\mathbf{x}\cdot\mathbf{y}) = (\mathbf{x}\cdot L\mathbf{y}) = (\mathbf{x}\cdot L^*\mathbf{y})$$

so by the above lemma, the eigenvalues are real and are therefore, in the field of scalars.

Now with this lemma, the following theorem is obtained. This is another major theorem. It is equivalent to the theorem in matrix theory which states every self adjoint matrix can be diagonalized.

Theorem 3.8.23 *Let H be a finite dimensional inner product space, real or complex, and let $L \in \mathcal{L}(H, H)$ be self adjoint. Then there exists an orthonormal basis $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ and real scalars, λ_k such that*

$$L = \sum_{k=1}^{n} \lambda_k\mathbf{w}_k\mathbf{w}_k.$$

The scalars are the eigenvalues and $\mathbf{w}_k$ is an eigenvector for λ_k for each k.

Proof: By Theorem 3.8.20, there exists an orthonormal basis, $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ such that

$$L = \sum_{j=1}^{n}\sum_{i=1}^{n} c_{ij}\mathbf{w}_i\mathbf{w}_j$$

where $c_{ij} = 0$ if $i > j$. Now using Lemma 3.8.21 and Proposition 3.8.16 along with the assumption that L is self adjoint,

$$L = \sum_{j=1}^{n}\sum_{i=1}^{n} c_{ij}\mathbf{w}_i\mathbf{w}_j = L^* = \sum_{j=1}^{n}\sum_{i=1}^{n} \overline{c_{ij}}\mathbf{w}_j\mathbf{w}_i = \sum_{i=1}^{n}\sum_{j=1}^{n} \overline{c_{ji}}\mathbf{w}_i\mathbf{w}_j$$

If $i < j$, then this shows $c_{ij} = \overline{c_{ji}}$ and the second number equals zero because $j > i$. Thus $c_{ij} = 0$ if $i < j$ and it is already known that $c_{ij} = 0$ if $i > j$. Therefore, let $\lambda_k = c_{kk}$ and the above reduces to

$$L = \sum_{j=1}^{n} \lambda_j\mathbf{w}_j\mathbf{w}_j = \sum_{j=1}^{n} \overline{\lambda_j}\mathbf{w}_j\mathbf{w}_j$$

showing that $\lambda_j = \overline{\lambda_j}$ so the eigenvalues are all real. Now

$$L\mathbf{w}_k = \sum_{j=1}^{n} \lambda_j\mathbf{w}_j\mathbf{w}_j\left(\mathbf{w}_k\right) = \sum_{j=1}^{n} \lambda_j\mathbf{w}_j\delta_{jk} = \lambda_k\mathbf{w}_k$$

which shows all the $\mathbf{w}_k$ are eigenvectors. This proves the theorem. ■

3.9 Polar Decompositions

An application of Theorem 3.8.23, is the following fundamental result, important in geometric measure theory and continuum mechanics. It is sometimes called the right polar decomposition. When the following theorem is applied in continuum mechanics, F is normally the deformation gradient, the derivative, discussed later, of a nonlinear map from some subset of three dimensional space to three dimensional space. In this context, U is called the right Cauchy Green strain tensor. It is a measure of how a body is stretched independent of rigid motions. First, here is a simple lemma.

Lemma 3.9.1 *Suppose $R \in \mathcal{L}(X, Y)$ where X, Y are finite dimensional inner product spaces and R preserves distances,*

$$|R\mathbf{x}|_Y = |\mathbf{x}|_X .$$

Then $R^*R = I$.

Proof: Since R preserves distances, $|R\mathbf{x}| = |\mathbf{x}|$ for every $\mathbf{x}$. Therefore from the axioms of the dot product,

$$\begin{aligned} & |\mathbf{x}|^2 + |\mathbf{y}|^2 + (\mathbf{x}\cdot\mathbf{y}) + (\mathbf{y}\cdot\mathbf{x}) \\ = \; & |\mathbf{x}+\mathbf{y}|^2 \\ = \; & (R(\mathbf{x}+\mathbf{y})\cdot R(\mathbf{x}+\mathbf{y})) \\ = \; & (R\mathbf{x}\cdot R\mathbf{x}) + (R\mathbf{y}\cdot R\mathbf{y}) + (R\mathbf{x}\cdot R\mathbf{y}) + (R\mathbf{y}\cdot R\mathbf{x}) \\ = \; & |\mathbf{x}|^2 + |\mathbf{y}|^2 + (R^*R\mathbf{x}\cdot\mathbf{y}) + (\mathbf{y}\cdot R^*R\mathbf{x}) \end{aligned}$$

and so for all $\mathbf{x}, \mathbf{y}$,

$$(R^*R\mathbf{x} - \mathbf{x}\cdot\mathbf{y}) + (\mathbf{y}\cdot R^*R\mathbf{x} - \mathbf{x}) = 0$$

Hence for all $\mathbf{x}, \mathbf{y}$,

$$\mathrm{Re}\,(R^*R\mathbf{x} - \mathbf{x}\cdot\mathbf{y}) = 0$$

Now for $\mathbf{x}, \mathbf{y}$ given, choose $\alpha \in \mathbb{C}$ such that

$$\alpha \left(R^* R\mathbf{x} - \mathbf{x} \cdot \mathbf{y} \right) = \left| (R^* R\mathbf{x} - \mathbf{x} \cdot \mathbf{y}) \right|$$

Then

$$\begin{aligned} 0 &= \operatorname{Re} \left(R^* R\mathbf{x} - \mathbf{x} \cdot \overline{\alpha}\mathbf{y} \right) = \operatorname{Re} \alpha \left(R^* R\mathbf{x} - \mathbf{x} \cdot \mathbf{y} \right) \\ &= \left| (R^* R\mathbf{x} - \mathbf{x} \cdot \mathbf{y}) \right| \end{aligned}$$

Thus $\left| (R^* R\mathbf{x} - \mathbf{x} \cdot \mathbf{y}) \right| = 0$ for all $\mathbf{x}, \mathbf{y}$ because the given $\mathbf{x}, \mathbf{y}$ were arbitrary. Let $\mathbf{y} = R^* R\mathbf{x} - \mathbf{x}$ to conclude that for all $\mathbf{x}$,

$$R^* R\mathbf{x} - \mathbf{x} = \mathbf{0}$$

which says $R^* R = I$ since $\mathbf{x}$ is arbitrary. This proves the lemma. ■

Definition 3.9.2 *In case $R \in \mathcal{L}(X, X)$ for X a real or complex inner product space of dimension n, R is said to be unitary if it preserves distances. Thus, from the above lemma, unitary transformations are those which satisfy*

$$R^* R = R R^* = \mathrm{id}$$

where id *is the identity map on X.*

Theorem 3.9.3 *Let X be a real or complex inner product space of dimension n, let Y be a real or complex inner product space of dimension $m \geq n$ and let $F \in \mathcal{L}(X, Y)$. Then there exists $R \in \mathcal{L}(X, Y)$ and $U \in \mathcal{L}(X, X)$ such that*

$$F = RU,\ U = U^*, (U \text{ is Hermitian}),$$

all eigenvalues of U are non negative,

$$U^2 = F^* F, R^* R = I,$$

and $|R\mathbf{x}| = |\mathbf{x}|$.

Proof: $(F^* F)^* = F^* F$ and so by Theorem 3.8.23, there is an orthonormal basis of eigenvectors for X, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ such that

$$F^* F = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i,\ F^* F \mathbf{v}_k = \lambda_k \mathbf{v}_k.$$

It is also clear that $\lambda_i \geq 0$ because

$$\lambda_i \left(\mathbf{v}_i \cdot \mathbf{v}_i \right) = \left(F^* F \mathbf{v}_i \cdot \mathbf{v}_i \right) = \left(F\mathbf{v}_i \cdot F\mathbf{v}_i \right) \geq 0.$$

Let

$$U \equiv \sum_{i=1}^{n} \lambda_i^{1/2} \mathbf{v}_i \mathbf{v}_i.$$

so U maps X to X and is self adjoint. Then from 3.13,

$$\begin{aligned} U^2 &= \sum_{ij} \lambda_i^{1/2} \lambda_j^{1/2} \left(\mathbf{v}_i \mathbf{v}_i \right) \left(\mathbf{v}_j \mathbf{v}_j \right) \\ &= \sum_{ij} \lambda_i^{1/2} \lambda_j^{1/2} \mathbf{v}_i \mathbf{v}_j \delta_{ij} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i = F^* F \end{aligned}$$

Let $\{U\mathbf{x}_1, \cdots, U\mathbf{x}_r\}$ be an orthonormal basis for $U(X)$. Extend this using the Gram Schmidt procedure to an orthonormal basis for X,

$$\{U\mathbf{x}_1, \cdots, U\mathbf{x}_r, \mathbf{y}_{r+1}, \cdots, \mathbf{y}_n\}.$$

Next note that $\{F\mathbf{x}_1, \cdots, F\mathbf{x}_r\}$ is also an orthonormal set of vectors in Y because

$$(F\mathbf{x}_k \cdot F\mathbf{x}_j) = (F^*F\mathbf{x}_k \cdot \mathbf{x}_j) = \left(U^2\mathbf{x}_k \cdot \mathbf{x}_j\right) = (U\mathbf{x}_k \cdot U\mathbf{x}_j) = \delta_{jk}.$$

Now extend $\{F\mathbf{x}_1, \cdots, F\mathbf{x}_r\}$ to an orthonormal basis for Y,

$$\{F\mathbf{x}_1, \cdots, F\mathbf{x}_r, \mathbf{z}_{r+1}, \cdots, \mathbf{z}_m\}.$$

Since $m \geq n$, there are at least as many $\mathbf{z}_k$ as there are $\mathbf{y}_k$.

Now define R as follows. For $\mathbf{x} \in X$, there exist unique scalars, c_k and d_k such that

$$\mathbf{x} = \sum_{k=1}^{r} c_k U\mathbf{x}_k + \sum_{k=r+1}^{n} d_k \mathbf{y}_k.$$

Then

$$R\mathbf{x} \equiv \sum_{k=1}^{r} c_k F\mathbf{x}_k + \sum_{k=r+1}^{n} d_k \mathbf{z}_k. \tag{3.46}$$

Thus, since $\{F\mathbf{x}_1, \cdots, F\mathbf{x}_r, \mathbf{z}_{r+1}, \cdots, \mathbf{z}_m\}$ is orthonormal, a short computation shows

$$|R\mathbf{x}|^2 = \sum_{k=1}^{r} |c_k|^2 + \sum_{k=r+1}^{n} |d_k|^2 = |\mathbf{x}|^2.$$

Now I need to verify $RU\mathbf{x} = F\mathbf{x}$. Since $\{U\mathbf{x}_1, \cdots, U\mathbf{x}_r\}$ is an orthonormal basis for UX, there exist scalars, b_k such that

$$U\mathbf{x} = \sum_{k=1}^{r} b_k U\mathbf{x}_k \tag{3.47}$$

and so from the definition of R given in 3.46,

$$RU\mathbf{x} \equiv \sum_{k=1}^{r} b_k F\mathbf{x}_k = F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right).$$

$RU = F$ is shown if $F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right) = F(\mathbf{x})$.

$$\begin{aligned}
&\left(F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right) - F(\mathbf{x}) \cdot F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right) - F(\mathbf{x})\right) \\
&= \left(F^*F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right) \cdot \sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right) \\
&= \left(U^2\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right) \cdot \left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right)\right) \\
&= \left(U\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right) \cdot U\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right)\right) \\
&= \left(\sum_{k=1}^{r} b_k U\mathbf{x}_k - U\mathbf{x} \cdot \sum_{k=1}^{r} b_k U\mathbf{x}_k - U\mathbf{x}\right) = 0
\end{aligned}$$

by 3.47.

Since $|R\mathbf{x}| = |\mathbf{x}|$, it follows $R^*R = I$ from Lemma 3.9.1. This proves the theorem. ■

The following corollary follows as a simple consequence of this theorem. It is called the left polar decomposition.

Corollary 3.9.4 *Let $F \in \mathcal{L}(X,Y)$ and suppose $n \geq m$ where X is a inner product space of dimension n and Y is a inner product space of dimension m. Then there exists a Hermitian $U \in \mathcal{L}(X,X)$, and an element of $\mathcal{L}(X,Y)$, R, such that*

$$F = UR,\ RR^* = I.$$

Proof: Recall that $L^{**} = L$ and $(ML)^* = L^*M^*$. Now apply Theorem 3.9.3 to $F^* \in \mathcal{L}(X,Y)$. Thus,

$$F^* = R^*U$$

where R^* and U satisfy the conditions of that theorem. Then

$$F = UR$$

and $RR^* = R^{**}R^* = I$. This proves the corollary. ■

This is a good place to consider a useful lemma.

Lemma 3.9.5 *Let X be a finite dimensional inner product space of dimension n and let $R \in \mathcal{L}(X,X)$ be unitary. Then $|\det(R)| = 1$.*

Proof: Let $\{\mathbf{w}_k\}$ be an orthonormal basis for X. Then to take the determinant it suffices to take the determinant of the matrix, (c_{ij}) where

$$R = \sum_{ij} c_{ij}\mathbf{w}_i\mathbf{w}_j.$$

$R\mathbf{w}_k = \sum_i c_{ik}\mathbf{w}_i$ and so

$$(R\mathbf{w}_k, \mathbf{w}_l) = c_{lk}.$$

and hence

$$R = \sum_{lk} (R\mathbf{w}_k, \mathbf{w}_l)\, \mathbf{w}_l\mathbf{w}_k$$

Similarly

$$R^* = \sum_{ij} (R^*\mathbf{w}_j, \mathbf{w}_i)\, \mathbf{w}_i\mathbf{w}_j.$$

Since R is given to be unitary,

$$\begin{aligned} RR^* &= \mathrm{id} = \sum_{lk}\sum_{ij} (R\mathbf{w}_k, \mathbf{w}_l)\,(R^*\mathbf{w}_j, \mathbf{w}_i)\,(\mathbf{w}_l\mathbf{w}_k)\,(\mathbf{w}_i\mathbf{w}_j) \\ &= \sum_{ijkl} (R\mathbf{w}_k, \mathbf{w}_l)\,(R^*\mathbf{w}_j, \mathbf{w}_i)\,\delta_{ki}\mathbf{w}_l\mathbf{w}_j \\ &= \sum_{jl} \left(\sum_i (R\mathbf{w}_i, \mathbf{w}_l)\,(R^*\mathbf{w}_j, \mathbf{w}_i) \right) \mathbf{w}_l\mathbf{w}_j \end{aligned}$$

Hence

$$\sum_i (R^*\mathbf{w}_j, \mathbf{w}_i)\,(R\mathbf{w}_i, \mathbf{w}_l) = \delta_{jl} = \sum_i \overline{(R\mathbf{w}_i, \mathbf{w}_j)}\,(R\mathbf{w}_i, \mathbf{w}_l) \tag{3.48}$$

because

$$\mathrm{id} = \sum_{jl} \delta_{jl} \mathbf{w}_l \mathbf{w}_j$$

Thus letting M be the matrix whose ij^{th} entry is $(R\mathbf{w}_i, \mathbf{w}_l)$, $\det(R)$ is defined as $\det(M)$ and 3.48 says

$$\sum_i \left(\overline{M^T}\right)_{ji} M_{il} = \delta_{jl}.$$

It follows

$$1 = \det(M) \det\left(\overline{M}^T\right) = \det(M) \det\left(\overline{M}\right) = \det(M) \overline{\det(M)} = |\det(M)|^2.$$

Thus $|\det(R)| = |\det(M)| = 1$ as claimed.

3.10 Exercises

1. For $\mathbf{u}, \mathbf{v}$ vectors in $\mathbb{F}^3$, define the product, $\mathbf{u} * \mathbf{v} \equiv u_1\overline{v_1} + 2u_2\overline{v_2} + 3u_3\overline{v_3}$. Show the axioms for a dot product all hold for this funny product. Prove

$$|\mathbf{u} * \mathbf{v}| \leq (\mathbf{u} * \mathbf{u})^{1/2} (\mathbf{v} * \mathbf{v})^{1/2}.$$

2. Suppose you have a real or complex vector space. Can it always be considered as an inner product space? What does this mean about Schur's theorem? **Hint:** Start with a basis and decree the basis is orthonormal. Then define an inner product accordingly.

3. Show that $(\mathbf{a} \cdot \mathbf{b}) = \frac{1}{4}\left[|\mathbf{a} + \mathbf{b}|^2 - |\mathbf{a} - \mathbf{b}|^2\right]$.

4. Prove from the axioms of the dot product the parallelogram identity, $|\mathbf{a} + \mathbf{b}|^2 + |\mathbf{a} - \mathbf{b}|^2 = 2|\mathbf{a}|^2 + 2|\mathbf{b}|^2$.

5. Suppose f, g are two Darboux Stieltjes integrable functions defined on $[0, 1]$. Define

$$(f \cdot g) = \int_0^1 f(x) \overline{g(x)} dF.$$

   Show this dot product satisfies the axioms for the inner product. Explain why the Cauchy Schwarz inequality continues to hold in this context and state the Cauchy Schwarz inequality in terms of integrals. Does the Cauchy Schwarz inequality still hold if

$$(f \cdot g) = \int_0^1 f(x) \overline{g(x)} p(x) dF$$

   where $p(x)$ is a given nonnegative function? If so, what would it be in terms of integrals.

6. If A is an $n \times n$ matrix considered as an element of $\mathcal{L}(\mathbb{C}^n, \mathbb{C}^n)$ by ordinary matrix multiplication, use the inner product in $\mathbb{C}^n$ to show that $(A^*)_{ij} = \overline{A_{ji}}$. In words, the adjoint is the transpose of the conjugate.

7. A symmetric matrix is a real $n \times n$ matrix A which satisfies $A^T = A$. Show every symmetric matrix is self adjoint and that there exists an orthonormal set of real vectors $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ such that

$$A = \sum_k \lambda_k \mathbf{x}_k \mathbf{x}_k$$

8. A normal matrix is an $n \times n$ matrix, A such that $A^*A = AA^*$. Show that for a normal matrix there is an orthonormal basis of $\mathbb{C}^n, \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ such that

$$A = \sum_i a_i \mathbf{x}_i \mathbf{x}_i$$

That is, with respect to this basis the matrix of A is diagonal. **Hint:** This is a harder version of what was done to prove Theorem 3.8.23. Use Schur's theorem to write $A = \sum_{j=1}^n \sum_{i=1}^n B_{ij} \mathbf{w}_i \mathbf{w}_j$ where B_{ij} is an upper triangular matrix. Then use the condition that A is normal and eventually get an equation

$$\sum_k B_{ik} \overline{B_{lk}} = \sum_k \overline{B_{ki}} B_{kl}$$

Next let $i = l$ and consider first $l = 1$, then $l = 2$, etc. If you are careful, you will find $B_{ij} = 0$ unless $i = j$.

9. Suppose $A \in \mathcal{L}(H, H)$ where H is an inner product space and

$$A = \sum_i a_i \mathbf{w}_i \mathbf{w}_i$$

where the vectors $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ are an orthonormal set. Show that A must be normal. In other words, you can't represent $A \in \mathcal{L}(H, H)$ in this very convenient way unless it is normal.

10. If L is a self adjoint operator defined on an inner product space, H such that L has all only nonnegative eigenvalues. Explain how to define $L^{1/n}$ and show why what you come up with is indeed the n^{th} root of the operator. For a self adjoint operator L on an inner product space, can you define $\sin(L) \equiv \sum_{k=0}^{\infty} (-1)^k L^{2k+1} / (2k+1)!$? What does the infinite series mean? Can you make some sense of this using the representation of L given in Theorem 3.8.23?

11. If L is a self adjoint linear transformation defined on $\mathcal{L}(H, H)$ for H an inner product space which has all eigenvalues nonnegative, show the square root is unique.

12. Using Problem 11 show $F \in \mathcal{L}(H, H)$ for H an inner product space is normal if and only if $RU = UR$ where $F = RU$ is the right polar decomposition defined above. Recall R preserves distances and U is self adjoint. What is the geometric significance of a linear transformation being normal?

13. Suppose you have a basis, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ in an inner product space, X. The Grammian matrix is the $n \times n$ matrix whose ij^{th} entry is $(\mathbf{v}_i \cdot \mathbf{v}_j)$. Show this matrix is invertible. **Hint:** You might try to show that the inner product of two vectors, $\sum_k a_k \mathbf{v}_k$ and $\sum_k b_k \mathbf{v}_k$ has something to do with the Grammian.

14. Suppose you have a basis, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ in an inner product space, X. Show there exists a "dual basis" $\{\mathbf{v}^1, \cdots, \mathbf{v}^n\}$ which satisfies $\mathbf{v}^k \cdot \mathbf{v}_j = \delta_j^k$, which equals 0 if $j \neq k$ and equals 1 if $j = k$.

Chapter 4

Sequences

4.1 Vector Valued Sequences And Their Limits

Functions defined on the set of integers larger than a given integer which have values in a vector space are called vector valued sequences. I will always assume the vector space is a normed vector space. Actually, it will specialized even more to $\mathbb{F}^n$, although everything can be done for an arbitrary vector space and when it creates no difficulties, I will state certain definitions and easy theorems in the more general context and use the symbol $||\cdot||$ to refer to the norm. Other than this, the notation is almost the same as it was when the sequences had values in $\mathbb{C}$. The main difference is that certain variables are placed in bold face to indicate they are vectors. Even this is not really necessary but it is conventional to do it.The concept of subsequence is also the same as it was for sequences of numbers. To review,

Definition 4.1.1 *Let $\{\mathbf{a}_n\}$ be a sequence and let $n_1 < n_2 < n_3, \cdots$ be any strictly increasing list of integers such that n_1 is at least as large as the first number in the domain of the function. Then if $\mathbf{b}_k \equiv \mathbf{a}_{n_k}$, $\{\mathbf{b}_k\}$ is called a subsequence of $\{\mathbf{a}_n\}$.*

Example 4.1.2 *Let $\mathbf{a}_n = \left(n+1, \sin\left(\frac{1}{n}\right)\right)$. Then $\{\mathbf{a}_n\}_{n=1}^{\infty}$ is a vector valued sequence.*

The definition of a limit of a vector valued sequence is given next. It is just like the definition given for sequences of scalars. However, here the symbol $|\cdot|$ refers to the usual norm in $\mathbb{F}^n$. In a general normed vector space, it will be denoted by $||\cdot||$.

Definition 4.1.3 *A vector valued sequence $\{\mathbf{a}_n\}_{n=1}^{\infty}$ converges to $\mathbf{a}$ in a normed vector space V, written as*

$$\lim_{n\to\infty} \mathbf{a}_n = \mathbf{a} \text{ or } \mathbf{a}_n \to \mathbf{a}$$

if and only if for every $\varepsilon > 0$ there exists n_ε such that whenever $n \geq n_\varepsilon$,

$$||\mathbf{a}_n - \mathbf{a}|| < \varepsilon.$$

In words the definition says that given any measure of closeness ε, the terms of the sequence are eventually this close to $\mathbf{a}$. Here, the word "eventually" refers to n being sufficiently large.

Theorem 4.1.4 *If $\lim_{n\to\infty} \mathbf{a}_n = \mathbf{a}$ and $\lim_{n\to\infty} \mathbf{a}_n = \mathbf{a}_1$ then $\mathbf{a}_1 = \mathbf{a}$.*

Proof: Suppose $\mathbf{a}_1 \neq \mathbf{a}$. Then let $0 < \varepsilon < ||\mathbf{a}_1 - \mathbf{a}||/2$ in the definition of the limit. It follows there exists n_ε such that if $n \geq n_\varepsilon$, then $||\mathbf{a}_n - \mathbf{a}|| < \varepsilon$ and $|\mathbf{a}_n - \mathbf{a}_1| < \varepsilon$. Therefore, for such n,

$$\begin{aligned} ||\mathbf{a}_1 - \mathbf{a}|| &\leq ||\mathbf{a}_1 - \mathbf{a}_n|| + ||\mathbf{a}_n - \mathbf{a}|| \\ &< \varepsilon + \varepsilon < ||\mathbf{a}_1 - \mathbf{a}||/2 + ||\mathbf{a}_1 - \mathbf{a}||/2 = ||\mathbf{a}_1 - \mathbf{a}||, \end{aligned}$$

a contradiction.

Theorem 4.1.5 *Suppose $\{\mathbf{a}_n\}$ and $\{\mathbf{b}_n\}$ are vector valued sequences and that*

$$\lim_{n\to\infty} \mathbf{a}_n = \mathbf{a} \text{ and } \lim_{n\to\infty} \mathbf{b}_n = \mathbf{b}.$$

Also suppose x and y are scalars in $\mathbb{F}$. Then

$$\lim_{n\to\infty} x\mathbf{a}_n + y\mathbf{b}_n = x\mathbf{a} + y\mathbf{b} \tag{4.1}$$

Also,

$$\lim_{n\to\infty} (\mathbf{a}_n \cdot \mathbf{b}_n) = (\mathbf{a} \cdot \mathbf{b}) \tag{4.2}$$

If $\{x_n\}$ is a sequence of scalars in $\mathbb{F}$ converging to x and if $\{\mathbf{a}_n\}$ is a sequence of vectors in $\mathbb{F}^n$ converging to $\mathbf{a}$, then

$$\lim_{n\to\infty} x_n\mathbf{a}_n = x\mathbf{a}. \tag{4.3}$$

Also if $\{\mathbf{x}_k\}$ is a sequence of vectors in $\mathbb{F}^n$ then $\mathbf{x}_k \to \mathbf{x}$, if and only if for each j,

$$\lim_{k\to\infty} x_k^j = x^j. \tag{4.4}$$

where here

$$\mathbf{x}_k = \left(x_k^1, \cdots, x_k^n\right), \ \mathbf{x} = \left(x^1, \cdots, x^n\right).$$

Proof: Consider the first claim. By the triangle inequality

$$||x\mathbf{a} + y\mathbf{b} - (x\mathbf{a}_n + y\mathbf{b}_n)|| \leq |x|\,||\mathbf{a} - \mathbf{a}_n|| + |y|\,||\mathbf{b} - \mathbf{b}_n||.$$

By definition, there exists n_ε such that if $n \geq n_\varepsilon$,

$$||\mathbf{a} - \mathbf{a}_n||, ||\mathbf{b} - \mathbf{b}_n|| < \frac{\varepsilon}{2(1 + |x| + |y|)}$$

so for $n > n_\varepsilon$,

$$||x\mathbf{a} + y\mathbf{b} - (x\mathbf{a}_n + y\mathbf{b}_n)|| < |x| \frac{\varepsilon}{2(1 + |x| + |y|)} + |y| \frac{\varepsilon}{2(1 + |x| + |y|)} \leq \varepsilon.$$

Now consider the second. Let $\varepsilon > 0$ be given and choose n_1 such that if $n \geq n_1$ then

$$|\mathbf{a}_n - \mathbf{a}| < 1.$$

For such n, it follows from the Cauchy Schwarz inequality and properties of the inner product that

$$\begin{aligned} |\mathbf{a}_n \cdot \mathbf{b}_n - \mathbf{a} \cdot \mathbf{b}| &\leq |(\mathbf{a}_n \cdot \mathbf{b}_n) - (\mathbf{a}_n \cdot \mathbf{b})| + |(\mathbf{a}_n \cdot \mathbf{b}) - (\mathbf{a} \cdot \mathbf{b})| \\ &\leq |\mathbf{a}_n|\,|\mathbf{b}_n - \mathbf{b}| + |\mathbf{b}|\,|\mathbf{a}_n - \mathbf{a}| \\ &\leq (|\mathbf{a}| + 1)\,|\mathbf{b}_n - \mathbf{b}| + |\mathbf{b}|\,|\mathbf{a}_n - \mathbf{a}|. \end{aligned}$$

Now let n_2 be large enough that for $n \geq n_2$,

$$|\mathbf{b}_n - \mathbf{b}| < \frac{\varepsilon}{2(|\mathbf{a}|+1)}, \text{ and } |\mathbf{a}_n - \mathbf{a}| < \frac{\varepsilon}{2(|\mathbf{b}|+1)}.$$

Such a number exists because of the definition of limit. Therefore, let

$$n_\varepsilon > \max(n_1, n_2).$$

For $n \geq n_\varepsilon$,

$$\begin{aligned} |\mathbf{a}_n \cdot \mathbf{b}_n - \mathbf{a} \cdot \mathbf{b}| &\leq (|\mathbf{a}|+1)|\mathbf{b}_n - \mathbf{b}| + |\mathbf{b}||\mathbf{a}_n - \mathbf{a}| \\ &< (|\mathbf{a}|+1)\frac{\varepsilon}{2(|\mathbf{a}|+1)} + |\mathbf{b}|\frac{\varepsilon}{2(|\mathbf{b}|+1)} \leq \varepsilon. \end{aligned}$$

This proves 4.2. The claim, 4.3 is left for you to do.

Finally consider the last claim. If 4.4 holds, then from the definition of distance in $\mathbb{F}^n$,

$$\lim_{k\to\infty} |\mathbf{x} - \mathbf{x}_k| \equiv \lim_{k\to\infty} \sqrt{\sum_{j=1}^{n} \left(x^j - x_k^j\right)^2} = 0.$$

On the other hand, if $\lim_{k\to\infty} |\mathbf{x} - \mathbf{x}_k| = 0$, then since $\left|x_k^j - x^j\right| \leq |\mathbf{x} - \mathbf{x}_k|$, it follows from the squeezing theorem that

$$\lim_{k\to\infty} \left|x_k^j - x^j\right| = 0.$$

This proves the theorem. ■

An important theorem is the one which states that if a sequence converges, so does every subsequence. You should review Definition 4.1.1 at this point. The proof is identical to the one involving sequences of numbers.

Theorem 4.1.6 *Let $\{\mathbf{x}_n\}$ be a vector valued sequence with $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}$ and let $\{\mathbf{x}_{n_k}\}$ be a subsequence. Then $\lim_{k\to\infty} \mathbf{x}_{n_k} = \mathbf{x}$.*

Proof: Let $\varepsilon > 0$ be given. Then there exists n_ε such that if $n > n_\varepsilon$, then $||\mathbf{x}_n - \mathbf{x}|| < \varepsilon$. Suppose $k > n_\varepsilon$. Then $n_k \geq k > n_\varepsilon$ and so

$$||\mathbf{x}_{n_k} - \mathbf{x}|| < \varepsilon$$

showing $\lim_{k\to\infty} \mathbf{x}_{n_k} = \mathbf{x}$ as claimed.

Theorem 4.1.7 *Let $\{x_n\}$ be a sequence of real numbers and suppose each $x_n \leq l$ $(\geq l)$ and $\lim_{n\to\infty} x_n = x$. Then $x \leq l$ $(\geq l)$. More generally, suppose $\{x_n\}$ and $\{y_n\}$ are two sequences such that $\lim_{n\to\infty} x_n = x$ and $\lim_{n\to\infty} y_n = y$. Then if $x_n \leq y_n$ for all n sufficiently large, then $x \leq y$.*

Proof: Let $\varepsilon > 0$ be given. Then for n large enough,

$$l \geq x_n > x - \varepsilon$$

and so

$$l + \varepsilon \geq x.$$

Since $\varepsilon > 0$ is arbitrary, this requires $l \geq x$. The other case is entirely similar or else you could consider $-l$ and $\{-x_n\}$ and apply the case just considered.

Consider the last claim. There exists N such that if $n \geq N$ then $x_n \leq y_n$ and

$$|x - x_n| + |y - y_n| < \varepsilon/2.$$

Then considering $n > N$ in what follows,

$$x - y \leq x_n + \varepsilon/2 - (y_n - \varepsilon/2) = x_n - y_n + \varepsilon \leq \varepsilon.$$

Since ε was arbitrary, it follows $x - y \leq 0$. This proves the theorem. ■

Theorem 4.1.8 *Let $\{\mathbf{x}_n\}$ be a sequence vectors and suppose each $||\mathbf{x}_n|| \leq l$ $(\geq l)$ and $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}$. Then $\mathbf{x} \leq l\,(\geq l)$. More generally, suppose $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$ are two sequences such that $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}$ and $\lim_{n\to\infty} \mathbf{y}_n = \mathbf{y}$. Then if $||\mathbf{x}_n|| \leq ||\mathbf{y}_n||$ for all n sufficiently large, then $||\mathbf{x}|| \leq ||\mathbf{y}||$.*

Proof: It suffices to just prove the second part since the first part is similar. By the triangle inequality,

$$|||\mathbf{x}_n|| - ||\mathbf{x}||| \leq ||\mathbf{x}_n - \mathbf{x}||$$

and for large n this is given to be small. Thus $\{||\mathbf{x}_n||\}$ converges to $||\mathbf{x}||$. Similarly $\{||\mathbf{y}_n||\}$ converges to $||\mathbf{y}||$. Now the desired result follows from Theorem 4.1.7. This proves the theorem. ■

4.2 Sequential Compactness

The following is the definition of sequential compactness. It is a very useful notion which can be used to prove existence theorems.

Definition 4.2.1 *A set, $K \subseteq V$, a normed vector space is sequentially compact if whenever $\{\mathbf{a}_n\} \subseteq K$ is a sequence, there exists a subsequence, $\{\mathbf{a}_{n_k}\}$ such that this subsequence converges to a point of K.*

First of all, it is convenient to consider the sequentially compact sets in $\mathbb{F}$.

Lemma 4.2.2 *Let $I_k = \left[a^k, b^k\right]$ and suppose that for all $k = 1, 2, \cdots$,*

$$I_k \supseteq I_{k+1}.$$

Then there exists a point, $c \in \mathbb{R}$ which is an element of every I_k.

Proof: Since $I_k \supseteq I_{k+1}$, this implies

$$a^k \leq a^{k+1},\ b^k \geq b^{k+1}. \tag{4.5}$$

Consequently, if $k \leq l$,

$$a^l \leq a^l \leq b^l \leq b^k. \tag{4.6}$$

Now define

$$c \equiv \sup\left\{a^l : l = 1, 2, \cdots\right\}$$

By the first inequality in 4.5, and 4.6

$$a^k \leq c = \sup\left\{a^l : l = k, k+1, \cdots\right\} \leq b^k \tag{4.7}$$

for each $k = 1, 2 \cdots$. Thus $c \in I_k$ for every k and This proves the lemma. ■ If this went too fast, the reason for the last inequality in 4.7 is that from 4.6, b^k is an upper bound to $\left\{a^l : l = k, k+1, \cdots\right\}$. Therefore, it is at least as large as the least upper bound.

Theorem 4.2.3 *Every closed interval, $[a,b]$ is sequentially compact.*

Proof: Let $\{x_n\} \subseteq [a,b] \equiv I_0$. Consider the two intervals $\left[a, \frac{a+b}{2}\right]$ and $\left[\frac{a+b}{2}, b\right]$ each of which has length $(b-a)/2$. At least one of these intervals contains x_n for infinitely many values of n. Call this interval I_1. Now do for I_1 what was done for I_0. Split it in half and let I_2 be the interval which contains x_n for infinitely many values of n. Continue this way obtaining a sequence of nested intervals $I_0 \supseteq I_1 \supseteq I_2 \supseteq I_3 \cdots$ where the length of I_n is $(b-a)/2^n$. Now pick n_1 such that $x_{n_1} \in I_1$, n_2 such that $n_2 > n_1$ and $x_{n_2} \in I_2, n_3$ such that $n_3 > n_2$ and $x_{n_3} \in I_3$, etc. (This can be done because in each case the intervals contained x_n for infinitely many values of n.) By the nested interval lemma there exists a point, c contained in all these intervals. Furthermore,

$$|x_{n_k} - c| < (b-a)\, 2^{-k}$$

and so $\lim_{k\to\infty} x_{n_k} = c \in [a,b]$. This proves the theorem. ■

Theorem 4.2.4 *Let*

$$I = \prod_{k=1}^{n} K_k$$

where K_k is a sequentially compact set in $\mathbb{F}$. Then I is a sequentially compact set in $\mathbb{F}^n$.

Proof: Let $\{\mathbf{x}_k\}_{k=1}^{\infty}$ be a sequence of points in I. Let

$$\mathbf{x}_k = \left(x_k^1, \cdots, x_k^n\right)$$

Thus $\left\{x_k^i\right\}_{k=1}^{\infty}$ is a sequence of points in K_i. Since K_i is sequentially compact, there exists a subsequence of $\{\mathbf{x}_k\}_{k=1}^{\infty}$ denoted by $\{\mathbf{x}_{1k}\}$ such that $\left\{x_{1k}^1\right\}$ converges to x^1 for some $x^1 \in K_1$. Now there exists a further subsequence, $\{\mathbf{x}_{2k}\}$ such that $\left\{x_{2k}^1\right\}$ converges to x^1, because by Theorem 4.1.6, subsequences of convergent sequences converge to the same limit as the convergent sequence, and in addition, $\left\{x_{2k}^2\right\}$ converges to some $x^2 \in K_2$. Continue taking subsequences such that for $\{\mathbf{x}_{jk}\}_{k=1}^{\infty}$, it follows $\left\{x_{jk}^r\right\}$ converges to some $x^r \in K_r$ for all $r \leq j$. Then $\{\mathbf{x}_{nk}\}_{k=1}^{\infty}$ is the desired subsequence such that the sequence of numbers in $\mathbb{F}$ obtained by taking the j^{th} component of this subsequence converges to some $x^j \in K_j$. It follows from Theorem 4.1.5 that $\mathbf{x} \equiv \left(x^1, \cdots, x^n\right) \in I$ and is the limit of $\{\mathbf{x}_{nk}\}_{k=1}^{\infty}$. This proves the theorem. ■

Corollary 4.2.5 *Any box of the form*

$$[a,b] + i\,[c,d] \equiv \{x + iy : x \in [a,b], y \in [c,d]\}$$

is sequentially compact in $\mathbb{C}$.

Proof: The given box is essentially $[a,b] \times [c,d]$.

$$\{x_k + iy_k\}_{k=1}^{\infty} \subseteq [a,b] + i\,[c,d]$$

is the same as saying $(x_k, y_k) \in [a,b] \times [c,d]$. Therefore, there exists $(x,y) \in [a,b] \times [c,d]$ such that $x_k \to x$ and $y_k \to y$. In other words $x_k + iy_k \to x + iy$ and $x + iy \in [a,b] + i\,[c,d]$. This proves the corollary. ■

4.3 Closed And Open Sets

The definition of open and closed sets is next.

Definition 4.3.1 *Let U be a set of points in a normed vector space, V. A point, $\mathbf{p} \in U$ is said to be an interior point if whenever $||\mathbf{x}-\mathbf{p}||$ is sufficiently small, it follows $\mathbf{x} \in U$ also. The set of points, $\mathbf{x}$ which are closer to $\mathbf{p}$ than δ is denoted by*

$$B(\mathbf{p},\delta) \equiv \{\mathbf{x} \in V : ||\mathbf{x}-\mathbf{p}|| < \delta\}.$$

This symbol, $B(\mathbf{p},\delta)$ is called an open ball of radius δ. Thus a point, $\mathbf{p}$ is an interior point of U if there exists $\delta > 0$ such that $\mathbf{p} \in B(\mathbf{p},\delta) \subseteq U$. An open set is one for which every point of the set is an interior point. Closed sets are those which are complements of open sets. Thus H is closed means H^C is open.

Theorem 4.3.2 *The intersection of any finite collection of open sets is open. The union of any collection of open sets is open. The intersection of any collection of closed sets is closed and the union of any finite collection of closed sets is closed.*

Proof: To see that any union of open sets is open, note that every point of the union is in at least one of the open sets. Therefore, it is an interior point of that set and hence an interior point of the entire union.

Now let $\{U_1,\cdots,U_m\}$ be some open sets and suppose $\mathbf{p} \in \cap_{k=1}^m U_k$. Then there exists $r_k > 0$ such that $B(\mathbf{p},r_k) \subseteq U_k$. Let $0 < r \leq \min(r_1,r_2,\cdots,r_m)$. Then $B(\mathbf{p},r) \subseteq \cap_{k=1}^m U_k$ and so the finite intersection is open. Note that if the finite intersection is empty, there is nothing to prove because it is certainly true in this case that every point in the intersection is an interior point because there aren't any such points.

Suppose $\{H_1,\cdots,H_m\}$ is a finite set of closed sets. Then $\cup_{k=1}^m H_k$ is closed if its complement is open. However, from DeMorgan's laws,

$$(\cup_{k=1}^m H_k)^C = \cap_{k=1}^m H_k^C,$$

a finite intersection of open sets which is open by what was just shown.

Next let $\mathcal{C}$ be some collection of closed sets. Then

$$(\cap\mathcal{C})^C = \cup\left\{H^C : H \in \mathcal{C}\right\},$$

a union of open sets which is therefore open by the first part of the proof. Thus $\cap\mathcal{C}$ is closed. This proves the theorem. ■

Next there is the concept of a limit point which gives another way of characterizing closed sets.

Definition 4.3.3 *Let A be any nonempty set and let $\mathbf{x}$ be a point. Then $\mathbf{x}$ is said to be a limit point of A if for every $r > 0, B(\mathbf{x},r)$ contains a point of A which is not equal to $\mathbf{x}$.*

Example 4.3.4 *Consider $A = B(\mathbf{x},\delta)$, an open ball in a normed vector space. Then every point of $B(\mathbf{x},\delta)$ is a limit point. There are more general situations than normed vector spaces in which this assertion is false.*

If $\mathbf{z} \in B(\mathbf{x},\delta)$, consider $\mathbf{z}+\frac{1}{k}(\mathbf{x}-\mathbf{z}) \equiv \mathbf{w}_k$ for $k \in \mathbb{N}$. Then

$$\begin{aligned} ||\mathbf{w}_k - \mathbf{x}|| &= \left|\left|\mathbf{z}+\frac{1}{k}(\mathbf{x}-\mathbf{z})-\mathbf{x}\right|\right| \\ &= \left|\left|\left(1-\frac{1}{k}\right)\mathbf{z}-\left(1-\frac{1}{k}\right)\mathbf{x}\right|\right| \\ &= \frac{k-1}{k}||\mathbf{z}-\mathbf{x}|| < \delta \end{aligned}$$

and also

$$||\mathbf{w}_k - \mathbf{z}|| \leq \frac{1}{k} ||\mathbf{x} - \mathbf{z}|| < \delta/k$$

so $\mathbf{w}_k \to \mathbf{z}$. Furthermore, the $\mathbf{w}_k$ are distinct. Thus $\mathbf{z}$ is a limit point of A as claimed. This is because every ball containing $\mathbf{z}$ contains infinitely many of the $\mathbf{w}_k$ and since they are all distinct, they can't all be equal to $\mathbf{z}$.

Similarly, the following holds in any normed vector space.

Theorem 4.3.5 *Let A be a nonempty set in V, a normed vector space. A point $\mathbf{a}$ is a limit point of A if and only if there exists a sequence of distinct points of A, $\{\mathbf{a}_n\}$ which converges to $\mathbf{a}$. Also a nonempty set, A is closed if and only if it contains all its limit points.*

Proof: Suppose first $\mathbf{a}$ is a limit point of A. There exists $\mathbf{a}_1 \in B(\mathbf{a}, 1) \cap A$ such that $\mathbf{a}_1 \neq \mathbf{a}$. Now supposing distinct points, $\mathbf{a}_1, \cdots, \mathbf{a}_n$ have been chosen such that none are equal to $\mathbf{a}$ and for each $k \leq n$, $\mathbf{a}_k \in B(\mathbf{a}, 1/k)$, let

$$0 < r_{n+1} < \min\left\{\frac{1}{n+1}, ||\mathbf{a} - \mathbf{a}_1||, \cdots, ||\mathbf{a} - \mathbf{a}_n||\right\}.$$

Then there exists $\mathbf{a}_{n+1} \in B(\mathbf{a}, r_{n+1}) \cap A$ with $\mathbf{a}_{n+1} \neq \mathbf{a}$. Because of the definition of r_{n+1}, $\mathbf{a}_{n+1}$ is not equal to any of the other $\mathbf{a}_k$ for $k < n+1$. Also since $||\mathbf{a} - \mathbf{a}_m|| < 1/m$, it follows $\lim_{m\to\infty} \mathbf{a}_m = \mathbf{a}$. Conversely, if there exists a sequence of distinct points of A converging to $\mathbf{a}$, then $B(\mathbf{a}, r)$ contains all $\mathbf{a}_n$ for n large enough. Thus $B(\mathbf{a}, r)$ contains infinitely many points of A since all are distinct. Thus at least one of them is not equal to $\mathbf{a}$. This establishes the first part of the theorem.

Now consider the second claim. If A is closed then it is the complement of an open set. Since A^C is open, it follows that if $\mathbf{a} \in A^C$, then there exists $\delta > 0$ such that $B(\mathbf{a}, \delta) \subseteq A^C$ and so no point of A^C can be a limit point of A. In other words, every limit point of A must be in A. Conversely, suppose A contains all its limit points. Then A^C does not contain any limit points of A. It also contains no points of A. Therefore, if $\mathbf{a} \in A^C$, since it is not a limit point of A, there exists $\delta > 0$ such that $B(\mathbf{a}, \delta)$ contains no points of A different than $\mathbf{a}$. However, $\mathbf{a}$ itself is not in A because $\mathbf{a} \in A^C$. Therefore, $B(\mathbf{a}, \delta)$ is entirely contained in A^C. Since $\mathbf{a} \in A^C$ was arbitrary, this shows every point of A^C is an interior point and so A^C is open. This proves the theorem. ■

Closed subsets of sequentially compact sets are sequentially compact.

Theorem 4.3.6 *If K is a sequentially compact set in a normed vector space and if H is a closed subset of K then H is sequentially compact.*

Proof: Let $\{\mathbf{x}_n\} \subseteq H$. Then since K is sequentially compact, there is a subsequence, $\{\mathbf{x}_{n_k}\}$ which converges to a point, $\mathbf{x} \in K$. If $\mathbf{x} \notin H$, then since H^C is open, it follows there exists $B(\mathbf{x}, r)$ such that this open ball contains no points of H. However, this is a contradiction to having $\mathbf{x}_{n_k} \to \mathbf{x}$ which requires $\mathbf{x}_{n_k} \in B(\mathbf{x}, r)$ for all k large enough. Thus $\mathbf{x} \in H$ and this has shown H is sequentially compact.

Definition 4.3.7 *A set $S \subseteq V$, a normed vector space is bounded if there is some $r > 0$ such that $S \subseteq B(\mathbf{0}, r)$.*

Theorem 4.3.8 *Every closed and bounded set in $\mathbb{F}^n$ is sequentially compact. Conversely, every sequentially compact set in $\mathbb{F}^n$ is closed and bounded.*

Proof: Let H be a closed and bounded set in $\mathbb{F}^n$. Then $H \subseteq B(\mathbf{0}, r)$ for some r. Therefore, if $\mathbf{x} \in H, \mathbf{x} = (x_1, \cdots, x_n)$, it must be that

$$\sqrt{\sum_{i=1}^{n} |x_i|^2} < r$$

and so each $x_i \in [-r, r] + i[-r, r] \equiv R_r$, a sequentially compact set by Corollary 4.2.5. Thus H is a closed subset of

$$\prod_{i=1}^{n} R_r$$

which is a sequentially compact set by Theorem 4.2.4. Therefore, by Theorem 4.3.6 it follows H is sequentially compact.

Conversely, suppose K is a sequentially compact set in $\mathbb{F}^n$. If it is not bounded, then there exists a sequence, $\{\mathbf{k}_m\}$ such that $\mathbf{k}_m \in K$ but $\mathbf{k}_m \notin B(\mathbf{0}, m)$ for $m = 1, 2, \cdots$. However, this sequence cannot have any convergent subsequence because if $\mathbf{k}_{m_k} \to \mathbf{k}$, then for large enough m, $\mathbf{k} \in B(\mathbf{0}, m) \subseteq D(\mathbf{0}, m)$ and $\mathbf{k}_{m_k} \in B(\mathbf{0}, m)^C$ for all k large enough and this is a contradiction because there can only be finitely many points of the sequence in $B(\mathbf{0}, m)$. If K is not closed, then it is missing a limit point. Say $\mathbf{k}_\infty$ is a limit point of K which is not in K. Pick $\mathbf{k}_m \in B\left(\mathbf{k}_\infty, \frac{1}{m}\right)$. Then $\{\mathbf{k}_m\}$ converges to $\mathbf{k}_\infty$ and so every subsequence also converges to $\mathbf{k}_\infty$ by Theorem 4.1.6. Thus there is no point of K which is a limit of some subsequence of $\{\mathbf{k}_m\}$, a contradiction. This proves the theorem. ■

What are some examples of closed and bounded sets in a general normed vector space and more specifically $\mathbb{F}^n$?

Proposition 4.3.9 *Let $D(\mathbf{z}, r)$ denote the set of points,*

$$\{\mathbf{w} \in V : ||\mathbf{w} - \mathbf{z}|| \leq r\}$$

Then $D(\mathbf{z}, r)$ is closed and bounded. Also, let $S(\mathbf{z}, r)$ denote the set of points

$$\{\mathbf{w} \in V : ||\mathbf{w} - \mathbf{z}|| = r\}$$

Then $S(\mathbf{z}, r)$ is closed and bounded. It follows that if $V = \mathbb{F}^n$, then these sets are sequentially compact.

Proof: First note $D(\mathbf{z}, r)$ is bounded because

$$D(\mathbf{z}, r) \subseteq B(\mathbf{0}, ||\mathbf{z}|| + 2r)$$

Here is why. Let $\mathbf{x} \in D(\mathbf{z}, r)$. Then $||\mathbf{x} - \mathbf{z}|| \leq r$ and so

$$||\mathbf{x}|| \leq ||\mathbf{x} - \mathbf{z}|| + ||\mathbf{z}|| \leq r + ||\mathbf{z}|| < 2r + ||\mathbf{z}||.$$

It remains to verify it is closed. Suppose then that $\mathbf{y} \notin D(\mathbf{z}, r)$. This means $||\mathbf{y} - \mathbf{z}|| > r$. Consider the open ball $B(\mathbf{y}, ||\mathbf{y} - \mathbf{z}|| - r)$. If $\mathbf{x} \in B(\mathbf{y}, ||\mathbf{y} - \mathbf{z}|| - r)$, then

$$||\mathbf{x} - \mathbf{y}|| < ||\mathbf{y} - \mathbf{z}|| - r$$

and so by the triangle inequality,

$$||\mathbf{z} - \mathbf{x}|| \geq ||\mathbf{z} - \mathbf{y}|| - ||\mathbf{y} - \mathbf{x}|| > ||\mathbf{x} - \mathbf{y}|| + r - ||\mathbf{x} - \mathbf{y}|| = r$$

Thus the complement of $D(\mathbf{z}, r)$ is open and so $D(\mathbf{z}, r)$ is closed.

For the second type of set, note $S(\mathbf{z}, r)^C = B(\mathbf{z}, r) \cup D(\mathbf{z}, r)^C$, the union of two open sets which by Theorem 4.3.2 is open. Therefore, $S(\mathbf{z}, r)$ is a closed set which is clearly bounded because $S(\mathbf{z}, r) \subseteq D(\mathbf{z}, r)$.

4.4 Cauchy Sequences And Completeness

The concept of completeness is that every Cauchy sequence converges. Cauchy sequences are those sequences which have the property that ultimately the terms of the sequence are bunching up. More precisely,

Definition 4.4.1 *$\{\mathbf{a}_n\}$ is a Cauchy sequence in a normed vector space, V if for all $\varepsilon > 0$, there exists n_ε such that whenever $n, m \geq n_\varepsilon$,*

$$||\mathbf{a}_n - \mathbf{a}_m|| < \varepsilon.$$

Theorem 4.4.2 *The set of terms (values) of a Cauchy sequence in a normed vector space V is bounded.*

Proof: Let $\varepsilon = 1$ in the definition of a Cauchy sequence and let $n > n_1$. Then from the definition,

$$||\mathbf{a}_n - \mathbf{a}_{n_1}|| < 1.$$

It follows that for all $n > n_1$,

$$||\mathbf{a}_n|| < 1 + ||\mathbf{a}_{n_1}||.$$

Therefore, for all n,

$$||\mathbf{a}_n|| \leq 1 + ||\mathbf{a}_{n_1}|| + \sum_{k=1}^{n_1} ||\mathbf{a}_k||.$$

This proves the theorem. ■

Theorem 4.4.3 *If a sequence $\{\mathbf{a}_n\}$ in V, a normed vector space converges, then the sequence is a Cauchy sequence.*

Proof: Let $\varepsilon > 0$ be given and suppose $\mathbf{a}_n \to \mathbf{a}$. Then from the definition of convergence, there exists n_ε such that if $n > n_\varepsilon$, it follows that

$$||\mathbf{a}_n - \mathbf{a}|| < \frac{\varepsilon}{2}$$

Therefore, if $m, n \geq n_\varepsilon + 1$, it follows that

$$||\mathbf{a}_n - \mathbf{a}_m|| \leq ||\mathbf{a}_n - \mathbf{a}|| + ||\mathbf{a} - \mathbf{a}_m|| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

showing that, since $\varepsilon > 0$ is arbitrary, $\{\mathbf{a}_n\}$ is a Cauchy sequence.

The following theorem is very useful. It is identical to an earlier theorem. All that is required is to put things in bold face to indicate they are vectors.

Theorem 4.4.4 *Suppose $\{\mathbf{a}_n\}$ is a Cauchy sequence in any normed vector space and there exists a subsequence, $\{\mathbf{a}_{n_k}\}$ which converges to $\mathbf{a}$. Then $\{\mathbf{a}_n\}$ also converges to $\mathbf{a}$.*

Proof: Let $\varepsilon > 0$ be given. There exists N such that if $m, n > N$, then

$$||\mathbf{a}_m - \mathbf{a}_n|| < \varepsilon/2.$$

Also there exists K such that if $k > K$, then

$$||\mathbf{a} - \mathbf{a}_{n_k}|| < \varepsilon/2.$$

Then let $k > \max(K, N)$. Then for such k,

$$\begin{aligned} ||\mathbf{a}_k - \mathbf{a}|| &\leq ||\mathbf{a}_k - \mathbf{a}_{n_k}|| + ||\mathbf{a}_{n_k} - \mathbf{a}|| \\ &< \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

This proves the theorem. ■

Definition 4.4.5 *If V is a normed vector space having the property that every Cauchy sequence converges, then V is called complete. It is also referred to as a Banach space.*

Example 4.4.6 $\mathbb{R}$ *is given to be complete. This is a fundamental axiom on which calculus is developed.*

Given $\mathbb{R}$ is complete, the following lemma is easily obtained.

Lemma 4.4.7 $\mathbb{C}$ *is complete.*

Proof: Let $\{x_k + iy_k\}_{k=1}^{\infty}$ be a Cauchy sequence in $\mathbb{C}$. This requires $\{x_k\}$ and $\{y_k\}$ are both Cauchy sequences in $\mathbb{R}$. This follows from the obvious estimates

$$|x_k - x_m|, |y_k - y_m| \leq |(x_k + iy_k) - (x_m + iy_m)|.$$

By completeness of $\mathbb{R}$ there exists $x \in \mathbb{R}$ such that $x_k \to x$ and similarly there exists $y \in \mathbb{R}$ such that $y_k \to y$. Therefore, since

$$\begin{aligned} |(x_k + iy_k) - (x + iy)| &\leq \sqrt{(x_k - x)^2 + (y_k - y)^2} \\ &\leq |x_k - x| + |y_k - y| \end{aligned}$$

it follows $(x_k + iy_k) \to (x + iy)$.■

A simple generalization of this idea yields the following theorem.

Theorem 4.4.8 $\mathbb{F}^n$ *is complete.*

Proof: By 4.4.7, $\mathbb{F}$ is complete. Now let $\{\mathbf{a}_m\}$ be a Cauchy sequence in $\mathbb{F}^n$. Then by the definition of the norm

$$\left|a_m^j - a_k^j\right| \leq |\mathbf{a}_m - \mathbf{a}_k|$$

where a_m^j denotes the j^{th} component of $\mathbf{a}_m$. Thus for each $j = 1, 2, \cdots, n$, $\left\{a_m^j\right\}_{m=1}^{\infty}$ is a Cauchy sequence. It follows from Theorem 4.4.7, the completeness of $\mathbb{F}$, there exists a^j such that

$$\lim_{m \to \infty} a_m^j = a^j$$

Theorem 4.1.5 implies that $\lim_{m \to \infty} \mathbf{a}_m = \mathbf{a}$ where

$$\mathbf{a} = \left(a^1, \cdots, a^n\right).$$

This proves the theorem. ■

4.5 Shrinking Diameters

It is useful to consider another version of the nested interval lemma. This involves a sequence of sets such that set $(n+1)$ is contained in set n and such that their diameters converge to 0. It turns out that if the sets are also closed, then often there exists a unique point in all of them.

Definition 4.5.1 *Let S be a nonempty set in a normed vector space, V. Then* $\operatorname{diam}(S)$ *is defined as*

$$\operatorname{diam}(S) \equiv \sup\{||\mathbf{x} - \mathbf{y}|| : \mathbf{x}, \mathbf{y} \in S\}.$$

This is called the diameter of S.

Theorem 4.5.2 *Let $\{F_n\}_{n=1}^{\infty}$ be a sequence of closed sets in $\mathbb{F}^n$ such that*

$$\lim_{n\to\infty} \operatorname{diam}(F_n) = 0$$

and $F_n \supseteq F_{n+1}$ for each n. Then there exists a unique $\mathbf{p} \in \cap_{k=1}^{\infty} F_k$.

Proof: Pick $\mathbf{p}_k \in F_k$. This is always possible because by assumption each set is nonempty. Then $\{\mathbf{p}_k\}_{k=m}^{\infty} \subseteq F_m$ and since the diameters converge to 0, it follows $\{\mathbf{p}_k\}$ is a Cauchy sequence. Therefore, it converges to a point, $\mathbf{p}$ by completeness of $\mathbb{F}^n$discussed in Theorem 4.4.8. Since each F_k is closed, $\mathbf{p} \in F_k$ for all k. This is because it is a limit of a sequence of points only finitely many of which are not in the closed set F_k. Therefore, $\mathbf{p} \in \cap_{k=1}^{\infty} F_k$. If $\mathbf{q} \in \cap_{k=1}^{\infty} F_k$, then since both $\mathbf{p}, \mathbf{q} \in F_k$,

$$|\mathbf{p} - \mathbf{q}| \leq \operatorname{diam}(F_k).$$

It follows since these diameters converge to 0, $|\mathbf{p} - \mathbf{q}| \leq \varepsilon$ for every ε. Hence $p = q$. This proves the theorem. ■

A sequence of sets $\{G_n\}$ which satisfies $G_n \supseteq G_{n+1}$ for all n is called a nested sequence of sets.

4.6 Exercises

1. For a nonempty set, S in a normed vector space, V, define a function

$$\mathbf{x} \to \operatorname{dist}(\mathbf{x},S) \equiv \inf\{||\mathbf{x} - \mathbf{y}|| : \mathbf{y} \in S\}.$$

   Show

$$||\operatorname{dist}(\mathbf{x}, S) - \operatorname{dist}(\mathbf{y}, S)|| \leq ||\mathbf{x} - \mathbf{y}||.$$

2. Let A be a nonempty set in $\mathbb{F}^n$ or more generally in a normed vector space. Define the closure of A to equal the intersection of all closed sets which contain A. This is usually denoted by $\overline{A}$. Show $\overline{A} = A \cup A'$ where A' consists of the set of limit points of A. Also explain why $\overline{A}$ is closed.

3. The interior of a set was defined above. Tell why the interior of a set is always an open set. The interior of a set A is sometimes denoted by A^0.

4. Given an example of a set A whose interior is empty but whose closure is all of $\mathbb{R}^n$.

5. A point, p is said to be in the boundary of a nonempty set, A if for every $r > 0$, $B(p, r)$ contains points of A as well as points of A^C. Sometimes this is denoted as ∂A. In a normed vector space, is it always the case that $A \cup \partial A = \overline{A}$? Prove or disprove.

6. Give an example of a finite dimensional normed vector space where the field of scalars is the rational numbers which is not complete.

7. Explain why as far as the theorems of this chapter are concerned, $\mathbb{C}^n$ is essentially the same as $\mathbb{R}^{2n}$.

8. A set, $A \subseteq \mathbb{R}^n$ is said to be convex if whenever $\mathbf{x}, \mathbf{y} \in A$ it follows $t\mathbf{x} + (1-t)\mathbf{y} \in A$ whenever $t \in [0, 1]$. Show $B(\mathbf{z}, r)$ is convex. Also show $D(\mathbf{z},r)$ is convex. If A is convex, does it follow $\overline{A}$ is convex? Explain why or why not.

9. Let A be any nonempty subset of $\mathbb{R}^n$. The convex hull of A, usually denoted by $\mathrm{co}(A)$ is defined as the set of all convex combinations of points in A. A convex combination is of the form $\sum_{k=1}^{p} t_k \mathbf{a}_k$ where each $t_k \geq 0$ and $\sum_k t_k = 1$. Note that p can be any finite number. Show $\mathrm{co}(A)$ is convex.

10. Suppose $A \subseteq \mathbb{R}^n$ and $\mathbf{z} \in \mathrm{co}(A)$. Thus $\mathbf{z} = \sum_{k=1}^{p} t_k \mathbf{a}_k$ for $t_k \geq 0$ and $\sum_k t_k = 1$. Show there exists $n+1$ of the points $\{\mathbf{a}_1, \cdots, \mathbf{a}_p\}$ such that $\mathbf{z}$ is a convex combination of these $n+1$ points. **Hint:** Show that if $p > n+1$ then the vectors $\{\mathbf{a}_k - \mathbf{a}_1\}_{k=2}^{p}$ must be linearly dependent. Conclude from this the existence of scalars $\{\alpha_i\}$ such that $\sum_{i=1}^{p} \alpha_i \mathbf{a}_i = \mathbf{0}$. Now for $s \in \mathbb{R}$, $\mathbf{z} = \sum_{k=1}^{p} (t_k + s\alpha_i)\, \mathbf{a}_k$. Consider small s and adjust till one or more of the $t_k + s\alpha_k$ vanish. Now you are in the same situation as before but with only $p-1$ of the $\mathbf{a}_k$. Repeat the argument till you end up with only $n+1$ at which time you can't repeat again.

11. Show that any uncountable set of points in $\mathbb{F}^n$ must have a limit point.

12. Let V be any finite dimensional vector space having a basis $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$. For $\mathbf{x} \in V$, let

$$\mathbf{x} = \sum_{k=1}^{n} x_k \mathbf{v}_k$$

so that the scalars, x_k are the components of $\mathbf{x}$ with respect to the given basis. Define for $\mathbf{x}, \mathbf{y} \in V$

$$(\mathbf{x} \cdot \mathbf{y}) \equiv \sum_{i=1}^{n} x_i \overline{y_i}$$

Show this is a dot product for V satisfying all the axioms of a dot product presented earlier.

13. In the context of Problem 12 let $|\mathbf{x}|$ denote the norm of $\mathbf{x}$ which is produced by this inner product and suppose $||\cdot||$ is some other norm on V. Thus

$$|\mathbf{x}| \equiv \left(\sum_i |x_i|^2 \right)^{1/2}$$

where

$$\mathbf{x} = \sum_k x_k \mathbf{v}_k. \tag{4.8}$$

Show there exist positive numbers $\delta < \Delta$ independent of $\mathbf{x}$ such that

$$\delta\, |\mathbf{x}| \leq ||\mathbf{x}|| \leq \Delta\, |\mathbf{x}|$$

This is referred to by saying the two norms are equivalent. **Hint:** The top half is easy using the Cauchy Schwarz inequality. The bottom half is somewhat harder. Argue that if it is not so, there exists a sequence $\{\mathbf{x}_k\}$ such that $|\mathbf{x}_k| = 1$ but $k^{-1} |\mathbf{x}_k| = k^{-1} \geq ||\mathbf{x}_k||$ and then note the vector of components of $\mathbf{x}_k$ is on $S(\mathbf{0}, 1)$ which was shown to be sequentially compact. Pass to a limit in 4.8 and use the assumed inequality to get a contradiction to $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ being a basis.

14. It was shown above that in $\mathbb{F}^n$, the sequentially compact sets are exactly those which are closed and bounded. Show that in any finite dimensional normed vector space, V the closed and bounded sets are those which are sequentially compact.

15. Two norms on a finite dimensional vector space, $||\cdot||_1$ and $||\cdot||_2$ are said to be equivalent if there exist positive numbers $\delta < \Delta$ such that

$$\delta ||\mathbf{x}||_1 \leq ||\mathbf{x}||_2 \leq \Delta ||\mathbf{x}_1||_1 .$$

Show the statement that two norms are equivalent is an equivalence relation. Explain using the result of Problem 13 why any two norms on a finite dimensional vector space are equivalent.

16. A normed vector space, V is separable if there is a countable set $\{\mathbf{w}_k\}_{k=1}^{\infty}$ such that whenever $B(\mathbf{x}, \delta)$ is an open ball in V, there exists some $\mathbf{w}_k$ in this open ball. Show that $\mathbb{F}^n$ is separable. This set of points is called a countable dense set.

17. Let V be any normed vector space with norm $||\cdot||$. Using Problem 13 show that V is separable.

18. Suppose V is a normed vector space. Show there exists a countable set of open balls $\mathcal{B} \equiv \{B(\mathbf{x}_k, r_k)\}_{k=1}^{\infty}$ having the remarkable property that any open set, U is the union of some subset of $\mathcal{B}$. This collection of balls is called a countable basis. **Hint:** Use Problem 17 to get a countable dense dense set of points, $\{\mathbf{x}_k\}_{k=1}^{\infty}$ and then consider balls of the form $B\left(\mathbf{x}_k, \frac{1}{r}\right)$ where $r \in \mathbb{N}$. Show this collection of balls is countable and then show it has the remarkable property mentioned.

19. Suppose S is any nonempty set in V a finite dimensional normed vector space. Suppose $\mathcal{C}$ is a set of open sets such that $\cup\mathcal{C} \supseteq S$. (Such a collection of sets is called an open cover.) Show using Problem 18 that there are countably many sets from $\mathcal{C}$, $\{U_k\}_{k=1}^{\infty}$ such that $S \subseteq \cup_{k=1}^{\infty} U_k$. This is called the Lindeloff property when every open cover can be reduced to a countable sub cover.

20. A set, H in a normed vector space is said to be compact if whenever $\mathcal{C}$ is a set of open sets such that $\cup\mathcal{C} \supseteq H$, there are finitely many sets of $\mathcal{C}$, $\{U_1, \cdots, U_p\}$ such that

$$H \subseteq \cup_{i=1}^{p} U_i.$$

Show using Problem 19 that if a set in a normed vector space is sequentially compact, then it must be compact. Next show using Problem 14 that a set in a normed vector space is compact if and only if it is closed and bounded. Explain why the sets which are compact, closed and bounded, and sequentially compact are the same sets in any finite dimensional normed vector space

Chapter 5

Continuous Functions

Continuous functions are defined as they are for a function of one variable.

Definition 5.0.1 *Let V, W be normed vector spaces. A function f: $D(\mathbf{f}) \subseteq V \to W$ is continuous at $\mathbf{x} \in D(\mathbf{f})$ if for each $\varepsilon > 0$ there exists $\delta > 0$ such that whenever $\mathbf{y} \in D(\mathbf{f})$ and*

$$||\mathbf{y} - \mathbf{x}||_V < \delta$$

it follows that

$$||\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})||_W < \varepsilon.$$

A function, f is continuous if it is continuous at every point of $D(\mathbf{f})$.

There is a theorem which makes it easier to verify certain functions are continuous without having to always go to the above definition. The statement of this theorem is purposely just a little vague. Some of these things tend to hold in almost any context, certainly for any normed vector space.

Theorem 5.0.2 *The following assertions are valid*

1. *The function, $af+bg$ is continuous at x when f, g are continuous at $x \in D(\mathbf{f}) \cap D(\mathbf{g})$ and $a, b \in \mathbb{F}$.*
2. *If and $\mathbf{f}$ and $\mathbf{g}$ have values in $\mathbb{F}^n$ and they are each continuous at $\mathbf{x}$, then $\mathbf{f} \cdot \mathbf{g}$ is continuous at $\mathbf{x}$. If g has values in $\mathbb{F}$ and $g(\mathbf{x}) \neq 0$ with g continuous, then $\mathbf{f}/g$ is continuous at $\mathbf{x}$.*
3. *If $\mathbf{f}$ is continuous at $\mathbf{x}$, $\mathbf{f}(\mathbf{x}) \in D(\mathbf{g})$, and $\mathbf{g}$ is continuous at $\mathbf{f}(\mathbf{x})$,then $\mathbf{g} \circ \mathbf{f}$ is continuous at $\mathbf{x}$.*
4. *If V is any normed vector space, the function $\mathbf{f} : V \to \mathbb{R}$, given by $\mathbf{f}(\mathbf{x}) = ||\mathbf{x}||$ is continuous.*
5. *$\mathbf{f}$ is continuous at every point of V if and only if whenever U is an open set in $W, \mathbf{f}^{-1}(W)$ is open.*

Proof: First consider 1.) Let $\varepsilon > 0$ be given. By assumption, there exist $\delta_1 > 0$ such that whenever $|\mathbf{x} - \mathbf{y}| < \delta_1$, it follows $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < \frac{\varepsilon}{2(|a|+|b|+1)}$ and there exists $\delta_2 > 0$ such that whenever $|\mathbf{x} - \mathbf{y}| < \delta_2$, it follows that $|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})| < \frac{\varepsilon}{2(|a|+|b|+1)}$. Then let $0 < \delta \leq \min(\delta_1, \delta_2)$. If $|\mathbf{x} - \mathbf{y}| < \delta$, then everything happens at once. Therefore, using the triangle inequality

$$|a\mathbf{f}(\mathbf{x}) + b\mathbf{f}(\mathbf{x}) - (a\mathbf{g}(\mathbf{y}) + b\mathbf{g}(\mathbf{y}))|$$

$$\leq |a| |\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| + |b| |\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})|$$
$$< |a| \left(\frac{\varepsilon}{2(|a| + |b| + 1)} \right) + |b| \left(\frac{\varepsilon}{2(|a| + |b| + 1)} \right) < \varepsilon.$$

Now consider 2.) There exists $\delta_1 > 0$ such that if $|\mathbf{y} - \mathbf{x}| < \delta_1$, then $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < 1$. Therefore, for such $\mathbf{y}$,

$$|\mathbf{f}(\mathbf{y})| < 1 + |\mathbf{f}(\mathbf{x})|.$$

It follows that for such $\mathbf{y}$,

$$|\mathbf{f} \cdot \mathbf{g}(\mathbf{x}) - \mathbf{f} \cdot \mathbf{g}(\mathbf{y})| \leq |\mathbf{f}(\mathbf{x}) \cdot \mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{x}) \cdot \mathbf{f}(\mathbf{y})| + |\mathbf{g}(\mathbf{x}) \cdot \mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{y}) \cdot \mathbf{g}(\mathbf{y})|$$

$$\leq |\mathbf{g}(\mathbf{x})| |\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| + |\mathbf{f}(\mathbf{y})| |\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})|$$
$$\leq (1 + |\mathbf{g}(\mathbf{x})| + |\mathbf{f}(\mathbf{y})|) [|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})| + |\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})|]$$
$$\leq (2 + |\mathbf{g}(\mathbf{x})| + |\mathbf{f}(\mathbf{x})|) [|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})| + |\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})|]$$

Now let $\varepsilon > 0$ be given. There exists δ_2 such that if $|\mathbf{x} - \mathbf{y}| < \delta_2$, then

$$|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})| < \frac{\varepsilon}{2(2 + |\mathbf{g}(\mathbf{x})| + |\mathbf{f}(\mathbf{x})|)},$$

and there exists δ_3 such that if $|\mathbf{x} - \mathbf{y}| < \delta_3$, then

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < \frac{\varepsilon}{2(2 + |\mathbf{g}(\mathbf{x})| + |\mathbf{f}(\mathbf{x})|)}$$

Now let $0 < \delta \leq \min(\delta_1, \delta_2, \delta_3)$. Then if $|\mathbf{x} - \mathbf{y}| < \delta$, all the above hold at once and so

$$|\mathbf{f} \cdot \mathbf{g}(\mathbf{x}) - \mathbf{f} \cdot \mathbf{g}(\mathbf{y})| \leq$$

$$(2 + |\mathbf{g}(\mathbf{x})| + |\mathbf{f}(\mathbf{x})|) [|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})| + |\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})|]$$
$$< (2 + |\mathbf{g}(\mathbf{x})| + |\mathbf{f}(\mathbf{x})|) \left(\frac{\varepsilon}{2(2 + |\mathbf{g}(\mathbf{x})| + |\mathbf{f}(\mathbf{x})|)} + \frac{\varepsilon}{2(2 + |\mathbf{g}(\mathbf{x})| + |\mathbf{f}(\mathbf{x})|)} \right) = \varepsilon.$$

This proves the first part of 2.) To obtain the second part, let δ_1 be as described above and let $\delta_0 > 0$ be such that for $|\mathbf{x} - \mathbf{y}| < \delta_0$,

$$|g(\mathbf{x}) - g(\mathbf{y})| < |g(\mathbf{x})| / 2$$

and so by the triangle inequality,

$$-|g(\mathbf{x})| / 2 \leq |g(\mathbf{y})| - |g(\mathbf{x})| \leq |g(\mathbf{x})| / 2$$

which implies $|g(\mathbf{y})| \geq |g(\mathbf{x})| / 2$, and $|g(\mathbf{y})| < 3 |g(\mathbf{x})| / 2$.

Then if $|\mathbf{x} - \mathbf{y}| < \min(\delta_0, \delta_1)$,

$$\left| \frac{\mathbf{f}(\mathbf{x})}{g(\mathbf{x})} - \frac{\mathbf{f}(\mathbf{y})}{g(\mathbf{y})} \right| = \left| \frac{\mathbf{f}(\mathbf{x}) g(\mathbf{y}) - \mathbf{f}(\mathbf{y}) g(\mathbf{x})}{g(\mathbf{x}) g(\mathbf{y})} \right|$$
$$\leq \frac{|\mathbf{f}(\mathbf{x}) g(\mathbf{y}) - \mathbf{f}(\mathbf{y}) g(\mathbf{x})|}{\left(\frac{|g(\mathbf{x})|^2}{2} \right)}$$
$$= \frac{2 |\mathbf{f}(\mathbf{x}) g(\mathbf{y}) - \mathbf{f}(\mathbf{y}) g(\mathbf{x})|}{|g(\mathbf{x})|^2}$$

$$\begin{aligned}
&\leq \frac{2}{|g(\mathbf{x})|^2}\left[|\mathbf{f}(\mathbf{x})\, g(\mathbf{y}) - \mathbf{f}(\mathbf{y})\, g(\mathbf{y}) + \mathbf{f}(\mathbf{y})\, g(\mathbf{y}) - \mathbf{f}(\mathbf{y})\, g(\mathbf{x})|\right] \\
&\leq \frac{2}{|g(\mathbf{x})|^2}\left[|g(\mathbf{y})|\,|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| + |\mathbf{f}(\mathbf{y})|\,|g(\mathbf{y}) - g(\mathbf{x})|\right] \\
&\leq \frac{2}{|g(\mathbf{x})|^2}\left[\frac{3}{2}|g(\mathbf{x})|\,|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| + (1 + |\mathbf{f}(\mathbf{x})|)\,|g(\mathbf{y}) - g(\mathbf{x})|\right] \\
&\leq \frac{2}{|g(\mathbf{x})|^2}(1 + 2|\mathbf{f}(\mathbf{x})| + 2|g(\mathbf{x})|)\left[|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| + |g(\mathbf{y}) - g(\mathbf{x})|\right] \\
&\equiv M\left[|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| + |g(\mathbf{y}) - g(\mathbf{x})|\right]
\end{aligned}$$

where M is defined by

$$M \equiv \frac{2}{|g(\mathbf{x})|^2}(1 + 2|\mathbf{f}(\mathbf{x})| + 2|g(\mathbf{x})|)$$

Now let δ_2 be such that if $|\mathbf{x} - \mathbf{y}| < \delta_2$, then

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < \frac{\varepsilon}{2}M^{-1}$$

and let δ_3 be such that if $|\mathbf{x} - \mathbf{y}| < \delta_3$, then

$$|g(\mathbf{y}) - g(\mathbf{x})| < \frac{\varepsilon}{2}M^{-1}.$$

Then if $0 < \delta \leq \min(\delta_0, \delta_1, \delta_2, \delta_3)$, and $|\mathbf{x} - \mathbf{y}| < \delta$, everything holds and

$$\begin{aligned}
\left|\frac{\mathbf{f}(\mathbf{x})}{g(\mathbf{x})} - \frac{\mathbf{f}(\mathbf{y})}{g(\mathbf{y})}\right| &\leq M\left[|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| + |g(\mathbf{y}) - g(\mathbf{x})|\right] \\
&< M\left[\frac{\varepsilon}{2}M^{-1} + \frac{\varepsilon}{2}M^{-1}\right] = \varepsilon.
\end{aligned}$$

This completes the proof of the second part of 2.)

Note that in these proofs no effort is made to find some sort of "best" δ. The problem is one which has a yes or a no answer. Either it is or it is not continuous.

Now consider 3.). If $\mathbf{f}$ is continuous at $\mathbf{x}$, $\mathbf{f}(\mathbf{x}) \in D(\mathbf{g})$, and $\mathbf{g}$ is continuous at $\mathbf{f}(\mathbf{x})$,then $\mathbf{g} \circ \mathbf{f}$ is continuous at $\mathbf{x}$. Let $\varepsilon > 0$ be given. Then there exists $\eta > 0$ such that if $|\mathbf{y} - \mathbf{f}(\mathbf{x})| < \eta$ and $\mathbf{y} \in D(\mathbf{g})$, it follows that $|\mathbf{g}(\mathbf{y}) - \mathbf{g}(\mathbf{f}(\mathbf{x}))| < \varepsilon$. From continuity of $\mathbf{f}$ at $\mathbf{x}$, there exists $\delta > 0$ such that if $|\mathbf{x} - \mathbf{z}| < \delta$ and $z \in D(\mathbf{f})$, then $|\mathbf{f}(\mathbf{z}) - \mathbf{f}(\mathbf{x})| < \eta$. Then if $|\mathbf{x} - \mathbf{z}| < \delta$ and $\mathbf{z} \in D(\mathbf{g} \circ \mathbf{f}) \subseteq D(\mathbf{f})$, all the above hold and so

$$|\mathbf{g}(\mathbf{f}(\mathbf{z})) - \mathbf{g}(\mathbf{f}(\mathbf{x}))| < \varepsilon.$$

This proves part 3.)

To verify part 4.), let $\varepsilon > 0$ be given and let $\delta = \varepsilon$. Then if $||\mathbf{x} - \mathbf{y}|| < \delta$, the triangle inequality implies

$$\begin{aligned}
|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| &= |\,||\mathbf{x}|| - ||\mathbf{y}||\,| \\
&\leq ||\mathbf{x} - \mathbf{y}|| < \delta = \varepsilon.
\end{aligned}$$

This proves part 4.)

Next consider 5.) Suppose first $\mathbf{f}$ is continuous. Let U be open and let $\mathbf{x} \in \mathbf{f}^{-1}(U)$. This means $\mathbf{f}(\mathbf{x}) \in U$. Since U is open, there exists $\varepsilon > 0$ such that $B(\mathbf{f}(\mathbf{x}), \varepsilon) \subseteq U$. By continuity, there exists $\delta > 0$ such that if $\mathbf{y} \in B(\mathbf{x}, \delta)$, then $\mathbf{f}(\mathbf{y}) \in B(\mathbf{f}(\mathbf{x}), \varepsilon)$ and so this shows $B(\mathbf{x},\delta) \subseteq \mathbf{f}^{-1}(U)$ which implies $\mathbf{f}^{-1}(U)$ is open since $\mathbf{x}$ is an arbitrary point

of $\mathbf{f}^{-1}(U)$. Next suppose the condition about inverse images of open sets are open. Then apply this condition to the open set $B(\mathbf{f}(\mathbf{x}),\varepsilon)$. The condition says $\mathbf{f}^{-1}(B(\mathbf{f}(\mathbf{x}),\varepsilon))$ is open and since $\mathbf{x}\in\mathbf{f}^{-1}(B(\mathbf{f}(\mathbf{x}),\varepsilon))$, it follows $\mathbf{x}$ is an interior point of $\mathbf{f}^{-1}(B(\mathbf{f}(\mathbf{x}),\varepsilon))$ so there exists $\delta>0$ such that $B(\mathbf{x},\delta)\subseteq\mathbf{f}^{-1}(B(\mathbf{f}(\mathbf{x}),\varepsilon))$. This says $\mathbf{f}(B(\mathbf{x},\delta))\subseteq B(\mathbf{f}(\mathbf{x}),\varepsilon)$. In other words, whenever $||\mathbf{y}-\mathbf{x}||<\delta, ||\mathbf{f}(\mathbf{y})-\mathbf{f}(\mathbf{x})||<\varepsilon$ which is the condition for continuity at the point $\mathbf{x}$. Since $\mathbf{x}$ is arbitrary, This proves the theorem. ■

5.1 Continuity And The Limit Of A Sequence

There is a very useful way of thinking of continuity in terms of limits of sequences found in the following theorem. In words, it says a function is continuous if it takes convergent sequences to convergent sequences whenever possible.

Theorem 5.1.1 *A function $\mathbf{f}:D(\mathbf{f})\to W$ is continuous at $\mathbf{x}\in D(\mathbf{f})$ if and only if, whenever $\mathbf{x}_n\to\mathbf{x}$ with $\mathbf{x}_n\in D(\mathbf{f})$, it follows $\mathbf{f}(\mathbf{x}_n)\to\mathbf{f}(\mathbf{x})$.*

Proof: Suppose first that $\mathbf{f}$ is continuous at $\mathbf{x}$ and let $\mathbf{x}_n\to\mathbf{x}$. Let $\varepsilon>0$ be given. By continuity, there exists $\delta>0$ such that if $||\mathbf{y}-\mathbf{x}||<\delta$, then $||\mathbf{f}(\mathbf{x})-\mathbf{f}(\mathbf{y})||<\varepsilon$. However, there exists n_δ such that if $n\geq n_\delta$, then $||\mathbf{x}_n-\mathbf{x}||<\delta$ and so for all n this large,

$$||\mathbf{f}(\mathbf{x})-\mathbf{f}(\mathbf{x}_n)||<\varepsilon$$

which shows $\mathbf{f}(\mathbf{x}_n)\to\mathbf{f}(\mathbf{x})$.

Now suppose the condition about taking convergent sequences to convergent sequences holds at $\mathbf{x}$. Suppose $\mathbf{f}$ fails to be continuous at $\mathbf{x}$. Then there exists $\varepsilon>0$ and $\mathbf{x}_n\in D(\mathbf{f})$ such that $||\mathbf{x}-\mathbf{x}_n||<\frac{1}{n}$, yet

$$||\mathbf{f}(\mathbf{x})-\mathbf{f}(\mathbf{x}_n)||\geq\varepsilon.$$

But this is clearly a contradiction because, although $\mathbf{x}_n\to\mathbf{x}$, $\mathbf{f}(\mathbf{x}_n)$ fails to converge to $\mathbf{f}(\mathbf{x})$. It follows $\mathbf{f}$ must be continuous after all. This proves the theorem. ■

Theorem 5.1.2 *Suppose $\mathbf{f}:D(\mathbf{f})\to\mathbb{R}$ is continuous at $\mathbf{x}\in D(\mathbf{f})$ and suppose*

$$||\mathbf{f}(\mathbf{x}_n)||\leq l\,(\geq l)$$

where $\{\mathbf{x}_n\}$ is a sequence of points of $D(\mathbf{f})$ which converges to $\mathbf{x}$. Then

$$||\mathbf{f}(\mathbf{x})||\leq l\,(\geq l).$$

Proof: Since $||\mathbf{f}(\mathbf{x}_n)||\leq l$ and $\mathbf{f}$ is continuous at $\mathbf{x}$, it follows from the triangle inequality, Theorem 4.1.8 and Theorem 5.1.1,

$$||\mathbf{f}(\mathbf{x})||=\lim_{n\to\infty}||\mathbf{f}(\mathbf{x}_n)||\leq l.$$

The other case is entirely similar. This proves the theorem. ■

Another very useful idea involves the automatic continuity of the inverse function under certain conditions.

Theorem 5.1.3 *Let K be a sequentially compact set and suppose $\mathbf{f}:K\to\mathbf{f}(K)$ is continuous and one to one. Then $\mathbf{f}^{-1}$ must also be continuous.*

Proof: Suppose $\mathbf{f}(\mathbf{k}_n) \to \mathbf{f}(\mathbf{k})$. Does it follow $\mathbf{k}_n \to \mathbf{k}$? If this does not happen, then there exists $\varepsilon > 0$ and a subsequence still denoted as $\{\mathbf{k}_n\}$ such that

$$|\mathbf{k}_n - \mathbf{k}| \geq \varepsilon \tag{5.1}$$

Now since K is compact, there exists a further subsequence, still denoted as $\{\mathbf{k}_n\}$ such that

$$\mathbf{k}_n \to \mathbf{k}' \in K$$

However, the continuity of $\mathbf{f}$ requires

$$\mathbf{f}(\mathbf{k}_n) \to \mathbf{f}(\mathbf{k}')$$

and so $\mathbf{f}(\mathbf{k}') = \mathbf{f}(\mathbf{k})$. Since $\mathbf{f}$ is one to one, this requires $\mathbf{k}' = \mathbf{k}$, a contradiction to 5.1. This proves the theorem. ■

5.2 The Extreme Values Theorem

The extreme values theorem says continuous functions achieve their maximum and minimum provided they are defined on a sequentially compact set.

The next theorem is known as the max min theorem or extreme value theorem.

Theorem 5.2.1 *Let $K \subseteq \mathbb{F}^n$ be sequentially compact. Thus K is closed and bounded, and let $f : K \to \mathbb{R}$ be continuous. Then f achieves its maximum and its minimum on K. This means there exist, $\mathbf{x}_1, \mathbf{x}_2 \in K$ such that for all $\mathbf{x} \in K$,*

$$f(\mathbf{x}_1) \leq f(\mathbf{x}) \leq f(\mathbf{x}_2).$$

Proof: Let $\lambda = \sup\{f(\mathbf{x}) : \mathbf{x} \in K\}$. Next let $\{\lambda_k\}$ be an increasing sequence which converges to λ but each $\lambda_k < \lambda$. Therefore, for each k, there exists $\mathbf{x}_k \in K$ such that

$$f(\mathbf{x}_k) > \lambda_k.$$

Since K is sequentially compact, there exists a subsequence, $\{\mathbf{x}_{k_l}\}$ such that $\lim_{l\to\infty} \mathbf{x}_{k_l} = \mathbf{x} \in K$. Then by continuity of f,

$$f(\mathbf{x}) = \lim_{l\to\infty} f(\mathbf{x}_{k_l}) \geq \lim_{l\to\infty} \lambda_{k_l} = \lambda$$

which shows f achieves its maximum on K. To see it achieves its minimum, you could repeat the argument with a minimizing sequence or else you could consider $-f$ and apply what was just shown to $-f$, $-f$ having its minimum when f has its maximum. This proves the theorem. ■

5.3 Connected Sets

Stated informally, connected sets are those which are in one piece. In order to define what is meant by this, I will first consider what it means for a set to not be in one piece.

Definition 5.3.1 *Let A be a nonempty subset of V a normed vector space. Then $\overline{A}$ is defined to be the intersection of all closed sets which contain A. This is called the closure of A. Note the whole space, V is one such closed set which contains A.*

Lemma 5.3.2 *Let A be a nonempty set in a normed vector space V. Then $\overline{A}$ is a closed set and*

$$\overline{A} = A \cup A'$$

where A' denotes the set of limit points of A.

Proof: First of all, denote by $\mathcal{C}$ the set of closed sets which contain A. Then

$$\overline{A} = \cap\mathcal{C}$$

and this will be closed if its complement is open. However,

$$\overline{A}^C = \cup\left\{H^C : H \in \mathcal{C}\right\}.$$

Each H^C is open and so the union of all these open sets must also be open. This is because if $\mathbf{x}$ is in this union, then it is in at least one of them. Hence it is an interior point of that one. But this implies it is an interior point of the union of them all which is an even larger set. Thus $\overline{A}$ is closed.

The interesting part is the next claim. First note that from the definition, $A \subseteq \overline{A}$ so if $\mathbf{x} \in A$, then $\mathbf{x} \in \overline{A}$. Now consider $\mathbf{y} \in A'$ but $\mathbf{y} \notin A$. If $\mathbf{y} \notin \overline{A}$, a closed set, then there exists $B(\mathbf{y}, r) \subseteq \overline{A}^C$. Thus $\mathbf{y}$ cannot be a limit point of A, a contradiction. Therefore,

$$A \cup A' \subseteq \overline{A}$$

Next suppose $\mathbf{x} \in \overline{A}$ and suppose $\mathbf{x} \notin A$. Then if $B(\mathbf{x}, r)$ contains no points of A different than $\mathbf{x}$, since $\mathbf{x}$ itself is not in A, it would follow that $B(\mathbf{x},r) \cap A = \emptyset$ and so recalling that open balls are open, $B(\mathbf{x}, r)^C$ is a closed set containing A so from the definition, it also contains $\overline{A}$ which is contrary to the assertion that $\mathbf{x} \in \overline{A}$. Hence if $\mathbf{x} \notin A$, then $\mathbf{x} \in A'$ and so

$$A \cup A' \supseteq \overline{A}$$

This proves the lemma. ■

Now that the closure of a set has been defined it is possible to define what is meant by a set being separated.

Definition 5.3.3 *A set, S in a normed vector space is separated if there exist sets A, B such that*

$$S = A \cup B,\ A, B \neq \emptyset,\ \text{and } \overline{A} \cap B = \overline{B} \cap A = \emptyset.$$

In this case, the sets A and B are said to separate S. A set is connected if it is not separated. Remember $\overline{A}$ denotes the closure of the set A.

Note that the concept of connected sets is defined in terms of what it is not. This makes it somewhat difficult to understand. One of the most important theorems about connected sets is the following.

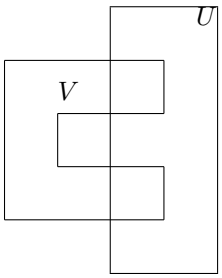
Theorem 5.3.4 *Suppose U and V are connected sets having nonempty intersection. Then $U \cup V$ is also connected.*

Proof: Suppose $U \cup V = A \cup B$ where $\overline{A} \cap B = \overline{B} \cap A = \emptyset$. Consider the sets $A \cap U$ and $B \cap U$. Since

$$\overline{(A \cap U)} \cap (B \cap U) = (A \cap U) \cap \left(\overline{B \cap U}\right) = \emptyset,$$

It follows one of these sets must be empty since otherwise, U would be separated. It follows that U is contained in either A or B. Similarly, V must be contained in either A or B. Since U and V have nonempty intersection, it follows that both V and U are contained in one of the sets A, B. Therefore, the other must be empty and this shows $U \cup V$ cannot be separated and is therefore, connected. ■

The intersection of connected sets is not necessarily connected as is shown by the following picture.



Theorem 5.3.5 *Let $\mathbf{f} : X \to Y$ be continuous where Y is a normed vector space and X is connected. Then $\mathbf{f}(X)$ is also connected.*

Proof: To do this you show $\mathbf{f}(X)$ is not separated. Suppose to the contrary that $\mathbf{f}(X) = A \cup B$ where A and B separate $\mathbf{f}(X)$. Then consider the sets $\mathbf{f}^{-1}(A)$ and $\mathbf{f}^{-1}(B)$. If $\mathbf{z} \in \mathbf{f}^{-1}(B)$, then $\mathbf{f}(\mathbf{z}) \in B$ and so $\mathbf{f}(\mathbf{z})$ is not a limit point of A. Therefore, there exists an open set, U containing $\mathbf{f}(\mathbf{z})$ such that $U \cap A = \emptyset$. But then, the continuity of $\mathbf{f}$ and Theorem 5.0.2 implies that $\mathbf{f}^{-1}(U)$ is an open set containing $\mathbf{z}$ such that $\mathbf{f}^{-1}(U) \cap \mathbf{f}^{-1}(A) = \emptyset$. Therefore, $\mathbf{f}^{-1}(B)$ contains no limit points of $\mathbf{f}^{-1}(A)$. Similar reasoning implies $\mathbf{f}^{-1}(A)$ contains no limit points of $\mathbf{f}^{-1}(B)$. It follows that X is separated by $\mathbf{f}^{-1}(A)$ and $\mathbf{f}^{-1}(B)$, contradicting the assumption that X was connected. ■

An arbitrary set can be written as a union of maximal connected sets called connected components. This is the concept of the next definition.

Definition 5.3.6 *Let S be a set and let $\mathbf{p} \in S$. Denote by $C_{\mathbf{p}}$ the union of all connected subsets of S which contain $\mathbf{p}$. This is called the connected component determined by $\mathbf{p}$.*

Theorem 5.3.7 *Let $C_{\mathbf{p}}$ be a connected component of a set S in a normed vector space. Then $C_{\mathbf{p}}$ is a connected set and if $C_{\mathbf{p}} \cap C_{\mathbf{q}} \neq \emptyset$, then $C_{\mathbf{p}} = C_{\mathbf{q}}$.*

Proof: Let $\mathcal{C}$ denote the connected subsets of S which contain $\mathbf{p}$. If $C_{\mathbf{p}} = A \cup B$ where

$$\overline{A} \cap B = \overline{B} \cap A = \emptyset,$$

then $\mathbf{p}$ is in one of A or B. Suppose without loss of generality $\mathbf{p} \in A$. Then every set of $\mathcal{C}$ must also be contained in A since otherwise, as in Theorem 5.3.4, the set would be separated. But this implies B is empty. Therefore, $C_{\mathbf{p}}$ is connected. From this, and Theorem 5.3.4, the second assertion of the theorem is proved.■

This shows the connected components of a set are equivalence classes and partition the set.

A set, I is an interval in $\mathbb{R}$ if and only if whenever $x, y \in I$ then $(x, y) \subseteq I$. The following theorem is about the connected sets in $\mathbb{R}$.

Theorem 5.3.8 *A set, C in $\mathbb{R}$ is connected if and only if C is an interval.*

Proof: Let C be connected. If C consists of a single point, p, there is nothing to prove. The interval is just $[p, p]$. Suppose $p < q$ and $p, q \in C$. You need to show $(p, q) \subseteq C$. If

$$x \in (p, q) \setminus C$$

let $C \cap (-\infty, x) \equiv A$, and $C \cap (x, \infty) \equiv B$. Then $C = A \cup B$ and the sets A and B separate C contrary to the assumption that C is connected.

Conversely, let I be an interval. Suppose I is separated by A and B. Pick $x \in A$ and $y \in B$. Suppose without loss of generality that $x < y$. Now define the set,

$$S \equiv \{t \in [x, y] : [x, t] \subseteq A\}$$

and let l be the least upper bound of S. Then $l \in \overline{A}$ so $l \notin B$ which implies $l \in A$. But if $l \notin \overline{B}$, then for some $\delta > 0$,

$$(l, l + \delta) \cap B = \emptyset$$

contradicting the definition of l as an upper bound for S. Therefore, $l \in \overline{B}$ which implies $l \notin A$ after all, a contradiction. It follows I must be connected. ■

This yields a generalization of the intermediate value theorem from one variable calculus.

Corollary 5.3.9 *Let E be a connected set in a normed vector space and suppose $f : E \to \mathbb{R}$ and that $y \in (f(e_1), f(e_2))$ where $e_i \in E$. Then there exists $e \in E$ such that $f(e) = y$.*

Proof: From Theorem 5.3.5, $f(E)$ is a connected subset of $\mathbb{R}$. By Theorem 5.3.8 $f(E)$ must be an interval. In particular, it must contain y. This proves the corollary. ■

The following theorem is a very useful description of the open sets in $\mathbb{R}$.

Theorem 5.3.10 *Let U be an open set in $\mathbb{R}$. Then there exist countably many disjoint open sets $\{(a_i, b_i)\}_{i=1}^{\infty}$ such that $U = \cup_{i=1}^{\infty} (a_i, b_i)$.*

Proof: Let $p \in U$ and let $z \in C_p$, the connected component determined by p. Since U is open, there exists, $\delta > 0$ such that $(z - \delta, z + \delta) \subseteq U$. It follows from Theorem 5.3.4 that

$$(z - \delta, z + \delta) \subseteq C_p.$$

This shows C_p is open. By Theorem 5.3.8, this shows C_p is an open interval, (a, b) where $a, b \in [-\infty, \infty]$. There are therefore at most countably many of these connected components because each must contain a rational number and the rational numbers are countable. Denote by $\{(a_i, b_i)\}_{i=1}^{\infty}$ the set of these connected components. This proves the theorem. ■

Definition 5.3.11 *A set E in a normed vector space is arcwise connected if for any two points, $\mathbf{p}, \mathbf{q} \in E$, there exists a closed interval, $[a, b]$ and a continuous function, $\gamma : [a, b] \to E$ such that $\boldsymbol{\gamma}(a) = p$ and $\boldsymbol{\gamma}(b) = q$.*

An example of an arcwise connected topological space would be any subset of $\mathbb{R}^n$ which is the continuous image of an interval. Arcwise connected is not the same as connected. A well known example is the following.

$$\left\{\left(x, \sin\frac{1}{x}\right) : x \in (0, 1]\right\} \cup \{(0, y) : y \in [-1, 1]\} \tag{5.2}$$

You can verify that this set of points in the normed vector space $\mathbb{R}^2$ is not arcwise connected but is connected.

Lemma 5.3.12 *In a normed vector space, $B(\mathbf{z}, r)$ is arcwise connected.*

Proof: This is easy from the convexity of the set. If $\mathbf{x},\mathbf{y} \in B(\mathbf{z},r)$, then let $\boldsymbol{\gamma}(t) = \mathbf{x} + t(\mathbf{y}-\mathbf{x})$ for $t \in [0,1]$.

$$\begin{aligned} ||\mathbf{x} + t(\mathbf{y}-\mathbf{x}) - \mathbf{z}|| &= ||(1-t)(\mathbf{x}-\mathbf{z}) + t(\mathbf{y}-\mathbf{z})|| \\ &\leq (1-t)||\mathbf{x}-\mathbf{z}|| + t||\mathbf{y}-\mathbf{z}|| \\ &< (1-t)r + tr = r \end{aligned}$$

showing $\boldsymbol{\gamma}(t)$ stays in $B(\mathbf{z},r)$.■

Proposition 5.3.13 *If X is arcwise connected, then it is connected.*

Proof: Let X be an arcwise connected set and suppose it is separated. Then $X = A \cup B$ where A, B are two separated sets. Pick $\mathbf{p} \in A$ and $\mathbf{q} \in B$. Since X is given to be arcwise connected, there must exist a continuous function $\gamma : [a,b] \to X$ such that $\gamma(a) = \mathbf{p}$ and $\gamma(b) = \mathbf{q}$. But then $\gamma([a,b]) = (\gamma([a,b]) \cap A) \cup (\gamma([a,b]) \cap B)$ and the two sets $\gamma([a,b]) \cap A$ and $\gamma([a,b]) \cap B$ are separated thus showing that $\gamma([a,b])$ is separated and contradicting Theorem 5.3.8 and Theorem 5.3.5. It follows that X must be connected as claimed.■

Theorem 5.3.14 *Let U be an open subset of a normed vector space. Then U is arcwise connected if and only if U is connected. Also the connected components of an open set are open sets.*

Proof: By Proposition 5.3.13 it is only necessary to verify that if U is connected and open in the context of this theorem, then U is arcwise connected. Pick $\mathbf{p} \in U$. Say $\mathbf{x} \in U$ satisfies $\mathcal{P}$ if there exists a continuous function, $\boldsymbol{\gamma} : [a,b] \to U$ such that $\gamma(a) = \mathbf{p}$ and $\boldsymbol{\gamma}(\mathbf{b}) = \mathbf{x}$.

$$A \equiv \{\mathbf{x} \in U \text{ such that } \mathbf{x} \text{ satisfies } \mathcal{P}.\}$$

If $\mathbf{x} \in A$, then Lemma 5.3.12 implies $B(\mathbf{x},r) \subseteq U$ is arcwise connected for small enough r. Thus letting $\mathbf{y} \in B(\mathbf{x},r)$, there exist intervals, $[a,b]$ and $[c,d]$ and continuous functions having values in U, $\boldsymbol{\gamma}, \boldsymbol{\eta}$ such that $\boldsymbol{\gamma}(a) = \mathbf{p}, \boldsymbol{\gamma}(b) = \mathbf{x}, \boldsymbol{\eta}(c) = \mathbf{x}$, and $\eta(d) = \mathbf{y}$. Then let $\boldsymbol{\gamma}_1 : [a, b+d-c] \to U$ be defined as

$$\boldsymbol{\gamma}_1(t) \equiv \begin{cases} \boldsymbol{\gamma}(t) & \text{if } t \in [a,b] \\ \boldsymbol{\eta}(t+c-b) & \text{if } t \in [b, b+d-c] \end{cases}$$

Then it is clear that $\boldsymbol{\gamma}_1$ is a continuous function mapping $\mathbf{p}$ to $\mathbf{y}$ and showing that $B(\mathbf{x},r) \subseteq A$. Therefore, A is open. $A \neq \emptyset$ because since U is open there is an open set, $B(\mathbf{p},\delta)$ containing $\mathbf{p}$ which is contained in U and is arcwise connected.

Now consider $B \equiv U \setminus A$. I claim this is also open. If B is not open, there exists a point $\mathbf{z} \in B$ such that every open set containing $\mathbf{z}$ is not contained in B. Therefore, letting $B(\mathbf{z},\delta)$ be such that $\mathbf{z} \in B(\mathbf{z},\delta) \subseteq U$, there exist points of A contained in $B(\mathbf{z},\delta)$. But then, a repeat of the above argument shows $z \in A$ also. Hence B is open and so if $B \neq \emptyset$, then $U = B \cup A$ and so U is separated by the two sets B and A contradicting the assumption that U is connected.

It remains to verify the connected components are open. Let $\mathbf{z} \in C_{\mathbf{p}}$ where $C_{\mathbf{p}}$ is the connected component determined by $\mathbf{p}$. Then picking $B(\mathbf{z},\delta) \subseteq U$, $C_{\mathbf{p}} \cup B(\mathbf{z},\delta)$ is connected and contained in U and so it must also be contained in $C_{\mathbf{p}}$. Thus $\mathbf{z}$ is an interior point of $C_{\mathbf{p}}$. This proves the theorem. ■

As an application, consider the following corollary.

Corollary 5.3.15 *Let $f : \Omega \to \mathbb{Z}$ be continuous where Ω is a connected open set in a normed vector space. Then f must be a constant.*

Proof: Suppose not. Then it achieves two different values, k and $l \neq k$. Then $\Omega = f^{-1}(l) \cup f^{-1}(\{m \in \mathbb{Z} : m \neq l\})$ and these are disjoint nonempty open sets which separate Ω. To see they are open, note

$$f^{-1}(\{m \in \mathbb{Z} : m \neq l\}) = f^{-1}\left(\cup_{m \neq l}\left(m - \frac{1}{6}, m + \frac{1}{6}\right)\right)$$

which is the inverse image of an open set while $f^{-1}(l) = f^{-1}\left(\left(l - \frac{1}{6}, l + \frac{1}{6}\right)\right)$ also an open set. ■

5.4 Uniform Continuity

The concept of uniform continuity is also similar to the one dimensional concept.

Definition 5.4.1 *Let* $\mathbf{f}$ *be a function. Then* $\mathbf{f}$ *is uniformly continuous if for every* $\varepsilon > 0$, *there exists a* δ ***depending only on*** ε *such that if* $||\mathbf{x} - \mathbf{y}|| < \delta$ *then* $||\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})|| < \varepsilon$.

Theorem 5.4.2 *Let* $\mathbf{f} : K \to \mathbb{F}$ *be continuous where* K *is a sequentially compact set in* $\mathbb{F}^n$ *or more generally a normed vector space. Then* $\mathbf{f}$ *is uniformly continuous on* K.

Proof: If this is not true, there exists $\varepsilon > 0$ such that for every $\delta > 0$ there exists a pair of points, $\mathbf{x}_\delta$ and $\mathbf{y}_\delta$ such that even though $||\mathbf{x}_\delta - \mathbf{y}_\delta|| < \delta$, $||\mathbf{f}(\mathbf{x}_\delta) - \mathbf{f}(\mathbf{y}_\delta)|| \geq \varepsilon$. Taking a succession of values for δ equal to $1, 1/2, 1/3, \cdots$, and letting the exceptional pair of points for $\delta = 1/n$ be denoted by $\mathbf{x}_n$ and $\mathbf{y}_n$,

$$||\mathbf{x}_n - \mathbf{y}_n|| < \frac{1}{n}, ||\mathbf{f}(\mathbf{x}_n) - \mathbf{f}(\mathbf{y}_n)|| \geq \varepsilon.$$

Now since K is sequentially compact, there exists a subsequence, $\{\mathbf{x}_{n_k}\}$ such that $\mathbf{x}_{n_k} \to \mathbf{z} \in K$. Now $n_k \geq k$ and so

$$||\mathbf{x}_{n_k} - \mathbf{y}_{n_k}|| < \frac{1}{k}.$$

Hence

$$\begin{aligned} ||\mathbf{y}_{n_k} - \mathbf{z}|| &\leq ||\mathbf{y}_{n_k} - \mathbf{x}_{n_k}|| + ||\mathbf{x}_{n_k} - \mathbf{z}|| \\ &< \frac{1}{k} + ||\mathbf{x}_{n_k} - \mathbf{z}|| \end{aligned}$$

Consequently, $\mathbf{y}_{n_k} \to \mathbf{z}$ also. By continuity of $\mathbf{f}$ and Theorem 5.1.2,

$$0 = ||\mathbf{f}(\mathbf{z}) - \mathbf{f}(\mathbf{z})|| = \lim_{k \to \infty} ||\mathbf{f}(\mathbf{x}_{n_k}) - \mathbf{f}(\mathbf{y}_{n_k})|| \geq \varepsilon,$$

an obvious contradiction. Therefore, the theorem must be true. ■

Recall the closed and bounded subsets of $\mathbb{F}^n$ are those which are sequentially compact.

5.5 Sequences And Series Of Functions

Now it is an easy matter to consider sequences of vector valued functions.

Definition 5.5.1 *A sequence of functions is a map defined on* $\mathbb{N}$ *or some set of integers larger than or equal to a given integer,* m *which has values which are functions. It is written in the form* $\{\mathbf{f}_n\}_{n=m}^{\infty}$ *where* $\mathbf{f}_n$ *is a function. It is assumed also that the domain of all these functions is the same.*

Here the functions have values in some normed vector space.

The definition of uniform convergence is exactly the same as earlier only now it is not possible to draw representative pictures so easily.

Definition 5.5.2 *Let $\{\mathbf{f}_n\}$ be a sequence of functions. Then the sequence converges pointwise to a function $\mathbf{f}$ if for all $\mathbf{x} \in D$, the domain of the functions in the sequence,*

$$\mathbf{f}(\mathbf{x}) = \lim_{n\to\infty} \mathbf{f}_n(\mathbf{x})$$

Thus you consider for each $\mathbf{x} \in D$ the sequence of numbers $\{\mathbf{f}_n(\mathbf{x})\}$ and if this sequence converges for each $\mathbf{x} \in D$, the thing it converges to is called $\mathbf{f}(\mathbf{x})$.

Definition 5.5.3 *Let $\{\mathbf{f}_n\}$ be a sequence of functions defined on D. Then $\{\mathbf{f}_n\}$ is said to converge uniformly to $\mathbf{f}$ if it converges pointwise to $\mathbf{f}$ and for every $\varepsilon > 0$ there exists N such that for all $n \geq N$*

$$||\mathbf{f}(\mathbf{x}) - \mathbf{f}_n(\mathbf{x})|| < \varepsilon$$

for all $\mathbf{x} \in D$.

Theorem 5.5.4 *Let $\{\mathbf{f}_n\}$ be a sequence of continuous functions defined on D and suppose this sequence converges uniformly to $\mathbf{f}$. Then $\mathbf{f}$ is also continuous on D. If each $\mathbf{f}_n$ is uniformly continuous on D, then $\mathbf{f}$ is also uniformly continuous on D.*

Proof: Let $\varepsilon > 0$ be given and pick $\mathbf{z} \in D$. By uniform convergence, there exists N such that if $n > N$, then for all $\mathbf{x} \in D$,

$$||\mathbf{f}(\mathbf{x}) - \mathbf{f}_n(\mathbf{x})|| < \varepsilon/3. \tag{5.3}$$

Pick such an n. By assumption, $\mathbf{f}_n$ is continuous at $\mathbf{z}$. Therefore, there exists $\delta > 0$ such that if $||\mathbf{z} - \mathbf{x}|| < \delta$ then

$$||\mathbf{f}_n(\mathbf{x}) - \mathbf{f}_n(\mathbf{z})|| < \varepsilon/3.$$

It follows that for $||\mathbf{x} - \mathbf{z}|| < \delta$,

$$\begin{aligned} ||\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{z})|| &\leq ||\mathbf{f}(\mathbf{x}) - \mathbf{f}_n(\mathbf{x})|| + ||\mathbf{f}_n(\mathbf{x}) - \mathbf{f}_n(\mathbf{z})|| + ||\mathbf{f}_n(\mathbf{z}) - \mathbf{f}(\mathbf{z})|| \\ &< \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon \end{aligned}$$

which shows that since ε was arbitrary, $\mathbf{f}$ is continuous at $\mathbf{z}$.

In the case where each $\mathbf{f}_n$ is uniformly continuous, and using the same $\mathbf{f}_n$ for which 5.3 holds, there exists a $\delta > 0$ such that if $||\mathbf{y} - \mathbf{z}|| < \delta$, then

$$||\mathbf{f}_n(\mathbf{z}) - \mathbf{f}_n(\mathbf{y})|| < \varepsilon/3.$$

Then for $||\mathbf{y} - \mathbf{z}|| < \delta$,

$$\begin{aligned} ||\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{z})|| &\leq ||\mathbf{f}(\mathbf{y}) - \mathbf{f}_n(\mathbf{y})|| + ||\mathbf{f}_n(\mathbf{y}) - \mathbf{f}_n(\mathbf{z})|| + ||\mathbf{f}_n(\mathbf{z}) - \mathbf{f}(\mathbf{z})|| \\ &< \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon \end{aligned}$$

This shows uniform continuity of $\mathbf{f}$. This proves the theorem. ■

Definition 5.5.5 *Let $\{\mathbf{f}_n\}$ be a sequence of functions defined on D. Then the sequence is said to be uniformly Cauchy if for every $\varepsilon > 0$ there exists N such that whenever $m, n \geq N$,*

$$||\mathbf{f}_m(\mathbf{x}) - \mathbf{f}_n(\mathbf{x})|| < \varepsilon$$

for all $\mathbf{x} \in D$.

Then the following theorem follows easily.

Theorem 5.5.6 *Let $\{\mathbf{f}_n\}$ be a uniformly Cauchy sequence of functions defined on D having values in a complete normed vector space such as $\mathbb{F}^n$ for example. Then there exists $\mathbf{f}$ defined on D such that $\{\mathbf{f}_n\}$ converges uniformly to $\mathbf{f}$.*

Proof: For each $\mathbf{x} \in D, \{\mathbf{f}_n(\mathbf{x})\}$ is a Cauchy sequence. Therefore, it converges to some vector $\mathbf{f}(\mathbf{x})$. Let $\varepsilon > 0$ be given and let N be such that if $n, m \geq N$,

$$||\mathbf{f}_m(\mathbf{x}) - \mathbf{f}_n(\mathbf{x})|| < \varepsilon/2$$

for all $\mathbf{x} \in D$. Then for any $\mathbf{x} \in D$, pick $n \geq N$ and it follows from Theorem 4.1.8

$$||\mathbf{f}(\mathbf{x}) - \mathbf{f}_n(\mathbf{x})|| = \lim_{m\to\infty} ||\mathbf{f}_m(\mathbf{x}) - \mathbf{f}_n(\mathbf{x})|| \leq \varepsilon/2 < \varepsilon.$$

This proves the theorem. ■

Corollary 5.5.7 *Let $\{\mathbf{f}_n\}$ be a uniformly Cauchy sequence of functions continuous on D having values in a complete normed vector space like $\mathbb{F}^n$. Then there exists $\mathbf{f}$ defined on D such that $\{\mathbf{f}_n\}$ converges uniformly to $\mathbf{f}$ and $\mathbf{f}$ is continuous. Also, if each $\mathbf{f}_n$ is uniformly continuous, then so is $\mathbf{f}$.*

Proof: This follows from Theorem 5.5.6 and Theorem 5.5.4. This proves the corollary. ■

Here is one more fairly obvious theorem.

Theorem 5.5.8 *Let $\{\mathbf{f}_n\}$ be a sequence of functions defined on D having values in a complete normed vector space like $\mathbb{F}^n$. Then it converges pointwise if and only if the sequence $\{\mathbf{f}_n(\mathbf{x})\}$ is a Cauchy sequence for every $\mathbf{x} \in D$. It converges uniformly if and only if $\{\mathbf{f}_n\}$ is a uniformly Cauchy sequence.*

Proof: If the sequence converges pointwise, then by Theorem 4.4.3 the sequence $\{\mathbf{f}_n(\mathbf{x})\}$ is a Cauchy sequence for each $\mathbf{x} \in D$. Conversely, if $\{\mathbf{f}_n(\mathbf{x})\}$ is a Cauchy sequence for each $\mathbf{x} \in D$, then $\{\mathbf{f}_n(\mathbf{x})\}$ converges for each $\mathbf{x} \in D$ because of completeness.

Now suppose $\{\mathbf{f}_n\}$ is uniformly Cauchy. Then from Theorem 5.5.6 there exists $\mathbf{f}$ such that $\{\mathbf{f}_n\}$ converges uniformly on D to $\mathbf{f}$. Conversely, if $\{\mathbf{f}_n\}$ converges uniformly to $\mathbf{f}$ on D, then if $\varepsilon > 0$ is given, there exists N such that if $n \geq N$,

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}_n(\mathbf{x})| < \varepsilon/2$$

for every $\mathbf{x} \in D$. Then if $m, n \geq N$ and $\mathbf{x} \in D$,

$$|\mathbf{f}_n(\mathbf{x}) - \mathbf{f}_m(\mathbf{x})| \leq |\mathbf{f}_n(\mathbf{x}) - \mathbf{f}(\mathbf{x})| + |\mathbf{f}(\mathbf{x}) - \mathbf{f}_m(\mathbf{x})| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Thus $\{\mathbf{f}_n\}$ is uniformly Cauchy. ■

Once you understand sequences, it is no problem to consider series.

Definition 5.5.9 *Let $\{\mathbf{f}_n\}$ be a sequence of functions defined on D. Then*

$$\left(\sum_{k=1}^{\infty} \mathbf{f}_k\right)(\mathbf{x}) \equiv \lim_{n\to\infty} \sum_{k=1}^{n} \mathbf{f}_k(\mathbf{x}) \tag{5.4}$$

whenever the limit exists. Thus there is a new function denoted by

$$\sum_{k=1}^{\infty} \mathbf{f}_k \tag{5.5}$$

and its value at $\mathbf{x}$ *is given by the limit of the sequence of partial sums in 5.4. If for all* $\mathbf{x} \in D$*, the limit in 5.4 exists, then 5.5 is said to converge pointwise.* $\sum_{k=1}^{\infty} \mathbf{f}_k$ *is said to converge uniformly on* D *if the sequence of partial sums,*

$$\left\{\sum_{k=1}^{n} \mathbf{f}_k\right\}$$

converges uniformly.If the indices for the functions start at some other value than 1, you make the obvious modification to the above definition.

Theorem 5.5.10 *Let* $\{\mathbf{f}_n\}$ *be a sequence of functions defined on* D *which have values in a complete normed vector space like* $\mathbb{F}^n$*. The series* $\sum_{k=1}^{\infty} \mathbf{f}_k$ *converges pointwise if and only if for each* $\varepsilon > 0$ *and* $\mathbf{x} \in D$*, there exists* $N_{\varepsilon,\mathbf{x}}$ *which may depend on* $\mathbf{x}$ *as well as* ε *such that when* $q > p \geq N_{\varepsilon,\mathbf{x}}$*,*

$$\left|\left|\sum_{k=p}^{q} \mathbf{f}_k(\mathbf{x})\right|\right| < \varepsilon$$

The series $\sum_{k=1}^{\infty} \mathbf{f}_k$ *converges uniformly on* D *if for every* $\varepsilon > 0$ *there exists* N_ε *such that if* $q > p \geq N$ *then*

$$\left|\left|\sum_{k=p}^{q} \mathbf{f}_k(\mathbf{x})\right|\right| < \varepsilon \tag{5.6}$$

for all $\mathbf{x} \in D$*.*

Proof: The first part follows from Theorem 5.5.8. The second part follows from observing the condition is equivalent to the sequence of partial sums forming a uniformly Cauchy sequence and then by Theorem 5.5.6, these partial sums converge uniformly to a function which is the definition of $\sum_{k=1}^{\infty} \mathbf{f}_k$. This proves the theorem. ■

Is there an easy way to recognize when 5.6 happens? Yes, there is. It is called the Weierstrass M test.

Theorem 5.5.11 *Let* $\{\mathbf{f}_n\}$ *be a sequence of functions defined on* D *having values in a complete normed vector space like* $\mathbb{F}^n$*. Suppose there exists* M_n *such that* $\sup\{|\mathbf{f}_n(\mathbf{x})| : \mathbf{x} \in D\} < M_n$ *and* $\sum_{n=1}^{\infty} M_n$ *converges. Then* $\sum_{n=1}^{\infty} \mathbf{f}_n$ *converges uniformly on* D*.*

Proof: Let $\mathbf{z} \in D$. Then letting $m < n$ and using the triangle inequality

$$\left|\left|\sum_{k=1}^{n} \mathbf{f}_k(\mathbf{z}) - \sum_{k=1}^{m} \mathbf{f}_k(\mathbf{z})\right|\right| \leq \sum_{k=m+1}^{n} ||\mathbf{f}_k(\mathbf{z})|| \leq \sum_{k=m+1}^{\infty} M_k < \varepsilon$$

whenever m is large enough because of the assumption that $\sum_{n=1}^{\infty} M_n$ converges. Therefore, the sequence of partial sums is uniformly Cauchy on D and therefore, converges uniformly to $\sum_{k=1}^{\infty} \mathbf{f}_k$ on D. This proves the theorem. ■

Theorem 5.5.12 *If* $\{\mathbf{f}_n\}$ *is a sequence of continuous functions defined on* D *and* $\sum_{k=1}^{\infty} \mathbf{f}_k$ *converges uniformly, then the function,* $\sum_{k=1}^{\infty} \mathbf{f}_k$ *must also be continuous.*

Proof: This follows from Theorem 5.5.4 applied to the sequence of partial sums of the above series which is assumed to converge uniformly to the function, $\sum_{k=1}^{\infty} \mathbf{f}_k$. ■

5.6 Polynomials

General considerations about what a function is have already been considered earlier. For functions of one variable, the special kind of functions known as a polynomial has a corresponding version when one considers a function of many variables. This is found in the next definition.

Definition 5.6.1 *Let α be an n dimensional multi-index. This means*

$$\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$$

where each α_i is a positive integer or zero. Also, let

$$|\boldsymbol{\alpha}| \equiv \sum_{i=1}^{n} |\alpha_i|$$

Then $\mathbf{x}^{\alpha}$ means

$$\mathbf{x}^{\alpha} \equiv x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_3^{\alpha_n}$$

where each $x_j \in \mathbb{F}$. An n dimensional polynomial of degree m is a function of the form

$$p(\mathbf{x}) = \sum_{|\alpha| \leq m} d_{\alpha} \mathbf{x}^{\alpha}.$$

where the d_{α} are complex or real numbers. Rational functions are defined as the quotient of two polynomials. Thus these functions are defined on $\mathbb{F}^n$.

For example, $f(\mathbf{x}) = x_1 x_2^2 + 7x_3^4 x_1$ is a polynomial of degree 5 and

$$\frac{x_1 x_2^2 + 7x_3^4 x_1 + x_2^3}{4x_1^3 x_2^2 + 7x_3^2 x_1 - x_2^3}$$

is a rational function.

Note that in the case of a rational function, the domain of the function might not be all of $\mathbb{F}^n$. For example, if

$$f(\mathbf{x}) = \frac{x_1 x_2^2 + 7x_3^4 x_1 + x_2^3}{x_2^2 + 3x_1^2 - 4},$$

the domain of f would be all complex numbers such that $x_2^2 + 3x_1^2 \neq 4$.

By Theorem 5.0.2 all polynomials are continuous. To see this, note that the function,

$$\pi_k(\mathbf{x}) \equiv x_k$$

is a continuous function because of the inequality

$$|\pi_k(\mathbf{x}) - \pi_k(\mathbf{y})| = |x_k - y_k| \leq |\mathbf{x} - \mathbf{y}|.$$

Polynomials are simple sums of scalars times products of these functions. Similarly, by this theorem, rational functions, quotients of polynomials, are continuous at points where the denominator is non zero. More generally, if V is a normed vector space, consider a V valued function of the form

$$\mathbf{f}(\mathbf{x}) \equiv \sum_{|\alpha| \leq m} \mathbf{d}_{\alpha} \mathbf{x}^{\alpha}$$

where $\mathbf{d}_{\alpha} \in V$, sort of a V valued polynomial. Then such a function is continuous by application of Theorem 5.0.2 and the above observation about the continuity of the functions π_k.

Thus there are lots of examples of continuous functions. However, it is even better than the above discussion indicates. As in the case of a function of one variable, an arbitrary continuous function can typically be approximated uniformly by a polynomial. This is the n dimensional version of the Weierstrass approximation theorem.

5.7 Sequences Of Polynomials, Weierstrass Approximation

An arbitrary continuous function defined on an interval can be approximated uniformly by a polynomial, there exists a similar theorem which is just a generalization of this which will hold for continuous functions defined on a box or more generally a closed and bounded set. However, we will settle for the case of a box. The proof is based on the following lemma.

Lemma 5.7.1 *The following estimate holds for $x \in [0,1]$ and $m \geq 2$.*

$$\sum_{k=0}^{m} \binom{m}{k} (k - mx)^2 x^k (1-x)^{m-k} \leq \frac{1}{4} m$$

Proof: First of all, from the binomial theorem

$$\sum_{k=0}^{m} \binom{m}{k} (tx)^k (1-x)^{m-k} = (1 - x + tx)^m$$

Take a derivative and then let $t = 1$.

$$\sum_{k=0}^{m} \binom{m}{k} k (tx)^{k-1} x (1-x)^{m-k} = mx (tx - x + 1)^{m-1}$$

$$\sum_{k=0}^{m} \binom{m}{k} k (x)^{k} (1-x)^{m-k} = mx$$

Then also,

$$\sum_{k=0}^{m} \binom{m}{k} k (tx)^{k} (1-x)^{m-k} = mxt (tx - x + 1)^{m-1}$$

Take another time derivative of both sides.

$$\begin{aligned} & \sum_{k=0}^{m} \binom{m}{k} k^2 (tx)^{k-1} x (1-x)^{m-k} \\ = \; & mx \left((tx - x + 1)^{m-1} - tx (tx - x + 1)^{m-2} + mtx (tx - x + 1)^{m-2} \right) \end{aligned}$$

Plug in $t = 1$.

$$\sum_{k=0}^{m} \binom{m}{k} k^2 x^k (1-x)^{m-k} = mx (mx - x + 1)$$

Then it follows

$$\begin{aligned} & \sum_{k=0}^{m} \binom{m}{k} (k - mx)^2 x^k (1-x)^{m-k} \\ = & \sum_{k=0}^{m} \binom{m}{k} \left(k^2 - 2kmx + x^2 m^2 \right) x^k (1-x)^{m-k} \end{aligned}$$

and from what was just shown along with the binomial theorem again, this equals

$$x^2 m^2 - x^2 m + mx - 2mx (mx) + x^2 m^2 = -x^2 m + mx = \frac{m}{4} - m \left(x - \frac{1}{2} \right)^2.$$

Thus the expression is maximized when $x = 1/2$ and yields $m/4$ in this case. This proves the lemma. ■

Now let f be a continuous function defined on $[0,1]$. Let p_n be the polynomial defined by

$$p_n(x) \equiv \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1-x)^{n-k}. \tag{5.7}$$

For f a continuous function defined on $[0,1]^n$ and for $\mathbf{x} = (x_1, \cdots, x_n)$,consider the polynomial,

$$p_{\mathbf{m}}(\mathbf{x}) \equiv \sum_{k_1=0}^{m_1} \cdots \sum_{k_n=0}^{m_n} \binom{m_1}{k_1}\binom{m_2}{k_2}\cdots\binom{m_n}{k_n} x_1^{k_1}(1-x_1)^{m_1-k_1} x_2^{k_2}(1-x_2)^{m_2-k_2}$$

$$\cdots x_n^{k_n}(1-x_n)^{m_n-k_n} f\left(\frac{k_1}{m_1}, \cdots, \frac{k_n}{m_n}\right). \tag{5.8}$$

Also define if I is a set in $\mathbb{R}^n$

$$||h||_I \equiv \sup\{|h(\mathbf{x})| : \mathbf{x} \in I\}.$$

Let

$$\min(\mathbf{m}) \equiv \min\{m_1, \cdots, m_n\}, \ \max(\mathbf{m}) \equiv \max\{m_1, \cdots, m_n\}$$

Definition 5.7.2 *Define $p_{\mathbf{m}}$ converges uniformly to f on a set, I if*

$$\lim_{\min(\mathbf{m})\to\infty} ||p_{\mathbf{m}} - f||_I = 0.$$

To simplify the notation, let $\mathbf{k} = (k_1, \cdots, k_n)$ where each $k_i \in [0, m_i]$,

$$\frac{\mathbf{k}}{\mathbf{m}} \equiv \left(\frac{k_1}{m_1}, \cdots, \frac{k_n}{m_n}\right),$$

and let

$$\binom{\mathbf{m}}{\mathbf{k}} \equiv \binom{m_1}{k_1}\binom{m_2}{k_2}\cdots\binom{m_n}{k_n}.$$

Also define for $\mathbf{k} = (k_1, \cdots, k_n)$

$$\mathbf{k} \le \mathbf{m} \text{ if } 0 \le k_i \le m_i \text{ for each } i$$

$$\mathbf{x}^{\mathbf{k}}(\mathbf{1}-\mathbf{x})^{\mathbf{m}-\mathbf{k}} \equiv x_1^{k_1}(1-x_1)^{m_1-k_1} x_2^{k_2}(1-x_2)^{m_2-k_2}\cdots x_n^{k_n}(1-x_n)^{m_n-k_n}.$$

Thus in terms of this notation,

$$p_{\mathbf{m}}(\mathbf{x}) = \sum_{\mathbf{k}\le\mathbf{m}} \binom{\mathbf{m}}{\mathbf{k}} \mathbf{x}^{\mathbf{k}}(\mathbf{1}-\mathbf{x})^{\mathbf{m}-\mathbf{k}} f\left(\frac{\mathbf{k}}{\mathbf{m}}\right)$$

This is the n dimensional version of the Bernstein polynomials which is what results in the case where $n = 1$.

Lemma 5.7.3 *For $\mathbf{x} \in [0,1]^n$, f a continuous $\mathbb{F}$ valued function defined on $[0,1]^n$, and $p_{\mathbf{m}}$ given in 5.8, $p_{\mathbf{m}}$ converges uniformly to f on $[0,1]^n$ as $m \to \infty$. More generally, one can have f a continuous function with values in an arbitrary real or complex normed linear space. There is no change in the conclusions and proof. You just write $\|\cdot\|$ instead of $|\cdot|$.*

Proof: The function f is uniformly continuous because it is continuous on a sequentially compact set $[0,1]^n$. Therefore, there exists $\delta > 0$ such that if $|\mathbf{x}-\mathbf{y}| < \delta$, then

$$|f(\mathbf{x}) - f(\mathbf{y})| < \varepsilon.$$

Denote by G the set of $\mathbf{k}$ such that $(k_i - m_i x_i)^2 < \eta^2 m^2$ for each i where $\eta = \delta/\sqrt{n}$. Note this condition is equivalent to saying that for each i, $\left|\frac{k_i}{m_i} - x_i\right| < \eta$ and

$$\left|\frac{\mathbf{k}}{\mathbf{m}} - \mathbf{x}\right| < \delta$$

A short computation shows that by the binomial theorem,

$$\sum_{\mathbf{k}\le\mathbf{m}} \binom{\mathbf{m}}{\mathbf{k}} \mathbf{x}^{\mathbf{k}} (\mathbf{1}-\mathbf{x})^{\mathbf{m}-\mathbf{k}} = 1$$

and so for $\mathbf{x} \in [0,1]^n$,

$$\begin{aligned} |p_{\mathbf{m}}(\mathbf{x}) - f(\mathbf{x})| &\le \sum_{\mathbf{k}\le\mathbf{m}} \binom{\mathbf{m}}{\mathbf{k}} \mathbf{x}^{\mathbf{k}} (\mathbf{1}-\mathbf{x})^{\mathbf{m}-\mathbf{k}} \left| f\left(\frac{\mathbf{k}}{\mathbf{m}}\right) - f(\mathbf{x})\right| \\ &\le \sum_{\mathbf{k}\in G} \binom{\mathbf{m}}{\mathbf{k}} \mathbf{x}^{\mathbf{k}} (\mathbf{1}-\mathbf{x})^{\mathbf{m}-\mathbf{k}} \left| f\left(\frac{\mathbf{k}}{\mathbf{m}}\right) - f(\mathbf{x})\right| \\ &+ \sum_{\mathbf{k}\in G^C} \binom{\mathbf{m}}{\mathbf{k}} \mathbf{x}^{\mathbf{k}} (\mathbf{1}-\mathbf{x})^{\mathbf{m}-\mathbf{k}} \left| f\left(\frac{\mathbf{k}}{\mathbf{m}}\right) - f(\mathbf{x})\right| \end{aligned} \tag{5.9}$$

Now for $\mathbf{k} \in G$ it follows that for each i

$$\left|\frac{k_i}{m_i} - x_i\right| < \frac{\delta}{\sqrt{n}} \tag{5.10}$$

and so $\left|f\left(\frac{\mathbf{k}}{\mathbf{m}}\right) - f(\mathbf{x})\right| < \varepsilon$ because the above implies $\left|\frac{\mathbf{k}}{\mathbf{m}} - \mathbf{x}\right| < \delta$. Therefore, the first sum on the right in 5.9 is no larger than

$$\sum_{\mathbf{k}\in G} \binom{\mathbf{m}}{\mathbf{k}} \mathbf{x}^{\mathbf{k}} (\mathbf{1}-\mathbf{x})^{\mathbf{m}-\mathbf{k}} \varepsilon \le \sum_{\mathbf{k}\le\mathbf{m}} \binom{\mathbf{m}}{\mathbf{k}} \mathbf{x}^{\mathbf{k}} (\mathbf{1}-\mathbf{x})^{\mathbf{m}-\mathbf{k}} \varepsilon = \varepsilon.$$

Letting $M \ge \max\{|f(\mathbf{x})| : \mathbf{x} \in [0,1]^n\}$ it follows that for some j,

$$\left|\frac{k_j}{m_j} - x_j\right| \ge \frac{\delta}{\sqrt{n}},\ (k_j - m_j x_j)^2 \ge m_j^2 \frac{\delta^2}{n}$$

by Lemma 5.7.1,

$$\begin{aligned} & |p_{\mathbf{m}}(\mathbf{x}) - f(\mathbf{x})| \\ \le\ & \varepsilon + 2M \sum_{\mathbf{k}\in G^C} \binom{\mathbf{m}}{\mathbf{k}} \mathbf{x}^{\mathbf{k}} (\mathbf{1}-\mathbf{x})^{\mathbf{m}-\mathbf{k}} \\ \le\ & \varepsilon + 2Mn \sum_{\mathbf{k}\in G^C} \binom{\mathbf{m}}{\mathbf{k}} \frac{(k_j - m_j x_j)^2}{\delta^2 m_j^2} \mathbf{x}^{\mathbf{k}} (\mathbf{1}-\mathbf{x})^{\mathbf{m}-\mathbf{k}} \\ \le\ & \varepsilon + 2Mn \frac{1}{\delta^2 m_j^2} \frac{1}{4} m_j = \varepsilon + \frac{1}{2} M \frac{n}{\delta^2 m_j} \le \varepsilon + \frac{1}{2} M \frac{n}{\delta^2 \min(\mathbf{m})} \end{aligned} \tag{5.11}$$

Therefore, since the right side does not depend on $\mathbf{x}$, it follows that if $\min(\mathbf{m})$ is large enough,

$$||p_{\mathbf{m}} - f||_{[0,1]^n} \leq 2\varepsilon$$

and since ε is arbitrary, this shows $\lim_{\min(\mathbf{m})\to\infty} ||p_{\mathbf{m}} - f||_{[0,1]^n} = 0$. This proves the lemma. ■

These Bernstein polynomials are very remarkable approximations. It turns out that if f is $C^1([0,1]^n)$, then

$$\lim_{\min(\mathbf{m})\to\infty} p_{\mathbf{m}\mathbf{x}_i}(\mathbf{x}) \to f_{x_i}(\mathbf{x}) \text{ uniformly on } [0,1]^n.$$

We show this first for the case that $n = 1$. From this, it is obvious for the general case.

Lemma 5.7.4 *Let $f \in C^1([0,1])$ and let*

$$p_m(x) \equiv \sum_{k=0}^{m} \binom{m}{k} x^k (1-x)^{m-k} f\left(\frac{k}{m}\right)$$

be the m^{th} Bernstein polynomial. Then in addition to $\|p_m - f\|_{[0,1]} \to 0$, it also follows that

$$\|p_m' - f'\|_{[0,1]} \to 0$$

Proof: From simple computations,

$$p_m'(x) = \sum_{k=1}^{m} \binom{m}{k} k x^{k-1} (1-x)^{m-k} f\left(\frac{k}{m}\right) - \sum_{k=0}^{m-1} \binom{m}{k} x^k (m-k)(1-x)^{m-1-k} f\left(\frac{k}{m}\right)$$

$$= \sum_{k=1}^{m} \frac{m(m-1)!}{(m-k)!(k-1)!} x^{k-1}(1-x)^{m-k} f\left(\frac{k}{m}\right) - \sum_{k=0}^{m-1} \binom{m}{k} x^k (m-k)(1-x)^{m-1-k} f\left(\frac{k}{m}\right)$$

$$= \sum_{k=0}^{m-1} \frac{m(m-1)!}{(m-1-k)!k!} x^k (1-x)^{m-1-k} f\left(\frac{k+1}{m}\right) - \sum_{k=0}^{m-1} \frac{m(m-1)!}{(m-1-k)!k!} x^k (1-x)^{m-1-k} f\left(\frac{k}{m}\right)$$

$$= \sum_{k=0}^{m-1} \frac{m(m-1)!}{(m-1-k)!k!} x^k (1-x)^{m-1-k} \left(f\left(\frac{k+1}{m}\right) - f\left(\frac{k}{m}\right)\right)$$

$$= \sum_{k=0}^{m-1} \binom{m-1}{k} x^k (1-x)^{m-1-k} \left(\frac{f\left(\frac{k+1}{m}\right) - f\left(\frac{k}{m}\right)}{1/m}\right)$$

By the mean value theorem,

$$\frac{f\left(\frac{k+1}{m}\right) - f\left(\frac{k}{m}\right)}{1/m} = f'(x_{k,m}),\ x_{k,m} \in \left(\frac{k}{m}, \frac{k+1}{m}\right)$$

Now the desired result follows as before from the uniform continuity of f' on $[0,1]$. Let $\delta > 0$ be such that if

$$|x - y| < \delta, \text{ then } |f'(x) - f'(y)| < \varepsilon$$

and let m be so large that $1/m < \delta/2$. Then if $\left|x - \frac{k}{m}\right| < \delta/2$, it follows that $|x - x_{k,m}| < \delta$ and so

$$|f'(x) - f'(x_{k,m})| = \left|f'(x) - \frac{f\left(\frac{k+1}{m}\right) - f\left(\frac{k}{m}\right)}{1/m}\right| < \varepsilon.$$

Now as before, letting $M \geq |f'(x)|$ for all x,

$$|p_m'(x) - f'(x)| \leq \sum_{k=0}^{m-1} \binom{m-1}{k} x^k (1-x)^{m-1-k} |f'(x_{k,m}) - f'(x)|$$

$$\leq \sum_{\{x:|x-\frac{k}{m}|<\frac{\delta}{2}\}} \binom{m-1}{k} x^k (1-x)^{m-1-k} \varepsilon$$
$$+M \sum_{k=0}^{m-1} \binom{m-1}{k} \frac{4(k-mx)^2}{m^2\delta^2} x^k (1-x)^{m-1-k}$$

$$\leq \varepsilon + 4M\frac{1}{4}m\frac{1}{m^2\delta^2} = \varepsilon + M\frac{1}{m\delta^2} < 2\varepsilon$$

whenever m is large enough. Thus this proves uniform convergence. ■

Now consider the case where $n \geq 1$. Applying the same manipulations to the sum which corresponds to the i^{th} variable,

$$p_{\mathbf{m}x_i}(\mathbf{x}) \equiv \sum_{k_1=0}^{m_1} \cdots \sum_{k_i=0}^{m_i-1} \cdots \sum_{k_n=0}^{m_n} \binom{m_1}{k_1} \cdots \binom{m_i-1}{k_i} \cdots \binom{m_n}{k_n} x_1^{k_1} (1-x_1)^{m_1-k_1} \cdots$$

$$x_i^{k_i} (1-x_i)^{m_i-1-k_i} \cdots x_n^{k_n} (1-x_n)^{m_n-k_n} \frac{f\left(\frac{k_1}{m_1}, \cdots \frac{k_i+1}{m_i}, \cdots \frac{k_n}{m_n}\right) - f\left(\frac{k_1}{m_1}, \cdots \frac{k_i}{m_i}, \cdots \frac{k_n}{m_n}\right)}{1/m_i}$$

By the mean value theorem, the difference quotient is of the form

$$f_{x_i}(\mathbf{x}_{\mathbf{k},\mathbf{m}}),$$

the i^{th} component of $\mathbf{x}_{\mathbf{k},\mathbf{m}}$ being between $\frac{k_i}{m_i}$ and $\frac{k_i+1}{m_i}$. Therefore, a repeat of the above argument involving splitting the sum into two pieces, one for which $\mathbf{k}/\mathbf{m}$ is close to $\mathbf{x}$, hence close to $\mathbf{x}_{\mathbf{k},\mathbf{m}}$ and one for which some k_j/m_j is not close to x_j for some j yields the same conclusion about uniform convergence on $[0,1]^n$. This has essentially proved the following lemma.

Lemma 5.7.5 *Let f be in $C^k([0,1]^n)$. Then there exists a sequence of polynomials $p_m(\mathbf{x})$ such that each partial derivative up to order k converges uniformly to the corresponding partial derivative of f.*

Proof: It was shown above that letting $\mathbf{m} = (m_1, m_2, \cdots, m_n)$,

$$\lim_{\min(\mathbf{m})\to\infty} \|p_{\mathbf{m}} - f\|_{[0,1]^n} = 0, \lim_{\min(\mathbf{m})\to\infty} \|p_{\mathbf{m}x_i} - f_{x_i}\|_{[0,1]^n} = 0$$

for each x_i. Extending to higher derivatives is just a technical generalization of what was just shown. ■

Theorem 5.7.6 *Let f be a continuous function defined on*

$$R \equiv \prod_{k=1}^{n} [a_k, b_k].$$

Then there exists a sequence of polynomials $\{p_{\mathbf{m}}\}$ converging uniformly to f on R as $\min(\mathbf{m}) \to \infty$. If f is $C^k(R)$, then the partial derivatives of $p_{\mathbf{m}}$ up to order k converge uniformly to the corresponding partial derivatives of f.

Proof: Let $g_k : [0,1] \to [a_k, b_k]$ be linear, one to one, and onto and let

$$\mathbf{x} = \mathbf{g}(\mathbf{y}) \equiv (g_1(y_1), g_2(y_2), \cdots, g_n(y_n)).$$

Thus $\mathbf{g} : [0,1]^n \to \prod_{k=1}^n [a_k, b_k]$ is one to one, onto, and each component function is linear. Then $f \circ \mathbf{g}$ is a continuous function defined on $[0,1]^n$. It follows from Lemma 5.7.3 there exists a sequence of polynomials, $\{p_{\mathbf{m}}(\mathbf{y})\}$ each defined on $[0,1]^n$ which converges uniformly to $f \circ \mathbf{g}$ on $[0,1]^n$. Therefore, $\left\{p_{\mathbf{m}}\left(\mathbf{g}^{-1}(\mathbf{x})\right)\right\}$ converges uniformly to $f(\mathbf{x})$ on R. But

$$\mathbf{y} = (y_1, \cdots, y_n) = \left(g_1^{-1}(x_1), \cdots, g_n^{-1}(x_n)\right)$$

and each g_k^{-1} is linear. Therefore, $\left\{p_{\mathbf{m}}\left(\mathbf{g}^{-1}(\mathbf{x})\right)\right\}$ is a sequence of polynomials. As to the partial derivatives, it was shown above that

$$\lim_{\min(\mathbf{m})\to\infty} \|Dp_{\mathbf{m}} - D(f \circ \mathbf{g})\|_{[0,1]^n} = 0$$

Now the chain rule implies that

$$D\left(p_m \circ \mathbf{g}^{-1}\right)(\mathbf{x}) = Dp_m\left(\mathbf{g}^{-1}(\mathbf{x})\right) D\mathbf{g}^{-1}(\mathbf{x})$$

Therefore, the following convergences are uniform in $\mathbf{x} \in R$.

$$\begin{aligned}
& \lim_{\min(\mathbf{m})\to\infty} D\left(p_m \circ \mathbf{g}^{-1}\right)(\mathbf{x}) \\
= & \lim_{\min(\mathbf{m})\to\infty} Dp_m\left(\mathbf{g}^{-1}(\mathbf{x})\right) D\mathbf{g}^{-1}(\mathbf{x}) \\
= & \lim_{\min(\mathbf{m})\to\infty} D(f \circ \mathbf{g})\left(\mathbf{g}^{-1}(\mathbf{x})\right) D\mathbf{g}^{-1}(\mathbf{x}) \\
= & \lim_{\min(\mathbf{m})\to\infty} Df\left(\mathbf{g}\left(\mathbf{g}^{-1}(\mathbf{x})\right)\right) D\mathbf{g}\left(\mathbf{g}^{-1}(\mathbf{x})\right) D\mathbf{g}^{-1}(\mathbf{x}) \\
= & Df(\mathbf{x})
\end{aligned}$$

The claim about higher order derivatives is more technical but follows in the same way. ■

There is a more general version of this theorem which is easy to get. It depends on the Tietze extension theorem, a wonderful little result which is interesting for its own sake.

5.7.1 The Tietze Extension Theorem

To generalize the Weierstrass approximation theorem I will give a special case of the Tietze extension theorem, a very useful result in topology. When this is done, it will be possible to prove the Weierstrass approximation theorem for functions defined on a closed and bounded subset of $\mathbb{R}^n$ rather than a box.

Lemma 5.7.7 *Let $S \subseteq \mathbb{R}^n$ be a nonempty subset. Define*

$$\operatorname{dist}(\mathbf{x}, S) \equiv \inf\{|\mathbf{x} - \mathbf{y}| : \mathbf{y} \in S\}.$$

Then $\mathbf{x} \to \operatorname{dist}(\mathbf{x}, S)$ is a continuous function satisfying the inequality,

$$|\operatorname{dist}(\mathbf{x}, S) - \operatorname{dist}(\mathbf{y}, S)| \leq |\mathbf{x} - \mathbf{y}|. \tag{5.12}$$

Proof: The continuity of $\mathbf{x} \to \text{dist}(\mathbf{x}, S)$ is obvious if the inequality 5.12 is established. So let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Without loss of generality, assume $\text{dist}(\mathbf{x}, S) \geq \text{dist}(\mathbf{y}, S)$ and pick $\mathbf{z} \in S$ such that $|\mathbf{y} - \mathbf{z}| - \varepsilon < \text{dist}(\mathbf{y}, S)$. Then

$$\begin{aligned} |\text{dist}(\mathbf{x}, S) - \text{dist}(\mathbf{y}, S)| &= \text{dist}(\mathbf{x}, S) - \text{dist}(\mathbf{y}, S) \\ &\leq |\mathbf{x} - \mathbf{z}| - (|\mathbf{y} - \mathbf{z}| - \varepsilon) \\ &\leq |\mathbf{z} - \mathbf{y}| + |\mathbf{x} - \mathbf{y}| - |\mathbf{y} - \mathbf{z}| + \varepsilon = |\mathbf{x} - \mathbf{y}| + \varepsilon. \end{aligned}$$

Since ε is arbitrary, this proves 5.12. ■

Lemma 5.7.8 *Let H, K be two nonempty disjoint closed subsets of $\mathbb{R}^n$. Then there exists a continuous function, $g : \mathbb{R}^n \to [-1, 1]$ such that $g(H) = -1/3$, $g(K) = 1/3, g(\mathbb{R}^n) \subseteq [-1/3, 1/3]$.*

Proof: Let

$$f(\mathbf{x}) \equiv \frac{\text{dist}(\mathbf{x}, H)}{\text{dist}(\mathbf{x}, H) + \text{dist}(\mathbf{x}, K)}.$$

The denominator is never equal to zero because if $\text{dist}(\mathbf{x}, H) = 0$, then $\mathbf{x} \in H$ because H is closed. (To see this, pick $\mathbf{h}_k \in B(\mathbf{x}, 1/k) \cap H$. Then $\mathbf{h}_k \to \mathbf{x}$ and since H is closed, $\mathbf{x} \in H$.) Similarly, if $\text{dist}(\mathbf{x}, K) = 0$, then $\mathbf{x} \in K$ and so the denominator is never zero as claimed. Hence f is continuous and from its definition, $f = 0$ on H and $f = 1$ on K. Now let $g(\mathbf{x}) \equiv \frac{2}{3}\left(f(\mathbf{x}) - \frac{1}{2}\right)$. Then g has the desired properties. ■

Definition 5.7.9 *For f a real or complex valued bounded continuous function defined on $M \subseteq \mathbb{R}^n$.*

$$||f||_M \equiv \sup\{|f(\mathbf{x})| : \mathbf{x} \in M\}.$$

Lemma 5.7.10 *Suppose M is a closed set in $\mathbb{R}^n$ where $\mathbb{R}^n$ and suppose $f : M \to [-1, 1]$ is continuous at every point of M. Then there exists a function, g which is defined and continuous on all of $\mathbb{R}^n$ such that $||f - g||_M < \frac{2}{3}$, $g(\mathbb{R}^n) \subseteq [-1/3, 1/3]$.*

Proof: Let $H = f^{-1}([-1, -1/3]), K = f^{-1}([1/3, 1])$. Thus H and K are disjoint closed subsets of M. Suppose first H, K are both nonempty. Then by Lemma 5.7.8 there exists g such that g is a continuous function defined on all of $\mathbb{R}^n$ and $g(H) = -1/3$, $g(K) = 1/3$, and $g(\mathbb{R}^n) \subseteq [-1/3, 1/3]$. It follows $||f - g||_M < 2/3$. If $H = \emptyset$, then f has all its values in $[-1/3, 1]$ and so letting $g \equiv 1/3$, the desired condition is obtained. If $K = \emptyset$, let $g \equiv -1/3$. This proves the lemma. ■

Lemma 5.7.11 *Suppose M is a closed set in $\mathbb{R}^n$ and suppose $f : M \to [-1, 1]$ is continuous at every point of M. Then there exists a function, g which is defined and continuous on all of $\mathbb{R}^n$ such that $g = f$ on M and g has its values in $[-1, 1]$.*

Proof: Using Lemma 5.7.10, let g_1 be such that $g_1(\mathbb{R}^n) \subseteq [-1/3, 1/3]$ and

$$||f - g_1||_M \leq \frac{2}{3}.$$

Suppose $g_1, \cdots, g_m$ have been chosen such that $g_j(\mathbb{R}^n) \subseteq [-1/3, 1/3]$ and

$$\left|\left| f - \sum_{i=1}^{m} \left(\frac{2}{3}\right)^{i-1} g_i \right|\right|_M < \left(\frac{2}{3}\right)^m. \tag{5.13}$$

This has been done for $m = 1$. Then

$$\left|\left| \left(\frac{3}{2}\right)^m \left(f - \sum_{i=1}^{m} \left(\frac{2}{3}\right)^{i-1} g_i \right) \right|\right|_M \leq 1$$

and so $\left(\frac{3}{2}\right)^m \left(f - \sum_{i=1}^m \left(\frac{2}{3}\right)^{i-1} g_i\right)$ can play the role of f in the first step of the proof. Therefore, there exists g_{m+1} defined and continuous on all of $\mathbb{R}^n$ such that its values are in $[-1/3, 1/3]$ and

$$\left\| \left(\frac{3}{2}\right)^m \left(f - \sum_{i=1}^m \left(\frac{2}{3}\right)^{i-1} g_i\right) - g_{m+1} \right\|_M \leq \frac{2}{3}.$$

Hence

$$\left\| \left(f - \sum_{i=1}^m \left(\frac{2}{3}\right)^{i-1} g_i\right) - \left(\frac{2}{3}\right)^m g_{m+1} \right\|_M \leq \left(\frac{2}{3}\right)^{m+1}.$$

It follows there exists a sequence, $\{g_i\}$ such that each has its values in $[-1/3, 1/3]$ and for every m 5.13 holds. Then let

$$g(\mathbf{x}) \equiv \sum_{i=1}^{\infty} \left(\frac{2}{3}\right)^{i-1} g_i(\mathbf{x}).$$

It follows

$$|g(\mathbf{x})| \leq \left| \sum_{i=1}^{\infty} \left(\frac{2}{3}\right)^{i-1} g_i(\mathbf{x}) \right| \leq \sum_{i=1}^{m} \left(\frac{2}{3}\right)^{i-1} \frac{1}{3} \leq 1$$

and

$$\left| \left(\frac{2}{3}\right)^{i-1} g_i(\mathbf{x}) \right| \leq \left(\frac{2}{3}\right)^{i-1} \frac{1}{3}$$

so the Weierstrass M test applies and shows convergence is uniform. Therefore g must be continuous. The estimate 5.13 implies $f = g$ on M. ■

The following is the Tietze extension theorem.

Theorem 5.7.12 *Let M be a closed nonempty subset of $\mathbb{R}^n$ and let $f : M \to [a, b]$ be continuous at every point of M. Then there exists a function, g continuous on all of $\mathbb{R}^n$ which coincides with f on M such that $g(\mathbb{R}^n) \subseteq [a, b]$.*

Proof: Let $f_1(\mathbf{x}) = 1 + \frac{2}{b-a}(f(\mathbf{x}) - b)$. Then f_1 satisfies the conditions of Lemma 5.7.11 and so there exists $g_1 : \mathbb{R}^n \to [-1, 1]$ such that g is continuous on $\mathbb{R}^n$ and equals f_1 on M. Let $g(\mathbf{x}) = (g_1(\mathbf{x}) - 1)\left(\frac{b-a}{2}\right) + b$. This works. ■

With the Tietze extension theorem, here is a better version of the Weierstrass approximation theorem.

Theorem 5.7.13 *Let K be a closed and bounded subset of $\mathbb{R}^n$ and let $f : K \to \mathbb{R}$ be continuous. Then there exists a sequence of polynomials $\{p_m\}$ such that*

$$\lim_{m \to \infty} (\sup\{|f(\mathbf{x}) - p_m(\mathbf{x})| : \mathbf{x} \in K\}) = 0.$$

In other words, the sequence of polynomials converges uniformly to f on K.

Proof: By the Tietze extension theorem, there exists an extension of f to a continuous function g defined on all $\mathbb{R}^n$ such that $g = f$ on K. Now since K is bounded, there exist intervals, $[a_k, b_k]$ such that

$$K \subseteq \prod_{k=1}^{n} [a_k, b_k] = R$$

Then by the Weierstrass approximation theorem, Theorem 5.7.6 there exists a sequence of polynomials $\{p_m\}$ converging uniformly to g on R. Therefore, this sequence of polynomials converges uniformly to $g = f$ on K as well. This proves the theorem. ■

By considering the real and imaginary parts of a function which has values in $\mathbb{C}$ one can generalize the above theorem.

Corollary 5.7.14 *Let K be a closed and bounded subset of $\mathbb{R}^n$ and let $f : K \to \mathbb{F}$ be continuous. Then there exists a sequence of polynomials $\{p_m\}$ such that*

$$\lim_{m\to\infty} \left(\sup\left\{|f(\mathbf{x}) - p_m(\mathbf{x})| : \mathbf{x} \in K\right\}\right) = 0.$$

In other words, the sequence of polynomials converges uniformly to f on K.

5.8 The Operator Norm

It is important to be able to measure the size of a linear operator. The most convenient way is described in the next definition.

Definition 5.8.1 *Let V, W be two finite dimensional normed vector spaces having norms $||\cdot||_V$ and $||\cdot||_W$ respectively. Let $L \in \mathcal{L}(V, W)$. Then the operator norm of L, denoted by $||L||$ is defined as*

$$||L|| \equiv \sup\left\{||L\mathbf{x}||_W : ||\mathbf{x}||_V \leq 1\right\}.$$

Then the following theorem discusses the main properties of this norm. In the future, I will dispense with the subscript on the symbols for the norm because it is clear from the context which norm is meant. Here is a useful lemma.

Lemma 5.8.2 *Let V be a normed vector space having a basis $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$. Let*

$$A = \left\{\mathbf{a} \in \mathbb{F}^n : \left|\left|\sum_{k=1}^n a_k \mathbf{v}_k\right|\right| \leq 1\right\}$$

where $\mathbf{a} = (a_1, \cdots, a_n)$. Then A is a closed and bounded subset of $\mathbb{F}^n$.

Proof: First suppose $\mathbf{a} \notin A$. Then

$$\left|\left|\sum_{k=1}^n a_k \mathbf{v}_k\right|\right| > 1.$$

Then for $\mathbf{b} = (b_1, \cdots, b_n)$, and using the triangle inequality,

$$\left|\left|\sum_{k=1}^n b_k \mathbf{v}_k\right|\right| = \left|\left|\sum_{k=1}^n (a_k - (a_k - b_k))\, \mathbf{v}_k\right|\right|$$

$$\geq \left|\left|\sum_{k=1}^n a_k \mathbf{v}_k\right|\right| - \sum_{k=1}^n |a_k - b_k|\, ||\mathbf{v}_k||$$

and now it is apparent that if $|\mathbf{a} - \mathbf{b}|$ is sufficiently small so that each $|a_k - b_k|$ is small enough, this expression is larger than 1. Thus there exists $\delta > 0$ such that $B(\mathbf{a}, \delta) \subseteq A^C$ showing that A^C is open. Therefore, A is closed.

Next consider the claim that A is bounded. Suppose this is not so. Then there exists a sequence $\{\mathbf{a}_k\}$ of points of A,

$$\mathbf{a}_k = \left(a_k^1, \cdots, a_k^n\right),$$

such that $\lim_{k\to\infty} |\mathbf{a}_k| = \infty$. Then from the definition of A,

$$\left|\left|\sum_{j=1}^n \frac{a_k^j}{|\mathbf{a}_k|} \mathbf{v}_j\right|\right| \leq \frac{1}{|\mathbf{a}_k|}. \tag{5.14}$$

Let

$$\mathbf{b}_k = \left(\frac{a_k^1}{|\mathbf{a}_k|}, \cdots, \frac{a_k^n}{|\mathbf{a}_k|} \right)$$

Then $|\mathbf{b}_k| = 1$ so $\mathbf{b}_k$ is contained in the closed and bounded set, $S(\mathbf{0}, 1)$ which is sequentially compact in $\mathbb{F}^n$. It follows there exists a subsequence, still denoted by $\{\mathbf{b}_k\}$ such that it converges to $\mathbf{b} \in S(\mathbf{0},1)$. Passing to the limit in 5.14 using the following inequality,

$$\left|\left| \sum_{j=1}^n \frac{a_k^j}{|\mathbf{a}_k|} \mathbf{v}_j - \sum_{j=1}^n b_j \mathbf{v}_j \right|\right| \leq \sum_{j=1}^n \left| \frac{a_k^j}{|\mathbf{a}_k|} - b_j \right| ||\mathbf{v}_j||$$

to see that the sum converges to $\sum_{j=1}^n b_j \mathbf{v}_j$, it follows

$$\sum_{j=1}^n b_j \mathbf{v}_j = \mathbf{0}$$

and this is a contradiction because $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis and not all the b_j can equal zero. Therefore, A must be bounded after all. This proves the lemma. ■

Theorem 5.8.3 *The operator norm has the following properties.*

1. $||L|| < \infty$
2. *For all* $\mathbf{x} \in X, ||L\mathbf{x}|| \leq ||L||\,||\mathbf{x}||$ *and if* $L \in \mathcal{L}(V,W)$ *while* $M \in \mathcal{L}(W,Z)$, *then* $||ML|| \leq ||M||\,||L||$.
3. $||\cdot||$ *is a norm. In particular,*
   (a) $||L|| \geq 0$ *and* $||L|| = 0$ *if and only if* $L = 0$, *the linear transformation which sends every vector to* $\mathbf{0}$.
   (b) $||aL|| = |a|\,||L||$ *whenever* $a \in \mathbb{F}$
   (c) $||L + M|| \leq ||L|| + ||M||$
4. *If* $L \in \mathcal{L}(V,W)$ *for* V, W *normed vector spaces,* L *is continuous, meaning that* $L^{-1}(U)$ *is open whenever* U *is an open set in* W.

Proof: First consider 1.). Let A be as in the above lemma. Then

$$\begin{aligned} ||L|| &\equiv \sup \left\{ \left|\left| L\left(\sum_{j=1}^n a_j \mathbf{v}_j \right) \right|\right| : \mathbf{a} \in A \right\} \\ &= \sup \left\{ \left|\left| \sum_{j=1}^n a_j L(\mathbf{v}_j) \right|\right| : \mathbf{a} \in A \right\} < \infty \end{aligned}$$

because $\mathbf{a} \to \left|\left| \sum_{j=1}^n a_j L(\mathbf{v}_j) \right|\right|$ is a real valued continuous function defined on a sequentially compact set and so it achieves its maximum.

Next consider 2.). If $\mathbf{x} = \mathbf{0}$ there is nothing to show. Assume $\mathbf{x} \neq \mathbf{0}$. Then from the definition of $||L||$,

$$\left|\left| L\left(\frac{\mathbf{x}}{||\mathbf{x}||} \right) \right|\right| \leq ||L||$$

and so, since L is linear, you can multiply on both sides by $||\mathbf{x}||$ and conclude

$$||L(\mathbf{x})|| \leq ||L||\,||\mathbf{x}||.$$

For the other claim,

$$\begin{aligned} ||ML|| &\equiv \sup\{||ML(\mathbf{x})|| : ||\mathbf{x}|| \leq 1\} \\ &\leq ||M|| \sup\{||L\mathbf{x}|| : ||\mathbf{x}|| \leq 1\} \equiv ||M||\,||L||. \end{aligned}$$

Finally consider 3.) If $||L|| = 0$ then from 2.), $||L\mathbf{x}|| \leq 0$ and so $L\mathbf{x} = \mathbf{0}$ for every $\mathbf{x}$ which is the same as saying $L = 0$. If $L\mathbf{x} = \mathbf{0}$ for every $\mathbf{x}$, then $L = 0$ by definition. Let $a \in \mathbb{F}$. Then from the properties of the norm, in the vector space,

$$\begin{aligned} ||aL|| &\equiv \sup\{||aL\mathbf{x}|| : ||\mathbf{x}|| \leq 1\} \\ &= \sup\{|a|\,||L\mathbf{x}|| : ||\mathbf{x}|| \leq 1\} \\ &= |a| \sup\{||L\mathbf{x}|| : ||\mathbf{x}|| \leq 1\} \equiv |a|\,||L|| \end{aligned}$$

Finally consider the triangle inequality.

$$||L + M|| \equiv \sup\{||L\mathbf{x} + M\mathbf{x}|| : ||\mathbf{x}|| \leq 1\}$$

$$\leq \sup\{||M\mathbf{x}|| + ||L\mathbf{x}|| : ||\mathbf{x}|| \leq 1\}$$

$$\leq \sup\{||L\mathbf{x}|| : ||\mathbf{x}|| \leq 1\} + \sup\{||M\mathbf{x}|| : ||\mathbf{x}|| \leq 1\}$$

because $||L\mathbf{x}|| \leq \sup\{||L\mathbf{x}|| : ||\mathbf{x}|| \leq 1\}$ with a similar inequality holding for M. Therefore, by definition,

$$||L + M|| \leq ||L|| + ||M||.$$

Finally consider 4.). Let $L \in \mathcal{L}(V, W)$ and let U be open in W and $\mathbf{v} \in L^{-1}(U)$. Thus since U is open, there exists $\delta > 0$ such that

$$L(\mathbf{v}) \in B(L(\mathbf{v}), \delta) \subseteq U.$$

Then if $\mathbf{w} \in V$,

$$||L(\mathbf{v} - \mathbf{w})|| = ||L(\mathbf{v}) - L(\mathbf{w})|| \leq ||L||\,||\mathbf{v} - \mathbf{w}||$$

and so if $||\mathbf{v} - \mathbf{w}||$ is sufficiently small, $||\mathbf{v} - \mathbf{w}|| < \delta/||L||$, then $L(\mathbf{w}) \in B(L(\mathbf{v}), \delta)$ which shows $B(\mathbf{v}, \delta/||L||) \subseteq L^{-1}(U)$ and since $\mathbf{v} \in L^{-1}(U)$ was arbitrary, this shows $L^{-1}(U)$ is open. This proves the theorem. ■

The operator norm will be very important in the chapter on the derivative.

Part 1.) of Theorem 5.8.3 says that if $L \in \mathcal{L}(V, W)$ where V and W are two normed vector spaces, then there exists K such that for all $\mathbf{v} \in V$,

$$||L\mathbf{v}||_W \leq K\,||\mathbf{v}||_V$$

An obvious case is to let $L = \mathrm{id}$, the identity map on V and let there be two different norms on $V, ||\cdot||_1$ and $||\cdot||_2$. Thus $(V, ||\cdot||_1)$ is a normed vector space and so is $(V, ||\cdot||_2)$. Then Theorem 5.8.3 implies that

$$||\mathbf{v}||_2 = ||\mathrm{id}(\mathbf{v})||_2 \leq K_2\,||\mathbf{v}||_1 \tag{5.15}$$

while the same reasoning implies there exists K_1 such that

$$||\mathbf{v}||_1 \leq K_1\,||\mathbf{v}||_2. \tag{5.16}$$

This leads to the following important theorem.

Theorem 5.8.4 *Let V be a finite dimensional vector space and let $||\cdot||_1$ and $||\cdot||_2$ be two norms for V. Then these norms are equivalent which means there exist constants, δ, Δ such that for all $\mathbf{v} \in V$*

$$\delta ||\mathbf{v}||_1 \leq ||\mathbf{v}||_2 \leq \Delta ||\mathbf{v}||_1$$

A set, K is sequentially compact if and only if it is closed and bounded. Also every finite dimensional normed vector space is complete. Also any closed and bounded subset of a finite dimensional normed vector space is sequentially compact.

Proof: From 5.15 and 5.16

$$||\mathbf{v}||_1 \leq K_1 ||\mathbf{v}||_2 \leq K_1 K_2 ||\mathbf{v}||_1$$

and so

$$\frac{1}{K_1} ||\mathbf{v}||_1 \leq ||\mathbf{v}||_2 \leq K_2 ||\mathbf{v}||_1 .$$

Next consider the claim that all closed and bounded sets in a normed vector space are sequentially compact. Let $L : \mathbb{F}^n \to V$ be defined by

$$L(\mathbf{a}) \equiv \sum_{k=1}^{n} a_k \mathbf{v}_k$$

where $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis for V. Thus $L \in \mathcal{L}(\mathbb{F}^n, V)$ and so by Theorem 5.8.3 this is a continuous function. Hence if K is a closed and bounded subset of V it follows

$$L^{-1}(K) = \mathbb{F}^n \setminus L^{-1}\left(K^C\right) = \mathbb{F}^n \setminus (\text{an open set}) = \text{a closed set.}$$

Also $L^{-1}(K)$ is bounded. To see this, note that L^{-1} is one to one onto V and so $L^{-1} \in \mathcal{L}(V, \mathbb{F}^n)$. Therefore,

$$\left|L^{-1}(\mathbf{v})\right| \leq \left|\left|L^{-1}\right|\right| ||\mathbf{v}|| \leq \left|\left|L^{-1}\right|\right| r$$

where $K \subseteq B(\mathbf{0}, r)$. Since K is bounded, such an r exists. Thus $L^{-1}(K)$ is a closed and bounded subset of $\mathbb{F}^n$ and is therefore sequentially compact. It follows that if $\{\mathbf{v}_k\}_{k=1}^{\infty} \subseteq K$, there is a subsequence $\{\mathbf{v}_{k_l}\}_{l=1}^{\infty}$ such that $\left\{L^{-1}\mathbf{v}_{k_l}\right\}$ converges to a point, $\mathbf{a} \in L^{-1}(K)$. Hence by continuity of L,

$$\mathbf{v}_{k_l} = L\left(L^{-1}(\mathbf{v}_{k_l})\right) \to L\mathbf{a} \in K.$$

Conversely, suppose K is sequentially compact. I need to verify it is closed and bounded. If it is not closed, then it is missing a limit point, $\mathbf{k}_0$. Since $\mathbf{k}_0$ is a limit point, there exists $\mathbf{k}_n \in B\left(\mathbf{k}_0, \frac{1}{n}\right)$ such that $\mathbf{k}_n \neq \mathbf{k}_0$. Therefore, $\{\mathbf{k}_n\}$ has no limit point in K because $\mathbf{k}_0 \notin K$. It follows K must be closed. If K is not bounded, then you could pick $\mathbf{k}_n \in K$ such that $\mathbf{k}_m \notin B(\mathbf{0}, m)$ and it follows $\{\mathbf{k}_k\}$ cannot have a subsequence which converges because if $\mathbf{k} \in K$, then for large enough m, $\mathbf{k} \in B(\mathbf{0}, m/2)$ and so if $\left\{\mathbf{k}_{k_j}\right\}$ is any subsequence, $\mathbf{k}_{k_j} \notin B(\mathbf{0}, m)$ for all but finitely many j. In other words, for any $\mathbf{k} \in K$, it is not the limit of any subsequence. Thus K must also be bounded.

Finally consider the claim about completeness. Let $\{\mathbf{v}_k\}_{k=1}^{\infty}$ be a Cauchy sequence in V. Since L^{-1}, defined above is in $\mathcal{L}(V, \mathbb{F}^n)$, it follows $\left\{L^{-1}\mathbf{v}_k\right\}_{k=1}^{\infty}$ is a Cauchy sequence in $\mathbb{F}^n$. This follows from the inequality,

$$\left|L^{-1}\mathbf{v}_k - L^{-1}\mathbf{v}_l\right| \leq \left|\left|L^{-1}\right|\right| ||\mathbf{v}_k - \mathbf{v}_l|| .$$

therefore, there exists $\mathbf{a} \in \mathbb{F}^n$ such that $L^{-1}\mathbf{v}_k \to \mathbf{a}$ and since L is continuous,

$$\mathbf{v}_k = L\left(L^{-1}(\mathbf{v}_k)\right) \to L(\mathbf{a}) .$$

Next suppose K is a closed and bounded subset of V and let $\{\mathbf{x}_k\}_{k=1}^{\infty}$ be a sequence of vectors in K. Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis for V and let

$$\mathbf{x}_k = \sum_{j=1}^{n} x_k^j \mathbf{v}_j$$

Define a norm for V according to

$$||\mathbf{x}||^2 \equiv \sum_{j=1}^{n} \left|x^j\right|^2, \mathbf{x} = \sum_{j=1}^{n} x^j \mathbf{v}_j$$

It is clear most axioms of a norm hold. The triangle inequality also holds because by the triangle inequality for $\mathbb{F}^n$,

$$||\mathbf{x}+\mathbf{y}|| \equiv \left(\sum_{j=1}^{n} \left|x^j + y^j\right|^2\right)^{1/2}$$

$$\leq \left(\sum_{j=1}^{n} \left|x^j\right|^2\right)^{1/2} + \left(\sum_{j=1}^{n} \left|y^j\right|^2\right)^{1/2} \equiv ||\mathbf{x}|| + ||\mathbf{y}||.$$

By the first part of this theorem, this norm is equivalent to the norm on V. Thus K is closed and bounded with respect to this new norm. It follows that for each j, $\left\{x_k^j\right\}_{k=1}^{\infty}$ is a bounded sequence in $\mathbb{F}$ and so by the theorems about sequential compactness in $\mathbb{F}$ it follows upon taking subsequences n times, there exists a subsequence $\mathbf{x}_{k_l}$ such that for each j,

$$\lim_{l\to\infty} x_{k_l}^j = x^j$$

for some x^j. Hence

$$\lim_{l\to\infty} \mathbf{x}_{k_l} = \lim_{l\to\infty} \sum_{j=1}^{n} x_{k_l}^j \mathbf{v}_j = \sum_{j=1}^{n} x^j \mathbf{v}_j \in K$$

because K is closed. This proves the theorem. ■

Example 5.8.5 *Let V be a vector space and let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis. Define a norm on V as follows. For $\mathbf{v} = \sum_{k=1}^{n} a_k \mathbf{v}_k$,*

$$||\mathbf{v}|| \equiv \max\{|a_k| : k = 1, \cdots, n\}$$

In the above example, this is a norm on the vector space, V. It is clear $||a\mathbf{v}|| = |a|\,||\mathbf{v}||$ and that $||\mathbf{v}|| \geq 0$ and equals 0 if and only if $\mathbf{v} = \mathbf{0}$. The hard part is the triangle inequality. Let $\mathbf{v} = \sum_{k=1}^{n} a_k \mathbf{v}_k$ and $\mathbf{w} = \mathbf{v} = \sum_{k=1}^{n} b_k \mathbf{v}_k$.

$$\begin{aligned} ||\mathbf{v}+\mathbf{w}|| &\equiv \max_k \{|a_k + b_k|\} \leq \max_k \{|a_k| + |b_k|\} \\ &\leq \max_k |a_k| + \max_k |b_k| \equiv ||\mathbf{v}|| + ||\mathbf{w}||. \end{aligned}$$

This shows this is indeed a norm.

5.9 Ascoli Arzela Theorem

Let $\{\mathbf{f}_k\}_{k=1}^{\infty}$ be a sequence of functions defined on a compact set which have values in a finite dimensional normed vector space V. The following definition will be of the norm of such a function.

Definition 5.9.1 *Let $\mathbf{f} : K \to V$ be a continuous function which has values in a finite dimensional normed vector space V where here K is a compact set contained in some normed vector space. Define*

$$||\mathbf{f}|| \equiv \sup \{||\mathbf{f}(\mathbf{x})||_V : \mathbf{x} \in K\}.$$

Denote the set of such functions by $C(K;V)$.

Proposition 5.9.2 *The above definition yields a norm and in fact $C(K;V)$ is a complete normed linear space.*

Proof: This is obviously a vector space. Just verify the axioms. The main thing to show it that the above is a norm. First note that $||\mathbf{f}|| = 0$ if and only if $\mathbf{f} = \mathbf{0}$ and $||\alpha\mathbf{f}|| = |\alpha|\,||\mathbf{f}||$ whenever $\alpha \in \mathbb{F}$, the field of scalars, $\mathbb{C}$ or $\mathbb{R}$. As to the triangle inequality,

$$||\mathbf{f}+\mathbf{g}|| \equiv \sup \{||(\mathbf{f}+\mathbf{g})(\mathbf{x})|| : \mathbf{x} \in K\}$$

$$\begin{aligned} &\leq \sup \{||\mathbf{f}(\mathbf{x})||_V : \mathbf{x} \in K\} + \sup \{||\mathbf{g}(\mathbf{x})||_V : \mathbf{x} \in K\} \\ &\equiv ||\mathbf{g}|| + ||\mathbf{f}|| \end{aligned}$$

Furthermore, the function $\mathbf{x} \to ||\mathbf{f}(\mathbf{x})||_V$ is continuous thanks to the triangle inequality which implies

$$|||\mathbf{f}(\mathbf{x})||_V - ||\mathbf{f}(\mathbf{y})||_V| \leq ||\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})||_V.$$

Therefore, $||\mathbf{f}||$ is a well defined nonnegative real number.

It remains to verify completeness. Suppose then $\{\mathbf{f}_k\}$ is a Cauchy sequence with respect to this norm. Then from the definition it is a uniformly Cauchy sequence and since by Theorem 5.8.4 V is a complete normed vector space, it follows from Theorem 5.5.6, there exists $\mathbf{f} \in C(K;V)$ such that $\{\mathbf{f}_k\}$ converges uniformly to $\mathbf{f}$. That is,

$$\lim_{k\to\infty} ||\mathbf{f} - \mathbf{f}_k|| = 0.$$

This proves the proposition. ■

Theorem 5.8.4 says that closed and bounded sets in a finite dimensional normed vector space V are sequentially compact. This theorem typically doesn't apply to $C(K;V)$ because generally this is not a **finite dimensional** vector space although as shown above it is a complete normed vector space. It turns out you need more than closed and bounded in order to have a subset of $C(K;V)$ be sequentially compact.

Definition 5.9.3 *Let $\mathcal{F} \subseteq C(K;V)$. Then $\mathcal{F}$ is equicontinuous if for every $\varepsilon > 0$ there exists $\delta > 0$ such that whenever $|\mathbf{x} - \mathbf{y}| < \delta$, it follows*

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < \varepsilon$$

for all $\mathbf{f} \in \mathcal{F}$. $\mathcal{F}$ is bounded if there exists C such that

$$||\mathbf{f}|| \leq C$$

for all $\mathbf{f} \in \mathcal{F}$.

Lemma 5.9.4 *Let K be a sequentially compact nonempty subset of a finite dimensional normed vector space. Then there exists a countable set $D \equiv \{k_i\}_{i=1}^{\infty}$ such that for every $\varepsilon > 0$ and for every $\mathbf{x} \in K$,*

$$B(\mathbf{x}, \varepsilon) \cap D \neq \emptyset.$$

Proof: Let $n \in \mathbb{N}$. Pick $k_1^n \in K$. If $B\left(k_1^n, \frac{1}{n}\right) \supseteq K$, stop. Otherwise pick

$$k_2^n \in K \setminus B\left(k_1^n, \frac{1}{n}\right)$$

Continue this way till the process ends. It must end because if it didn't, there would exist a convergent subsequence which would imply two of the k_j^n would have to be closer than $1/n$ which is impossible from the construction. Denote this collection of points by D_n. Then $D \equiv \cup_{n=1}^{\infty} D_n$. This must work because if $\varepsilon > 0$ is given and $\mathbf{x} \in K$, let $1/n < \varepsilon/3$ and the construction implies $\mathbf{x} \in B(k_i^n, 1/n)$ for some $k_i^n \in D_n \cup D$. Then

$$k_i^n \in B(\mathbf{x}, \varepsilon).$$

D is countable because it is the countable union of finite sets. This proves the lemma. ■

Definition 5.9.5 *More generally, if K is any subset of a normed vector space and there exists D such that D is countable and for all $\mathbf{x} \in K$,*

$$B(\mathbf{x}, \varepsilon) \cap D \neq \emptyset$$

then K is called separable.

Now here is another remarkable result about equicontinuous functions.

Lemma 5.9.6 *Suppose $\{\mathbf{f}_k\}_{k=1}^{\infty}$ is equicontinuous and the functions are defined on a sequentially compact set K. Suppose also for each $\mathbf{x} \in K$,*

$$\lim_{k \to \infty} \mathbf{f}_k(\mathbf{x}) = \mathbf{f}(\mathbf{x}).$$

Then in fact $\mathbf{f}$ is continuous and the convergence is uniform. That is

$$\lim_{k \to \infty} ||\mathbf{f}_k - \mathbf{f}|| = 0.$$

Proof: Uniform convergence would say that for every $\varepsilon > 0$, there exists n_ε such that if $k, l \geq n_\varepsilon$, then for all $\mathbf{x} \in K$,

$$||\mathbf{f}_k(\mathbf{x}) - \mathbf{f}_l(\mathbf{x})|| < \varepsilon.$$

Thus if the given sequence does not converge uniformly, there exists $\varepsilon > 0$ such that for all n, there exists $k, l \geq n$ and $\mathbf{x}_n \in K$ such that

$$||\mathbf{f}_k(\mathbf{x}_n) - \mathbf{f}_l(\mathbf{x}_n)|| \geq \varepsilon$$

Since K is sequentially compact, there exists a subsequence, still denoted by $\{\mathbf{x}_n\}$ such that $\lim_{n \to \infty} \mathbf{x}_n = \mathbf{x} \in K$. Then letting k, l be associated with n as just described,

$$\begin{aligned} \varepsilon \leq\ & ||\mathbf{f}_k(\mathbf{x}_n) - \mathbf{f}_l(\mathbf{x}_n)||_V \leq ||\mathbf{f}_k(\mathbf{x}_n) - \mathbf{f}_k(\mathbf{x})||_V \\ & + ||\mathbf{f}_k(\mathbf{x}) - \mathbf{f}_l(\mathbf{x})||_V + ||\mathbf{f}_l(\mathbf{x}) - \mathbf{f}_l(\mathbf{x}_n)||_V \end{aligned}$$

By equicontinuity, if n is large enough, this implies

$$\varepsilon < \frac{\varepsilon}{3} + ||\mathbf{f}_k(\mathbf{x}) - \mathbf{f}_l(\mathbf{x})||_V + \frac{\varepsilon}{3}$$

and now taking n still larger if necessary, the middle term on the right in the above is also less than $\varepsilon/3$ which yields a contradiction. Hence convergence is uniform and so it follows from Theorem 5.5.6 the function $\mathbf{f}$ is actually continuous and

$$\lim_{k\to\infty} ||\mathbf{f} - \mathbf{f}_k|| = 0.$$

This proves the lemma. ■

The Ascoli Arzela theorem is the following.

Theorem 5.9.7 *Let K be a closed and bounded subset of a finite dimensional normed vector space and let $\mathcal{F} \subseteq C(K;V)$ where V is a finite dimensional normed vector space. Suppose also that $\mathcal{F}$ is bounded and equicontinuous. Then if $\{\mathbf{f}_k\}_{k=1}^{\infty} \subseteq \mathcal{F}$, there exists a subsequence $\{\mathbf{f}_{k_l}\}_{l=1}^{\infty}$ which converges to a function $\mathbf{f} \in C(K;V)$ in the sense that*

$$\lim_{l\to\infty} ||\mathbf{f} - \mathbf{f}_{k_l}||$$

Proof: Denote by $\left\{\mathbf{f}_{(k,n)}\right\}_{n=1}^{\infty}$ a subsequence of $\left\{\mathbf{f}_{(k-1,n)}\right\}_{n=1}^{\infty}$ where the index denoted by $(k-1,k-1)$ is always less than the index denoted by (k,k). Also let the countable dense subset of Lemma 5.9.4 be $D = \{\mathbf{d}_k\}_{k=1}^{\infty}$. Then consider the following diagram.

$$\begin{array}{c}
\mathbf{f}_{(1,1)}, \mathbf{f}_{(1,2)}, \mathbf{f}_{(1,3)}, \mathbf{f}_{(1,4)}, \cdots \to \mathbf{d}_1 \\
\mathbf{f}_{(2,1)}, \mathbf{f}_{(2,2)}, \mathbf{f}_{(2,3)}, \mathbf{f}_{(2,4)}, \cdots \to \mathbf{d}_1, \mathbf{d}_2 \\
\mathbf{f}_{(3,1)}, \mathbf{f}_{(3,2)}, \mathbf{f}_{(3,3)}, \mathbf{f}_{(3,4)}, \cdots \to \mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3 \\
\mathbf{f}_{(4,1)}, \mathbf{f}_{(4,2)}, \mathbf{f}_{(4,3)}, \mathbf{f}_{(4,4)}, \cdots \to \mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_4 \\
\vdots
\end{array}$$

The meaning is as follows. $\left\{\mathbf{f}_{(1,k)}\right\}_{k=1}^{\infty}$ is a subsequence of the original sequence which converges at $\mathbf{d}_1$. Such a subsequence exists because $\{\mathbf{f}_k(\mathbf{d}_1)\}_{k=1}^{\infty}$ is contained in a bounded set so a subsequence converges by Theorem 5.8.4. (It is given to be in a bounded set and so the closure of this bounded set is both closed and bounded, hence weakly compact.) Now $\left\{\mathbf{f}_{(2,k)}\right\}_{k=1}^{\infty}$ is a subsequence of the first subsequence which converges, at $\mathbf{d}_2$. Then by Theorem 4.1.6 this new subsequence continues to converge at $\mathbf{d}_1$. Thus, as indicated by the diagram, it converges at both $\mathbf{b}_1$ and $\mathbf{b}_2$. Continuing this way explains the meaning of the diagram. Now consider the subsequence of the original sequence $\left\{\mathbf{f}_{(k,k)}\right\}_{k=1}^{\infty}$. For $k \geq n$, this subsequence is a subsequence of the subsequence $\{\mathbf{f}_{n,k}\}_{k=1}^{\infty}$ and so it converges at $\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3 \cdots, \mathbf{d}_4$. This being true for all n, it follows $\left\{\mathbf{f}_{(k,k)}\right\}_{k=1}^{\infty}$ converges at every point of D. To save on notation, I shall simply denote this as $\left\{\mathbf{f}_k\right\}$.

Then letting $\mathbf{d} \in D$,

$$\begin{aligned}
||\mathbf{f}_k(\mathbf{x}) - \mathbf{f}_l(\mathbf{x})||_V &\leq ||\mathbf{f}_k(\mathbf{x}) - \mathbf{f}_k(\mathbf{d})||_V \\
&\quad + ||\mathbf{f}_k(\mathbf{d}) - \mathbf{f}_l(\mathbf{d})||_V + ||\mathbf{f}_l(\mathbf{d}) - \mathbf{f}_l(\mathbf{x})||_V
\end{aligned}$$

Picking $\mathbf{d}$ close enough to $\mathbf{x}$ and applying equicontinuity,

$$||\mathbf{f}_k(\mathbf{x}) - \mathbf{f}_l(\mathbf{x})||_V < 2\varepsilon/3 + ||\mathbf{f}_k(\mathbf{d}) - \mathbf{f}_l(\mathbf{d})||_V$$

Thus for k, l large enough, the right side is less than ε. This shows that for each $\mathbf{x} \in K$, $\{\mathbf{f}_k(\mathbf{x})\}_{k=1}^{\infty}$ is a Cauchy sequence and so by completeness of V this converges. Let $\mathbf{f}(\mathbf{x})$ be the thing to which it converges. Then $\mathbf{f}$ is continuous and the convergence is uniform by Lemma 5.9.6. This proves the theorem. ■

5.10 Exercises

1. In Theorem 5.7.6 it is assumed f has values in $\mathbb{F}$. Show there is no change if f has values in V, a normed vector space provided you redefine the definition of a polynomial to be something of the form $\sum_{|\boldsymbol{\alpha}|\leq m} a_\alpha \mathbf{x}^{\boldsymbol{\alpha}}$ where $a_{\boldsymbol{\alpha}} \in V$.

2. How would you generalize the conclusion of Corollary 5.7.14 to include the situation where f has values in a finite dimensional normed vector space?

3. Recall the Bernstein polynomials

$$p_{\mathbf{m}}(\mathbf{x}) = \sum_{\mathbf{k}\leq\mathbf{m}} \binom{\mathbf{m}}{\mathbf{k}} \mathbf{x}^{\mathbf{k}} (\mathbf{1}-\mathbf{x})^{\mathbf{m}-\mathbf{k}} f\left(\frac{\mathbf{k}}{\mathbf{m}}\right)$$

   It was shown that these converge uniformly to f provided $\min(\mathbf{m}) \to \infty$. Explain why it suffices to have $\max(\mathbf{m}) \to \infty$. See the estimate which was derived.

4. If $\{\mathbf{f}_n\}$ and $\{\mathbf{g}_n\}$ are sequences of $\mathbb{F}^n$ valued functions defined on D which converge uniformly, show that if a, b are constants, then $a\mathbf{f}_n + b\mathbf{g}_n$ also converges uniformly. If there exists a constant, M such that $|\mathbf{f}_n(\mathbf{x})|, |\mathbf{g}_n(\mathbf{x})| < M$ for all n and for all $\mathbf{x} \in D$, show $\{\mathbf{f}_n \cdot \mathbf{g}_n\}$ converges uniformly. Let $f_n(\mathbf{x}) \equiv 1/|\mathbf{x}|$ for $\mathbf{x} \in B(\mathbf{0},1)$ and let $g_n(\mathbf{x}) \equiv (n-1)/n$. Show $\{f_n\}$ converges uniformly on $B(\mathbf{0},1)$ and $\{g_n\}$ converges uniformly but $\{f_n g_n\}$ fails to converge uniformly.

5. Formulate a theorem for series of functions of n variables which will allow you to conclude the infinite series is uniformly continuous based on reasonable assumptions about the functions in the sum.

6. If f and g are real valued functions which are continuous on some set, D, show that

$$\min(f,g), \max(f,g)$$

   are also continuous. Generalize this to any finite collection of continuous functions. **Hint:** Note $\max(f,g) = \frac{|f-g|+f+g}{2}$. Now recall the triangle inequality which can be used to show $|\cdot|$ is a continuous function.

7. Find an example of a sequence of continuous functions defined on $\mathbb{R}^n$ such that each function is nonnegative and each function has a maximum value equal to 1 but the sequence of functions converges to 0 pointwise on $\mathbb{R}^n \setminus \{\mathbf{0}\}$, that is, the set of vectors in $\mathbb{R}^n$ excluding $\mathbf{0}$.

8. Theorem 5.3.14 says an open subset U of $\mathbb{R}^n$ is arcwise connected if and only if U is connected. Consider the usual Cartesian coordinates relative to axes $x_1, \cdots, x_n$. A square curve is one consisting of a succession of straight line segments each of which is parallel to some coordinate axis. Show an open subset U of $\mathbb{R}^n$ is connected if and only if every two points can be joined by a square curve.

9. Let $\mathbf{x} \to h(\mathbf{x})$ be a bounded continuous function. Show the function $f(\mathbf{x}) = \sum_{n=1}^{\infty} \frac{h(n\mathbf{x})}{n^2}$ is continuous.

10. Let S be a any countable subset of $\mathbb{R}^n$. Show there exists a function, $\mathbf{f}$ defined on $\mathbb{R}^n$ which is discontinuous at every point of S but continuous everywhere else. **Hint:** This is real easy if you do the right thing. It involves the Weierstrass M test.

11. By Theorem 5.7.13 there exists a sequence of polynomials converging uniformly to $f(\mathbf{x}) = |\mathbf{x}|$ on $R \equiv \prod_{k=1}^{n} [-M, M]$. Show there exists a sequence of polynomials, $\{p_n\}$ converging uniformly to f on R which has the additional property that for all n, $p_n(\mathbf{0}) = 0$.

12. If f is any continuous function defined on K a sequentially compact subset of $\mathbb{R}^n$, show there exists a series of the form $\sum_{k=1}^{\infty} p_k$, where each p_k is a polynomial, which converges uniformly to f on $[a, b]$. **Hint:** You should use the Weierstrass approximation theorem to obtain a sequence of polynomials. Then arrange it so the limit of this sequence is an infinite sum.

13. A function $\mathbf{f}$ is Holder continuous if there exists a constant, K such that

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| \leq K |\mathbf{x} - \mathbf{y}|^{\alpha}$$

for some $\alpha \leq 1$ for all $\mathbf{x}, \mathbf{y}$. Show every Holder continuous function is uniformly continuous.

14. Consider $f(\mathbf{x}) \equiv \operatorname{dist}(\mathbf{x}, S)$ where S is a nonempty subset of $\mathbb{R}^n$. Show f is uniformly continuous.

15. Let K be a sequentially compact set in a normed vector space V and let $\mathbf{f} : V \to W$ be continuous where W is also a normed vector space. Show $\mathbf{f}(K)$ is also sequentially compact.

16. If $\mathbf{f}$ is uniformly continuous, does it follow that $|\mathbf{f}|$ is also uniformly continuous? If $|\mathbf{f}|$ is uniformly continuous does it follow that $\mathbf{f}$ is uniformly continuous? Answer the same questions with "uniformly continuous" replaced with "continuous". Explain why.

17. Let $\mathbf{f} : D \to \mathbb{R}$ be a function. This function is said to be lower semicontinuous[1] at $\mathbf{x} \in D$ if for any sequence $\{\mathbf{x}_n\} \subseteq D$ which converges to $\mathbf{x}$ it follows

$$\mathbf{f}(\mathbf{x}) \leq \lim_{n \to \infty} \inf \mathbf{f}(\mathbf{x}_n).$$

Suppose D is sequentially compact and $\mathbf{f}$ is lower semicontinuous at every point of D. Show that then $\mathbf{f}$ achieves its minimum on D.

18. Let $f : D \to \mathbb{R}$ be a function. This function is said to be upper semicontinuous at $\mathbf{x} \in D$ if for any sequence $\{\mathbf{x}_n\} \subseteq D$ which converges to $\mathbf{x}$ it follows

$$f(\mathbf{x}) \geq \lim_{n \to \infty} \sup f(\mathbf{x}_n).$$

Suppose D is sequentially compact and f is upper semicontinuous at every point of D. Show that then f achieves its maximum on D.

19. Show that a real valued function defined on $D \subseteq \mathbb{R}^n$ is continuous if and only if it is both upper and lower semicontinuous.

20. Show that a real valued lower semicontinuous function defined on a sequentially compact set achieves its minimum and that an upper semicontinuous function defined on a sequentially compact set achieves its maximum.

21. Give an example of a lower semicontinuous function defined on $\mathbb{R}^n$ which is not continuous and an example of an upper semicontinuous function which is not continuous.

[1] The notion of lower semicontinuity is very important for functions which are defined on infinite dimensional sets. In more general settings, one formulates the concept differently.

22. Suppose $\{f_\alpha : \alpha \in \Lambda\}$ is a collection of continuous functions. Let

$$F(\mathbf{x}) \equiv \inf \{f_\alpha(\mathbf{x}) : \alpha \in \Lambda\}$$

Show F is an upper semicontinuous function. Next let

$$G(\mathbf{x}) \equiv \sup \{f_\alpha(\mathbf{x}) : \alpha \in \Lambda\}$$

Show G is a lower semicontinuous function.

23. Let f be a function. $\text{epi}(f)$ is defined as

$$\{(\mathbf{x}, y) : y \geq \mathbf{f}(\mathbf{x})\}.$$

It is called the epigraph of f. We say $\text{epi}(f)$ is closed if whenever $(\mathbf{x}_n, y_n) \in \text{epi}(f)$ and $\mathbf{x}_n \to \mathbf{x}$ and $y_n \to y$, it follows $(\mathbf{x}, y) \in \text{epi}(f)$. Show f is lower semicontinuous if and only if $\text{epi}(f)$ is closed. What would be the corresponding result equivalent to upper semicontinuous?

24. The operator norm was defined for $\mathcal{L}(V, W)$ above. This is the usual norm used for this vector space of linear transformations. Show that any other norm used on $\mathcal{L}(V, W)$ is equivalent to the operator norm. That is, show that if $||\cdot||_1$ is another norm, there exist scalars δ, Δ such that

$$\delta ||L|| \leq ||L||_1 \leq \Delta ||L||$$

for all $L \in \mathcal{L}(V, W)$ where here $||\cdot||$ denotes the operator norm.

25. One alternative norm which is very popular is as follows. Let $L \in \mathcal{L}(V, W)$ and let (l_{ij}) denote the matrix of L with respect to some bases. Then the Frobenius norm is defined by

$$\left(\sum_{ij} |l_{ij}|^2 \right)^{1/2} \equiv ||L||_F.$$

Show this is a norm. Other norms are of the form

$$\left(\sum_{ij} |l_{ij}|^p \right)^{1/p}$$

where $p \geq 1$ or even

$$||L||_\infty = \max_{ij} |l_{ij}|.$$

Show these are also norms.

26. Explain why $\mathcal{L}(V, W)$ is always a complete normed vector space whenever V, W are finite dimensional normed vector spaces for any choice of norm for $\mathcal{L}(V, W)$. Also explain why every closed and bounded subset of $\mathcal{L}(V, W)$ is sequentially compact for any choice of norm on this space.

27. Let $L \in \mathcal{L}(V, V)$ where V is a finite dimensional normed vector space. Define

$$e^L \equiv \sum_{k=1}^{\infty} \frac{L^k}{k!}$$

Explain the meaning of this infinite sum and show it converges in $\mathcal{L}(V, V)$ for any choice of norm on this space. Now tell how to define $\sin(L)$.

28. Let X be a finite dimensional normed vector space, real or complex. Show that X is separable. **Hint:** Let $\{v_i\}_{i=1}^n$ be a basis and define a map from $\mathbb{F}^n$ to X, θ, as follows. $\theta\left(\sum_{k=1}^n x_k \mathbf{e}_k\right) \equiv \sum_{k=1}^n x_k v_k$. Show θ is continuous and has a continuous inverse. Now let D be a countable dense set in $\mathbb{F}^n$ and consider $\theta(D)$.

29. Let $B(X;\mathbb{R}^n)$ be the space of functions $\mathbf{f}$, mapping X to $\mathbb{R}^n$ such that

$$\sup\{|\mathbf{f}(\mathbf{x})| : \mathbf{x} \in X\} < \infty.$$

Show $B(X;\mathbb{R}^n)$ is a complete normed linear space if we define

$$||\mathbf{f}|| \equiv \sup\{|\mathbf{f}(\mathbf{x})| : \mathbf{x} \in X\}.$$

30. Let $\alpha \in (0,1]$. Define, for X a compact subset of $\mathbb{R}^p$,

$$C^\alpha(X;\mathbb{R}^n) \equiv \{\mathbf{f} \in C(X;\mathbb{R}^n) : \rho_\alpha(\mathbf{f}) + ||\mathbf{f}|| \equiv ||\mathbf{f}||_\alpha < \infty\}$$

where

$$||\mathbf{f}|| \equiv \sup\{|\mathbf{f}(\mathbf{x})| : \mathbf{x} \in X\}$$

and

$$\rho_\alpha(\mathbf{f}) \equiv \sup\left\{\frac{|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})|}{|\mathbf{x}-\mathbf{y}|^\alpha} : \mathbf{x},\mathbf{y} \in X,\ \mathbf{x} \neq \mathbf{y}\right\}.$$

Show that $(C^\alpha(X;\mathbb{R}^n), ||\cdot||_\alpha)$ is a complete normed linear space. This is called a Holder space. What would this space consist of if $\alpha > 1$?

31. Let $\{\mathbf{f}_n\}_{n=1}^\infty \subseteq C^\alpha(X;\mathbb{R}^n)$ where X is a compact subset of $\mathbb{R}^p$ and suppose

$$||\mathbf{f}_n||_\alpha \leq M$$

for all n. Show there exists a subsequence, n_k, such that $\mathbf{f}_{n_k}$ converges in $C(X;\mathbb{R}^n)$. The given sequence is precompact when this happens. (This also shows the embedding of $C^\alpha(X;\mathbb{R}^n)$ into $C(X;\mathbb{R}^n)$ is a compact embedding.) **Hint:** You might want to use the Ascoli Arzela theorem.

32. This problem is for those who know about the derivative and the integral of a function of one variable. Let $\mathbf{f} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ be continuous and bounded and let $\mathbf{x}_0 \in \mathbb{R}^n$. If

$$\mathbf{x} : [0,T] \to \mathbb{R}^n$$

and $h > 0$, let

$$\tau_h \mathbf{x}(s) \equiv \begin{cases} \mathbf{x}_0 \text{ if } s \leq h, \\ \mathbf{x}(s-h), \text{ if } s > h. \end{cases}$$

For $t \in [0,T]$, let

$$\mathbf{x}_h(t) = \mathbf{x}_0 + \int_0^t \mathbf{f}(s, \tau_h \mathbf{x}_h(s))\, ds.$$

Show using the Ascoli Arzela theorem that there exists a sequence $h \to 0$ such that

$$\mathbf{x}_h \to \mathbf{x}$$

in $C([0,T];\mathbb{R}^n)$. Next argue

$$\mathbf{x}(t) = \mathbf{x}_0 + \int_0^t \mathbf{f}(s, \mathbf{x}(s))\, ds$$

and conclude the following theorem. If $\mathbf{f}:\mathbb{R}\times\mathbb{R}^n\to\mathbb{R}^n$ is continuous and bounded, and if $\mathbf{x}_0\in\mathbb{R}^n$ is given, there exists a solution to the following initial value problem.

$$\begin{aligned}\mathbf{x}' &= \mathbf{f}(t,\mathbf{x}),\quad t\in[0,T]\\ \mathbf{x}(0) &= \mathbf{x}_0.\end{aligned}$$

This is the Peano existence theorem for ordinary differential equations.

33. Let $D(\mathbf{x}_0,r)$ be the closed ball in $\mathbb{R}^n$,

$$\{\mathbf{x}:|\mathbf{x}-\mathbf{x}_0|\le r\}$$

where this is the usual norm coming from the dot product. Let $P:\mathbb{R}^n\to D(\mathbf{x}_0,r)$ be defined by

$$P(\mathbf{x})\equiv\begin{cases}\mathbf{x} \text{ if } \mathbf{x}\in D(\mathbf{x}_0,r)\\ \mathbf{x}_0+r\frac{\mathbf{x}-\mathbf{x}_0}{|\mathbf{x}-\mathbf{x}_0|} \text{ if } \mathbf{x}\notin D(\mathbf{x}_0,r)\end{cases}$$

Show that $|P\mathbf{x}-P\mathbf{y}|\le|\mathbf{x}-\mathbf{y}|$ for all $\mathbf{x}\in\mathbb{R}^n$.

34. Use Problem 32 to obtain local solutions to the initial value problem where $\mathbf{f}$ is not assumed to be bounded. It is only assumed to be continuous. This means there is a small interval whose length is perhaps not T such that the solution to the differential equation exists on this small interval.

Chapter 6

The Derivative

6.1 Limits Of A Function

As in the case of scalar valued functions of one variable, a concept closely related to continuity is that of the **limit of a function**. The notion of limit of a function makes sense at points $\mathbf{x}$, which are limit points of $D(\mathbf{f})$ and this concept is defined next. In all that follows $(V, \|\cdot\|)$ and $(W, \|\cdot\|)$ are two normed linear spaces. Recall the definition of limit point first.

Definition 6.1.1 *Let $A \subseteq W$ be a set. A point $\mathbf{x}$, is a limit point of A if $B(\mathbf{x}, r)$ contains infinitely many points of A for every $r > 0$.*

Definition 6.1.2 *Let $\mathbf{f} : D(\mathbf{f}) \subseteq V \to W$ be a function and let $\mathbf{x}$ be a **limit point** of $D(\mathbf{f})$. Then*

$$\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}$$

if and only if the following condition holds. For all $\varepsilon > 0$ there exists $\delta > 0$ such that if

$$0 < \|\mathbf{y} - \mathbf{x}\| < \delta, \text{ and } \mathbf{y} \in D(\mathbf{f})$$

then,

$$\|\mathbf{L} - \mathbf{f}(\mathbf{y})\| < \varepsilon.$$

Theorem 6.1.3 *If $\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}$ and $\lim_{y\to x} \mathbf{f}(\mathbf{y}) = \mathbf{L}_1$, then $\mathbf{L} = \mathbf{L}_1$.*

Proof: Let $\varepsilon > 0$ be given. There exists $\delta > 0$ such that if $0 < |\mathbf{y} - \mathbf{x}| < \delta$ and $\mathbf{y} \in D(\mathbf{f})$, then

$$\|\mathbf{f}(\mathbf{y}) - \mathbf{L}\| < \varepsilon, \ \|\mathbf{f}(\mathbf{y}) - \mathbf{L}_1\| < \varepsilon.$$

Pick such a $\mathbf{y}$. There exists one because $\mathbf{x}$ is a limit point of $D(\mathbf{f})$. Then

$$\|\mathbf{L} - \mathbf{L}_1\| \leq \|\mathbf{L} - \mathbf{f}(\mathbf{y})\| + \|\mathbf{f}(\mathbf{y}) - \mathbf{L}_1\| < \varepsilon + \varepsilon = 2\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this shows $\mathbf{L} = \mathbf{L}_1$. ■

As in the case of functions of one variable, one can define what it means for $\lim_{\mathbf{y}\to\mathbf{x}} f(\mathbf{x}) = \pm\infty$.

Definition 6.1.4 *If $f(\mathbf{x}) \in \mathbb{R}$, $\lim_{\mathbf{y}\to\mathbf{x}} f(\mathbf{x}) = \infty$ if for every number l, there exists $\delta > 0$ such that whenever $\|\mathbf{y} - \mathbf{x}\| < \delta$ and $\mathbf{y} \in D(\mathbf{f})$, then $f(\mathbf{x}) > l$. $\lim_{\mathbf{y}\to\mathbf{x}} f(\mathbf{x}) = -\infty$ if for every number l, there exists $\delta > 0$ such that whenever $\|\mathbf{y} - \mathbf{x}\| < \delta$ and $\mathbf{y} \in D(\mathbf{f})$, then $f(\mathbf{x}) < l$.*

The following theorem is just like the one variable version of calculus.

Theorem 6.1.5 *Suppose* $\mathbf{f} : D(\mathbf{f}) \subseteq V \to \mathbb{F}^m$. *Then for* $\mathbf{x}$ *a limit point of* $D(\mathbf{f})$,

$$\lim_{y \to x} \mathbf{f}(\mathbf{y}) = \mathbf{L} \tag{6.1}$$

if and only if

$$\lim_{y \to x} f_k(\mathbf{y}) = L_k \tag{6.2}$$

where $\mathbf{f}(\mathbf{y}) \equiv (f_1(\mathbf{y}), \cdots, f_p(\mathbf{y}))$ *and* $\mathbf{L} \equiv (L_1, \cdots, L_p)$.

Suppose here that f *has values in* W, *a normed linear space and*

$$\lim_{y \to \mathbf{x}} f(y) = L, \ \lim_{y \to \mathbf{x}} g(y) = K$$

where K,$L \in W$. *Then if* a, $b \in \mathbb{F}$,

$$\lim_{y \to x} (af(y) + bg(y)) = aL + bK, \tag{6.3}$$

If W *is an inner product space,*

$$\lim_{y \to x} (f, g)(y) = (L,K) \tag{6.4}$$

If g *is scalar valued with* $\lim_{y \to x} g(y) = K$,

$$\lim_{y \to x} f(y) g(y) = LK. \tag{6.5}$$

Also, if h *is a continuous function defined near* L, *then*

$$\lim_{y \to x} h \circ f(y) = h(L). \tag{6.6}$$

Suppose $\lim_{y \to \mathbf{x}} f(y) = L$. *If* $\|f(y) - b\| \leq r$ *for all* y *sufficiently close to* $\mathbf{x}$, *then* $|L - b| \leq r$ *also.*

Proof: Suppose 6.1. Then letting $\varepsilon > 0$ be given there exists $\delta > 0$ such that if $0 < \|y - x\| < \delta$, it follows

$$|f_k(y) - L_k| \leq \|\mathbf{f}(y) - \mathbf{L}\| < \varepsilon$$

which verifies 6.2.

Now suppose 6.2 holds. Then letting $\varepsilon > 0$ be given, there exists δ_k such that if $0 < \|y - x\| < \delta_k$, then

$$|f_k(y) - L_k| < \varepsilon.$$

Let $0 < \delta < \min(\delta_1, \cdots, \delta_p)$. Then if $0 < \|y - x\| < \delta$, it follows

$$\|\mathbf{f}(y) - \mathbf{L}\|_\infty < \varepsilon$$

Any other norm on $\mathbb{F}^m$ would work out the same way because the norms are all equivalent.

Each of the remaining assertions follows immediately from the coordinate descriptions of the various expressions and the first part. However, I will give a different argument for these.

The proof of 6.3 is left for you. Now 6.4 is to be verified. Let $\varepsilon > 0$ be given. Then by the triangle inequality,

$$\begin{aligned} |(f,g)(y) - (L,K)| &\leq |(f,g)(y) - (f(y), K)| + |(f(y), K) - (L,K)| \\ &\leq \|f(y)\| \|g(y) - K\| + \|K\| \|f(y) - L\|. \end{aligned}$$

There exists δ_1 such that if $0 < \|y-x\| < \delta_1$ and $y \in D(f)$, then

$$\|f(y) - L\| < 1,$$

and so for such y, the triangle inequality implies, $\|f(y)\| < 1 + \|L\|$. Therefore, for $0 < \|y-x\| < \delta_1$,

$$|(f,g)(y) - (L,K)| \leq (1 + \|K\| + \|L\|)\left[\|g(y) - K\| + \|f(y) - L\|\right]. \tag{6.7}$$

Now let $0 < \delta_2$ be such that if $y \in D(f)$ and $0 < \|x-y\| < \delta_2$,

$$\|f(y) - L\| < \frac{\varepsilon}{2(1 + \|K\| + \|L\|)},\ \|g(y) - K\| < \frac{\varepsilon}{2(1 + \|K\| + \|L\|)}.$$

Then letting $0 < \delta \leq \min(\delta_1, \delta_2)$, it follows from 6.7 that

$$|(f,g)(y) - (L,K)| < \varepsilon$$

and this proves 6.4.

The proof of 6.5 is left to you.

Consider 6.6. Since h is continuous near L, it follows that for $\varepsilon > 0$ given, there exists $\eta > 0$ such that if $\|y-L\| < \eta$, then

$$\|h(y) - h(L)\| < \varepsilon$$

Now since $\lim_{y\to x} f(y) = L$, there exists $\delta > 0$ such that if $0 < \|y-x\| < \delta$, then

$$\|f(y) - L\| < \eta.$$

Therefore, if $0 < \|y-x\| < \delta$,

$$\|h(f(y)) - h(L)\| < \varepsilon.$$

It only remains to verify the last assertion. Assume $\|f(y) - b\| \leq r$. It is required to show that $\|L-b\| \leq r$. If this is not true, then $\|L-b\| > r$. Consider $B(L, \|L-b\| - r)$. Since L is the limit of f, it follows $f(y) \in B(L, \|L-b\| - r)$ whenever $y \in D(f)$ is close enough to x. Thus, by the triangle inequality,

$$\|f(y) - L\| < \|L-b\| - r$$

and so

$$\begin{aligned} r &< \|L-b\| - \|f(y) - L\| \leq |\|b-L\| - \|f(y) - L\|| \\ &\leq \|b-f(y)\|, \end{aligned}$$

a contradiction to the assumption that $\|b-f(y)\| \leq r$. ■

The relation between continuity and limits is as follows.

Theorem 6.1.6 *For $f : D(f) \to W$ and $x \in D(f)$ a limit point of $D(f)$, f is continuous at x if and only if*

$$\lim_{y\to x} f(y) = f(x).$$

Proof: First suppose f is continuous at x a limit point of $D(f)$. Then for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\|x - y\| < \delta$ and $y \in D(f)$, then $|f(x) - f(y)| < \varepsilon$. In particular, this holds if $0 < \|x - y\| < \delta$ and this is just the definition of the limit. Hence $f(x) = \lim_{y\to x} f(y)$.

Next suppose x is a limit point of $D(f)$ and $\lim_{y\to x} f(y) = f(x)$. This means that if $\varepsilon > 0$ there exists $\delta > 0$ such that for $0 < \|x - y\| < \delta$ and $y \in D(f)$, it follows $|f(y) - f(x)| < \varepsilon$. However, if $y = x$, then $|f(y) - f(x)| = |f(x) - f(x)| = 0$ and so whenever $y \in D(f)$ and $\|x - y\| < \delta$, it follows $|f(x) - f(y)| < \varepsilon$, showing f is continuous at x. ■

Example 6.1.7 *Find* $\lim_{(x,y)\to(3,1)} \left(\frac{x^2-9}{x-3}, y\right)$.

It is clear that $\lim_{(x,y)\to(3,1)} \frac{x^2-9}{x-3} = 6$ and $\lim_{(x,y)\to(3,1)} y = 1$. Therefore, this limit equals $(6,1)$.

Example 6.1.8 *Find* $\lim_{(x,y)\to(0,0)} \frac{xy}{x^2+y^2}$.

First of all, observe the domain of the function is $\mathbb{R}^2 \setminus \{(0,0)\}$, every point in $\mathbb{R}^2$ except the origin. Therefore, $(0,0)$ is a limit point of the domain of the function so it might make sense to take a limit. However, just as in the case of a function of one variable, the limit may not exist. In fact, this is the case here. To see this, take points on the line $y = 0$. At these points, the value of the function equals 0. Now consider points on the line $y = x$ where the value of the function equals $1/2$. Since, arbitrarily close to $(0,0)$, there are points where the function equals $1/2$ and points where the function has the value 0, it follows there can be no limit. Just take $\varepsilon = 1/10$ for example. You cannot be within $1/10$ of $1/2$ and also within $1/10$ of 0 at the same time.

Note it is necessary to rely on the definition of the limit much more than in the case of a function of one variable and there are no easy ways to do limit problems for functions of more than one variable. It is what it is and you will not deal with these concepts without suffering and anguish.

6.2 Basic Definitions

The concept of derivative generalizes right away to functions of many variables. However, no attempt will be made to consider derivatives from one side or another. This is because when you consider functions of many variables, there isn't a well defined side. However, it is certainly the case that there are more general notions which include such things. I will present a fairly general notion of the derivative of a function which is defined on a normed vector space which has values in a normed vector space. The case of most interest is that of a function which maps $\mathbb{F}^n$ to $\mathbb{F}^m$ but it is no more trouble to consider the extra generality and it is sometimes useful to have this extra generality because sometimes you want to consider functions defined, for example on subspaces of $\mathbb{F}^n$and it is nice to not have to trouble with ad hoc considerations. Also, you might want to consider $\mathbb{F}^n$ with some norm other than the usual one.

In what follows, X, Y will denote normed vector spaces. Thanks to Theorem 5.8.4 all the definitions and theorems given below work the same for any norm given on the vector spaces.

Let U be an open set in X, and let $\mathbf{f} : U \to Y$ be a function.

Definition 6.2.1 *A function* $\mathbf{g}$ *is* $\mathbf{o}(\mathbf{v})$ *if*

$$\lim_{||\mathbf{v}||\to 0} \frac{\mathbf{g}(\mathbf{v})}{||\mathbf{v}||} = \mathbf{0} \tag{6.8}$$

A function $\mathbf{f} : U \to Y$ *is differentiable at* $\mathbf{x} \in U$ *if there exists a linear transformation* $L \in \mathcal{L}(X, Y)$ *such that*

$$\mathbf{f}(\mathbf{x}+\mathbf{v}) = \mathbf{f}(\mathbf{x}) + L\mathbf{v} + \mathbf{o}(\mathbf{v})$$

This linear transformation L *is the definition of* $D\mathbf{f}(\mathbf{x})$. *This derivative is often called the Frechet derivative.*

Note that from Theorem 5.8.4 the question whether a given function is differentiable is independent of the norm used on the finite dimensional vector space. That is, a

function is differentiable with one norm if and only if it is differentiable with another norm.

The definition 6.8 means the error,

$$\mathbf{f}(\mathbf{x}+\mathbf{v})-\mathbf{f}(\mathbf{x})-L\mathbf{v}$$

converges to $\mathbf{0}$ faster than $||\mathbf{v}||$. Thus the above definition is equivalent to saying

$$\lim_{||\mathbf{v}||\to 0}\frac{||\mathbf{f}(\mathbf{x}+\mathbf{v})-\mathbf{f}(\mathbf{x})-L\mathbf{v}||}{||\mathbf{v}||}=0 \tag{6.9}$$

or equivalently,

$$\lim_{\mathbf{y}\to\mathbf{x}}\frac{||\mathbf{f}(\mathbf{y})-\mathbf{f}(\mathbf{x})-D\mathbf{f}(\mathbf{x})(\mathbf{y}-\mathbf{x})||}{||\mathbf{y}-\mathbf{x}||}=0. \tag{6.10}$$

The symbol, $\mathbf{o}(\mathbf{v})$ should be thought of as an adjective. Thus, if t and k are constants,

$$\mathbf{o}(\mathbf{v})=\mathbf{o}(\mathbf{v})+\mathbf{o}(\mathbf{v}),\ \mathbf{o}(t\mathbf{v})=\mathbf{o}(\mathbf{v}),\ k\mathbf{o}(\mathbf{v})=\mathbf{o}(\mathbf{v})$$

and other similar observations hold.

Theorem 6.2.2 *The derivative is well defined.*

Proof: First note that for a fixed vector, $\mathbf{v}$, $\mathbf{o}(t\mathbf{v})=\mathbf{o}(t)$. This is because

$$\lim_{t\to 0}\frac{\mathbf{o}(t\mathbf{v})}{|t|}=\lim_{t\to 0}||\mathbf{v}||\frac{\mathbf{o}(t\mathbf{v})}{||t\mathbf{v}||}=\mathbf{0}$$

Now suppose both L_1 and L_2 work in the above definition. Then let $\mathbf{v}$ be any vector and let t be a real scalar which is chosen small enough that $t\mathbf{v}+\mathbf{x}\in U$. Then

$$\mathbf{f}(\mathbf{x}+t\mathbf{v})=\mathbf{f}(\mathbf{x})+L_1t\mathbf{v}+\mathbf{o}(t\mathbf{v}),\ \mathbf{f}(\mathbf{x}+t\mathbf{v})=\mathbf{f}(\mathbf{x})+L_2t\mathbf{v}+\mathbf{o}(t\mathbf{v}).$$

Therefore, subtracting these two yields $(L_2-L_1)(t\mathbf{v})=\mathbf{o}(t\mathbf{v})=\mathbf{o}(t)$. Therefore, dividing by t yields $(L_2-L_1)(\mathbf{v})=\frac{\mathbf{o}(t)}{t}$. Now let $t\to 0$ to conclude that $(L_2-L_1)(\mathbf{v})=0$. Since this is true for all $\mathbf{v}$, it follows $L_2=L_1$. This proves the theorem. ■

Lemma 6.2.3 *Let $\mathbf{f}$ be differentiable at $\mathbf{x}$. Then $\mathbf{f}$ is continuous at $\mathbf{x}$ and in fact, there exists $K>0$ such that whenever $||\mathbf{v}||$ is small enough,*

$$||\mathbf{f}(\mathbf{x}+\mathbf{v})-\mathbf{f}(\mathbf{x})||\le K||\mathbf{v}||$$

Also if $\mathbf{f}$ is differentiable at $\mathbf{x}$, then

$$\mathbf{o}(\|\mathbf{f}(\mathbf{x}+\mathbf{v})-\mathbf{f}(\mathbf{x})\|)=\mathbf{o}(\mathbf{v})$$

Proof: From the definition of the derivative,

$$\mathbf{f}(\mathbf{x}+\mathbf{v})-\mathbf{f}(\mathbf{x})=D\mathbf{f}(\mathbf{x})\mathbf{v}+\mathbf{o}(\mathbf{v}).$$

Let $||\mathbf{v}||$ be small enough that $\frac{\mathbf{o}(||\mathbf{v}||)}{||\mathbf{v}||}<1$ so that $||\mathbf{o}(\mathbf{v})||\le||\mathbf{v}||$. Then for such $\mathbf{v}$,

$$\begin{aligned}||\mathbf{f}(\mathbf{x}+\mathbf{v})-\mathbf{f}(\mathbf{x})|| &\le ||D\mathbf{f}(\mathbf{x})\mathbf{v}||+||\mathbf{v}||\\ &\le (||D\mathbf{f}(\mathbf{x})||+1)||\mathbf{v}||\end{aligned}$$

This proves the lemma with $K=||D\mathbf{f}(\mathbf{x})||+1$.

The last assertion is implied by the first as follows. Define

$$\mathbf{h}(\mathbf{v}) \equiv \begin{cases} \frac{\mathbf{o}(\|\mathbf{f}(\mathbf{x}+\mathbf{v})-\mathbf{f}(\mathbf{x})\|)}{\|\mathbf{f}(\mathbf{x}+\mathbf{v})-\mathbf{f}(\mathbf{x})\|} \text{ if } \|\mathbf{f}(\mathbf{x}+\mathbf{v})-\mathbf{f}(\mathbf{x})\| \neq 0 \\ \mathbf{0} \text{ if } \|\mathbf{f}(\mathbf{x}+\mathbf{v})-\mathbf{f}(\mathbf{x})\| = 0 \end{cases}$$

Then $\lim_{\|\mathbf{v}\|\to 0} \mathbf{h}(\mathbf{v}) = \mathbf{0}$ from continuity of $\mathbf{f}$ at $\mathbf{x}$ which is implied by the first part. Also from the above estimate,

$$\left\| \frac{\mathbf{o}(\|\mathbf{f}(\mathbf{x}+\mathbf{v})-\mathbf{f}(\mathbf{x})\|)}{\|\mathbf{v}\|} \right\| = \|\mathbf{h}(\mathbf{v})\| \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{v})-\mathbf{f}(\mathbf{x})\|}{\|\mathbf{v}\|} \leq \|\mathbf{h}(\mathbf{v})\| (||D\mathbf{f}(\mathbf{x})|| + 1)$$

This establishes the second claim. ■

Here $||D\mathbf{f}(\mathbf{x})||$ is the operator norm of the linear transformation, $D\mathbf{f}(\mathbf{x})$.

6.3 The Chain Rule

With the above lemma, it is easy to prove the chain rule.

Theorem 6.3.1 *(The chain rule) Let U and V be open sets $U \subseteq X$ and $V \subseteq Y$. Suppose $\mathbf{f} : U \to V$ is differentiable at $\mathbf{x} \in U$ and suppose $\mathbf{g} : V \to \mathbb{F}^q$ is differentiable at $\mathbf{f}(\mathbf{x}) \in V$. Then $\mathbf{g} \circ \mathbf{f}$ is differentiable at $\mathbf{x}$ and*

$$D(\mathbf{g} \circ \mathbf{f})(\mathbf{x}) = D(\mathbf{g}(\mathbf{f}(\mathbf{x}))) D(\mathbf{f}(\mathbf{x})).$$

Proof: This follows from a computation. Let $B(\mathbf{x},r) \subseteq U$ and let r also be small enough that for $||\mathbf{v}|| \leq r$, it follows that $\mathbf{f}(\mathbf{x}+\mathbf{v}) \in V$. Such an r exists because $\mathbf{f}$ is continuous at $\mathbf{x}$. For $||\mathbf{v}|| < r$, the definition of differentiability of $\mathbf{g}$ and $\mathbf{f}$ implies

$$\mathbf{g}(\mathbf{f}(\mathbf{x}+\mathbf{v})) - \mathbf{g}(\mathbf{f}(\mathbf{x})) =$$

$$\begin{aligned} & D\mathbf{g}(\mathbf{f}(\mathbf{x}))(\mathbf{f}(\mathbf{x}+\mathbf{v}) - \mathbf{f}(\mathbf{x})) + \mathbf{o}(\mathbf{f}(\mathbf{x}+\mathbf{v}) - \mathbf{f}(\mathbf{x})) \\ = \; & D\mathbf{g}(\mathbf{f}(\mathbf{x}))[D\mathbf{f}(\mathbf{x})\mathbf{v} + \mathbf{o}(\mathbf{v})] + \mathbf{o}(\mathbf{f}(\mathbf{x}+\mathbf{v}) - \mathbf{f}(\mathbf{x})) \\ = \; & D(\mathbf{g}(\mathbf{f}(\mathbf{x}))) D(\mathbf{f}(\mathbf{x}))\mathbf{v} + \mathbf{o}(\mathbf{v}) + \mathbf{o}(\mathbf{f}(\mathbf{x}+\mathbf{v}) - \mathbf{f}(\mathbf{x})) \\ = \; & D(\mathbf{g}(\mathbf{f}(\mathbf{x}))) D(\mathbf{f}(\mathbf{x}))\mathbf{v} + \mathbf{o}(\mathbf{v}) \end{aligned} \tag{6.11}$$

By Lemma 6.2.3. From the definition of the derivative $D(\mathbf{g} \circ \mathbf{f})(\mathbf{x})$ exists and equals $D(\mathbf{g}(\mathbf{f}(\mathbf{x})))D(\mathbf{f}(\mathbf{x}))$. ■

6.4 The Matrix Of The Derivative

Let X, Y be normed vector spaces, a basis for X being $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and a basis for Y being $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$. First note that if $\pi_i : X \to \mathbb{F}$ is defined by

$$\pi_i \mathbf{v} \equiv x_i \text{ where } \mathbf{v} = \sum_k x_k \mathbf{v}_k,$$

then $\pi_i \in \mathcal{L}(X, \mathbb{F})$ and so by Theorem 5.8.3, it follows that π_i is continuous and if $\lim_{s\to t} \mathbf{g}(s) = \mathbf{L}$, then $|\pi_i \mathbf{g}(s) - \pi_i \mathbf{L}| \leq ||\pi_i||\,||\mathbf{g}(s) - \mathbf{L}||$ and so the i^{th} components converge also.

Suppose that $\mathbf{f} : U \to Y$ is differentiable. What is the matrix of $D\mathbf{f}(\mathbf{x})$ with respect to the given bases? That is, if

$$D\mathbf{f}(\mathbf{x}) = \sum_{ij} J_{ij}(\mathbf{x}) \mathbf{w}_i \mathbf{v}_j,$$

what is $J_{ij}(\mathbf{x})$?

$$\begin{aligned} D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x}) &\equiv \lim_{t\to 0}\frac{\mathbf{f}(\mathbf{x}+t\mathbf{v}_k)-\mathbf{f}(\mathbf{x})}{t} = \lim_{t\to 0}\frac{D\mathbf{f}(\mathbf{x})(t\mathbf{v}_k)+\mathbf{o}(t\mathbf{v}_k)}{t} \\ &= D\mathbf{f}(\mathbf{x})(\mathbf{v}_k) = \sum_{ij} J_{ij}(\mathbf{x})\,\mathbf{w}_i\mathbf{v}_j(\mathbf{v}_k) = \sum_{ij} J_{ij}(\mathbf{x})\,\mathbf{w}_i\delta_{jk} \\ &= \sum_i J_{ik}(\mathbf{x})\,\mathbf{w}_i \end{aligned}$$

It follows

$$\begin{aligned} &\lim_{t\to 0}\pi_j\left(\frac{\mathbf{f}(\mathbf{x}+t\mathbf{v}_k)-\mathbf{f}(\mathbf{x})}{t}\right) \\ \equiv\ &\lim_{t\to 0}\frac{f_j(\mathbf{x}+t\mathbf{v}_k)-f_j(\mathbf{x})}{t} \equiv D_{\mathbf{v}_k}f_j(\mathbf{x}) \\ =\ &\pi_j\left(\sum_i J_{ik}(\mathbf{x})\,\mathbf{w}_i\right) = J_{jk}(\mathbf{x}) \end{aligned}$$

Thus $J_{ik}(\mathbf{x}) = D_{\mathbf{v}_k}f_i(\mathbf{x})$.

In the case where $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$ and $\mathbf{v}$ is a unit vector, $D_{\mathbf{v}}f_i(\mathbf{x})$ is the familiar directional derivative in the direction $\mathbf{v}$ of the function, f_i.

Of course the case where $X = \mathbb{F}^n$ and $\mathbf{f} : U \subseteq \mathbb{F}^n \to \mathbb{F}^m$, is differentiable and the basis vectors are the usual basis vectors is the case most commonly encountered. What is the matrix of $D\mathbf{f}(\mathbf{x})$ taken with respect to the usual basis vectors? Let $\mathbf{e}_i$ denote the vector of $\mathbb{F}^n$ which has a one in the i^{th} entry and zeroes elsewhere. This is the standard basis for $\mathbb{F}^n$. Denote by $J_{ij}(\mathbf{x})$ the matrix with respect to these basis vectors. Thus

$$D\mathbf{f}(\mathbf{x}) = \sum_{ij} J_{ij}(\mathbf{x})\,\mathbf{e}_i\mathbf{e}_j.$$

Then from what was just shown,

$$\begin{aligned} J_{ik}(\mathbf{x}) &= D_{\mathbf{e}_k}f_i(\mathbf{x}) \equiv \lim_{t\to 0}\frac{f_i(\mathbf{x}+t\mathbf{e}_k)-f_i(\mathbf{x})}{t} \\ &\equiv \frac{\partial f_i}{\partial x_k}(\mathbf{x}) \equiv f_{i,x_k}(\mathbf{x}) \equiv f_{i,k}(\mathbf{x}) \end{aligned}$$

where the last several symbols are just the usual notations for the partial derivative of the function, f_i with respect to the k^{th} variable where

$$\mathbf{f}(\mathbf{x}) \equiv \sum_{i=1}^m f_i(\mathbf{x})\,\mathbf{e}_i.$$

In other words, the matrix of $D\mathbf{f}(\mathbf{x})$ is nothing more than the matrix of partial derivatives. The k^{th} column of the matrix (J_{ij}) is

$$\frac{\partial \mathbf{f}}{\partial x_k}(\mathbf{x}) = \lim_{t\to 0}\frac{\mathbf{f}(\mathbf{x}+t\mathbf{e}_k)-\mathbf{f}(\mathbf{x})}{t} \equiv D_{\mathbf{e}_k}\mathbf{f}(\mathbf{x}).$$

Thus the matrix of $D\mathbf{f}(\mathbf{x})$ with respect to the usual basis vectors is the matrix of the form

$$\begin{pmatrix} f_{1,x_1}(\mathbf{x}) & f_{1,x_2}(\mathbf{x}) & \cdots & f_{1,x_n}(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ f_{m,x_1}(\mathbf{x}) & f_{m,x_2}(\mathbf{x}) & \cdots & f_{m,x_n}(\mathbf{x}) \end{pmatrix}$$

where the notation $g_{,x_k}$ denotes the k^{th} partial derivative given by the limit,

$$\lim_{t\to 0}\frac{g(\mathbf{x}+t\mathbf{e}_k)-g(\mathbf{x})}{t}\equiv\frac{\partial g}{\partial x_k}.$$

The above discussion is summarized in the following theorem.

Theorem 6.4.1 *Let $\mathbf{f}:\mathbb{F}^n\to\mathbb{F}^m$ and suppose $\mathbf{f}$ is differentiable at $\mathbf{x}$. Then all the partial derivatives $\frac{\partial f_i(\mathbf{x})}{\partial x_j}$ exist and if $J\mathbf{f}(\mathbf{x})$ is the matrix of the linear transformation, $D\mathbf{f}(\mathbf{x})$ with respect to the standard basis vectors, then the ij^{th} entry is given by $\frac{\partial f_i}{\partial x_j}(\mathbf{x})$ also denoted as $f_{i,j}$ or f_{i,x_j}.*

Definition 6.4.2 *In general, the symbol*

$$D_{\mathbf{v}}\mathbf{f}(\mathbf{x})$$

is defined by

$$\lim_{t\to 0}\frac{\mathbf{f}(\mathbf{x}+t\mathbf{v})-\mathbf{f}(\mathbf{x})}{t}$$

where $t\in\mathbb{F}$. This is often called the Gateaux derivative.

What if all the partial derivatives of $\mathbf{f}$ exist? Does it follow that $\mathbf{f}$ is differentiable? Consider the following function, $f:\mathbb{R}^2\to\mathbb{R}$,

$$f(x,y)=\begin{cases}\frac{xy}{x^2+y^2}\text{ if }(x,y)\neq(0,0)\\ 0\text{ if }(x,y)=(0,0)\end{cases}.$$

Then from the definition of partial derivatives,

$$\lim_{h\to 0}\frac{f(h,0)-f(0,0)}{h}=\lim_{h\to 0}\frac{0-0}{h}=0$$

and

$$\lim_{h\to 0}\frac{f(0,h)-f(0,0)}{h}=\lim_{h\to 0}\frac{0-0}{h}=0$$

However f is not even continuous at $(0,0)$ which may be seen by considering the behavior of the function along the line $y=x$ and along the line $x=0$. By Lemma 6.2.3 this implies f is not differentiable. Therefore, it is necessary to consider the correct definition of the derivative given above if you want to get a notion which generalizes the concept of the derivative of a function of one variable in such a way as to preserve continuity whenever the function is differentiable.

6.5 A Mean Value Inequality

The following theorem will be very useful in much of what follows. It is a version of the mean value theorem as is the next lemma.

Lemma 6.5.1 *Let Y be a normed vector space and suppose $\mathbf{h}:[0,1]\to Y$ is differentiable and satisfies*

$$||\mathbf{h}'(t)||\leq M.$$

Then

$$||\mathbf{h}(1)-\mathbf{h}(0)||\leq M.$$

Proof: Let $\varepsilon > 0$ be given and let

$$S \equiv \{t \in [0,1] : \text{ for all } s \in [0,t], ||\mathbf{h}(s) - \mathbf{h}(0)|| \leq (M+\varepsilon)\, s\}$$

Then $0 \in S$. Let $t = \sup S$. Then by continuity of $\mathbf{h}$ it follows

$$||\mathbf{h}(t) - \mathbf{h}(0)|| = (M+\varepsilon)\, t \tag{6.12}$$

Suppose $t < 1$. Then there exist positive numbers, h_k decreasing to 0 such that

$$||\mathbf{h}(t+h_k) - \mathbf{h}(0)|| > (M+\varepsilon)(t+h_k)$$

and now it follows from 6.12 and the triangle inequality that

$$\begin{array}{rl} & ||\mathbf{h}(t+h_k) - \mathbf{h}(t)|| + ||\mathbf{h}(t) - \mathbf{h}(0)|| \\ = & ||\mathbf{h}(t+h_k) - \mathbf{h}(t)|| + (M+\varepsilon)\, t > (M+\varepsilon)(t+h_k) \end{array}$$

and so

$$||\mathbf{h}(t+h_k) - \mathbf{h}(t)|| > (M+\varepsilon)\, h_k$$

Now dividing by h_k and letting $k \to \infty$

$$||\mathbf{h}'(t)|| \geq M + \varepsilon,$$

a contradiction. This proves the lemma. ■

Theorem 6.5.2 *Suppose U is an open subset of X and $\mathbf{f} : U \to Y$ has the property that $D\mathbf{f}(\mathbf{x})$ exists for all $\mathbf{x}$ in U and that, $\mathbf{x} + t(\mathbf{y} - \mathbf{x}) \in U$ for all $t \in [0,1]$. (The line segment joining the two points lies in U.) Suppose also that for all points on this line segment,*

$$||D\mathbf{f}(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))|| \leq M.$$

Then

$$||\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})|| \leq M\, |\mathbf{y} - \mathbf{x}|.$$

Proof: Let

$$\mathbf{h}(t) \equiv \mathbf{f}(\mathbf{x} + t(\mathbf{y} - \mathbf{x})).$$

Then by the chain rule,

$$\mathbf{h}'(t) = D\mathbf{f}(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})$$

and so

$$\begin{array}{rcl} ||\mathbf{h}'(t)|| & = & ||D\mathbf{f}(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})|| \\ & \leq & M\, ||\mathbf{y} - \mathbf{x}|| \end{array}$$

by Lemma 6.5.1

$$||\mathbf{h}(1) - \mathbf{h}(0)|| = ||\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})|| \leq M\, ||\mathbf{y} - \mathbf{x}||.$$

This proves the theorem. ■

6.6 Existence Of The Derivative, C^1 Functions

There is a way to get the differentiability of a function from the existence and continuity of the Gateaux derivatives. This is very convenient because these Gateaux derivatives are taken with respect to a one dimensional variable. The following theorem is the main result.

Theorem 6.6.1 *Let X be a normed vector space having basis $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and let Y be another normed vector space having basis $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$. Let U be an open set in X and let $\mathbf{f} : U \to Y$ have the property that the Gateaux derivatives,*

$$D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x}) \equiv \lim_{t\to 0} \frac{\mathbf{f}(\mathbf{x}+t\mathbf{v}_k) - \mathbf{f}(\mathbf{x})}{t}$$

exist and are continuous functions of $\mathbf{x}$. Then $D\mathbf{f}(\mathbf{x})$ exists and

$$D\mathbf{f}(\mathbf{x})\,\mathbf{v} = \sum_{k=1}^{n} D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x})\, a_k$$

where

$$\mathbf{v} = \sum_{k=1}^{n} a_k \mathbf{v}_k.$$

Furthermore, $\mathbf{x} \to D\mathbf{f}(\mathbf{x})$ is continuous; that is

$$\lim_{\mathbf{y}\to\mathbf{x}} ||D\mathbf{f}(\mathbf{y}) - D\mathbf{f}(\mathbf{x})|| = 0.$$

Proof: Let $\mathbf{v} = \sum_{k=1}^{n} a_k\mathbf{v}_k$. Then

$$\mathbf{f}(\mathbf{x}+\mathbf{v}) - \mathbf{f}(\mathbf{x}) = \mathbf{f}\left(\mathbf{x} + \sum_{k=1}^{n} a_k\mathbf{v}_k\right) - \mathbf{f}(\mathbf{x}).$$

Then letting $\sum_{k=1}^{0} \equiv 0$, $\mathbf{f}(\mathbf{x}+\mathbf{v}) - \mathbf{f}(\mathbf{x})$ is given by

$$\sum_{k=1}^{n}\left[\mathbf{f}\left(\mathbf{x} + \sum_{j=1}^{k} a_j\mathbf{v}_j\right) - \mathbf{f}\left(\mathbf{x} + \sum_{j=1}^{k-1} a_j\mathbf{v}_j\right)\right]$$

$$= \sum_{k=1}^{n}\left[\mathbf{f}(\mathbf{x}+a_k\mathbf{v}_k) - \mathbf{f}(\mathbf{x})\right] +$$

$$\sum_{k=1}^{n}\left[\left(\mathbf{f}\left(\mathbf{x} + \sum_{j=1}^{k} a_j\mathbf{v}_j\right) - \mathbf{f}(\mathbf{x}+a_k\mathbf{v}_k)\right) - \left(\mathbf{f}\left(\mathbf{x} + \sum_{j=1}^{k-1} a_j\mathbf{v}_j\right) - \mathbf{f}(\mathbf{x})\right)\right] \tag{6.13}$$

Consider the k^{th} term in 6.13. Let

$$\mathbf{h}(t) \equiv \mathbf{f}\left(\mathbf{x} + \sum_{j=1}^{k-1} a_j\mathbf{v}_j + ta_k\mathbf{v}_k\right) - \mathbf{f}(\mathbf{x} + ta_k\mathbf{v}_k)$$

for $t \in [0,1]$. Then

$$\begin{aligned}\mathbf{h}'(t) &= a_k \lim_{h\to 0} \frac{1}{a_k h}\left(\mathbf{f}\left(\mathbf{x} + \sum_{j=1}^{k-1} a_j\mathbf{v}_j + (t+h)a_k\mathbf{v}_k\right) - \mathbf{f}(\mathbf{x} + (t+h)a_k\mathbf{v}_k)\right.\\ &\quad \left. - \left(\mathbf{f}\left(\mathbf{x} + \sum_{j=1}^{k-1} a_j\mathbf{v}_j + ta_k\mathbf{v}_k\right) - \mathbf{f}(\mathbf{x} + ta_k\mathbf{v}_k)\right)\right)\end{aligned}$$

and this equals

$$\left(D_{\mathbf{v}_k}\mathbf{f}\left(\mathbf{x}+\sum_{j=1}^{k-1} a_j\mathbf{v}_j + ta_k\mathbf{v}_k \right) - D_{\mathbf{v}_k}\mathbf{f}\left(\mathbf{x}+ta_k\mathbf{v}_k\right) \right) a_k \tag{6.14}$$

Now without loss of generality, it can be assumed the norm on X is given by that of Example 5.8.5,

$$||\mathbf{v}|| \equiv \max\left\{ |a_k| : \mathbf{v} = \sum_{j=1}^{n} a_k\mathbf{v}_k \right\}$$

because by Theorem 5.8.4 all norms on X are equivalent. Therefore, from 6.14 and the assumption that the Gateaux derivatives are continuous,

$$\begin{array}{rcl} ||\mathbf{h}'(t)|| & = & \left|\left| \left(D_{\mathbf{v}_k}\mathbf{f}\left(\mathbf{x}+\sum_{j=1}^{k-1} a_j\mathbf{v}_j + ta_k\mathbf{v}_k \right) - D_{\mathbf{v}_k}\mathbf{f}\left(\mathbf{x}+ta_k\mathbf{v}_k\right) \right) a_k \right|\right| \\ & \leq & \varepsilon\,|a_k| \leq \varepsilon\,||\mathbf{v}|| \end{array}$$

provided $||\mathbf{v}||$ is sufficiently small. Since ε is arbitrary, it follows from Lemma 6.5.1 the expression in 6.13 is $\mathbf{o}(\mathbf{v})$ because this expression equals a finite sum of terms of the form $\mathbf{h}(1)-\mathbf{h}(0)$ where $||\mathbf{h}'(t)|| \leq \varepsilon\,||\mathbf{v}||$. Thus

$$\mathbf{f}(\mathbf{x}+\mathbf{v}) - \mathbf{f}(\mathbf{x}) = \sum_{k=1}^{n} \left[\mathbf{f}(\mathbf{x}+a_k\mathbf{v}_k) - \mathbf{f}(\mathbf{x})\right] + \mathbf{o}(\mathbf{v})$$

$$= \sum_{k=1}^{n} D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x})\,a_k + \sum_{k=1}^{n} \left[\mathbf{f}(\mathbf{x}+a_k\mathbf{v}_k) - \mathbf{f}(\mathbf{x}) - D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x})\,a_k\right] + \mathbf{o}(\mathbf{v}).$$

Consider the k^{th} term in the second sum.

$$\mathbf{f}(\mathbf{x}+a_k\mathbf{v}_k) - \mathbf{f}(\mathbf{x}) - D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x})\,a_k = a_k\left(\frac{\mathbf{f}(\mathbf{x}+a_k\mathbf{v}_k)-\mathbf{f}(\mathbf{x})}{a_k} - D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x}) \right)$$

where the expression in the parentheses converges to 0 as $a_k \to 0$. Thus whenever $||\mathbf{v}||$ is sufficiently small,

$$||\mathbf{f}(\mathbf{x}+a_k\mathbf{v}_k) - \mathbf{f}(\mathbf{x}) - D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x})\,a_k|| \leq \varepsilon\,|a_k| \leq \varepsilon\,||\mathbf{v}||$$

which shows the second sum is also $\mathbf{o}(\mathbf{v})$. Therefore,

$$\mathbf{f}(\mathbf{x}+\mathbf{v}) - \mathbf{f}(\mathbf{x}) = \sum_{k=1}^{n} D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x})\,a_k + \mathbf{o}(\mathbf{v}).$$

Defining

$$D\mathbf{f}(\mathbf{x})\,\mathbf{v} \equiv \sum_{k=1}^{n} D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x})\,a_k$$

where $\mathbf{v} = \sum_k a_k\mathbf{v}_k$, it follows $D\mathbf{f}(\mathbf{x}) \in \mathcal{L}(X,Y)$ and is given by the above formula.

It remains to verify $\mathbf{x} \to D\mathbf{f}(\mathbf{x})$ is continuous.

$$\begin{aligned} & ||(D\mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{y}))\mathbf{v}|| \\ \leq & \sum_{k=1}^{n} ||(D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x}) - D_{\mathbf{v}_k}\mathbf{f}(\mathbf{y}))a_k|| \\ \leq & \max\{|a_k|, k = 1, \cdots, n\} \sum_{k=1}^{n} ||D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x}) - D_{\mathbf{v}_k}\mathbf{f}(\mathbf{y})|| \\ = & ||\mathbf{v}|| \sum_{k=1}^{n} ||D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x}) - D_{\mathbf{v}_k}\mathbf{f}(\mathbf{y})|| \end{aligned}$$

and so

$$||D\mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{y})|| \leq \sum_{k=1}^{n} ||D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x}) - D_{\mathbf{v}_k}\mathbf{f}(\mathbf{y})||$$

which proves the continuity of $D\mathbf{f}$ because of the assumption the Gateaux derivatives are continuous. This proves the theorem. ■

This motivates the following definition of what it means for a function to be C^1.

Definition 6.6.2 *Let U be an open subset of a normed finite dimensional vector space, X and let $\mathbf{f} : U \to Y$ another finite dimensional normed vector space. Then $\mathbf{f}$ is said to be C^1 if there exists a basis for $X, \{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ such that the Gateaux derivatives,*

$$D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x})$$

exist on U and are continuous.

Here is another definition of what it means for a function to be C^1.

Definition 6.6.3 *Let U be an open subset of a normed finite dimensional vector space, X and let $\mathbf{f} : U \to Y$ another finite dimensional normed vector space. Then $\mathbf{f}$ is said to be C^1 if $\mathbf{f}$ is differentiable and $\mathbf{x} \to D\mathbf{f}(\mathbf{x})$ is continuous as a map from U to $\mathcal{L}(X, Y)$.*

Now the following major theorem states these two definitions are equivalent.

Theorem 6.6.4 *Let U be an open subset of a normed finite dimensional vector space, X and let $\mathbf{f} : U \to Y$ another finite dimensional normed vector space. Then the two definitions above are equivalent.*

Proof: It was shown in Theorem 6.6.1 that Definition 6.6.2 implies 6.6.3. Suppose then that Definition 6.6.3 holds. Then if $\mathbf{v}$ is any vector,

$$\begin{aligned} \lim_{t\to 0} \frac{\mathbf{f}(\mathbf{x} + t\mathbf{v}) - \mathbf{f}(\mathbf{x})}{t} &= \lim_{t\to 0} \frac{D\mathbf{f}(\mathbf{x})\, t\mathbf{v} + \mathbf{o}(t\mathbf{v})}{t} \\ &= D\mathbf{f}(\mathbf{x})\mathbf{v} + \lim_{t\to 0} \frac{\mathbf{o}(t\mathbf{v})}{t} = D\mathbf{f}(\mathbf{x})\mathbf{v} \end{aligned}$$

Thus $D_{\mathbf{v}}\mathbf{f}(\mathbf{x})$ exists and equals $D\mathbf{f}(\mathbf{x})\mathbf{v}$. By continuity of $\mathbf{x} \to D\mathbf{f}(\mathbf{x})$, this establishes continuity of $\mathbf{x} \to D_{\mathbf{v}}\mathbf{f}(\mathbf{x})$ and proves the theorem. ■

Note that the proof of the theorem also implies the following corollary.

Corollary 6.6.5 *Let U be an open subset of a normed finite dimensional vector space, X and let $\mathbf{f} : U \to Y$ another finite dimensional normed vector space. Then if there is a basis of $X, \{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ such that the Gateaux derivatives, $D_{\mathbf{v}_k}\mathbf{f}(\mathbf{x})$ exist and are continuous. Then all Gateaux derivatives, $D_{\mathbf{v}}\mathbf{f}(\mathbf{x})$ exist and are continuous for all $\mathbf{v} \in X$.*

From now on, whichever definition is more convenient will be used.

6.7 Higher Order Derivatives

If $f : U \subseteq X \to Y$ for U an open set, then

$$\mathbf{x} \to D\mathbf{f}(\mathbf{x})$$

is a mapping from U to $\mathcal{L}(X,Y)$, a normed vector space. Therefore, it makes perfect sense to ask whether this function is also differentiable.

Definition 6.7.1 *The following is the definition of the second derivative.*

$$D^2\mathbf{f}(\mathbf{x}) \equiv D(D\mathbf{f}(\mathbf{x})).$$

Thus,

$$D\mathbf{f}(\mathbf{x}+\mathbf{v}) - D\mathbf{f}(\mathbf{x}) = D^2\mathbf{f}(\mathbf{x})\mathbf{v} + \mathbf{o}(\mathbf{v}).$$

This implies

$$D^2\mathbf{f}(\mathbf{x}) \in \mathcal{L}(X, \mathcal{L}(X,Y)),\ D^2\mathbf{f}(\mathbf{x})(\mathbf{u})(\mathbf{v}) \in Y,$$

and the map

$$(\mathbf{u},\mathbf{v}) \to D^2\mathbf{f}(\mathbf{x})(\mathbf{u})(\mathbf{v})$$

is a bilinear map having values in Y. In other words, the two functions,

$$\mathbf{u} \to D^2\mathbf{f}(\mathbf{x})(\mathbf{u})(\mathbf{v}),\ \mathbf{v} \to D^2\mathbf{f}(\mathbf{x})(\mathbf{u})(\mathbf{v})$$

are both linear.

The same pattern applies to taking higher order derivatives. Thus,

$$D^3\mathbf{f}(\mathbf{x}) \equiv D\left(D^2\mathbf{f}(\mathbf{x})\right)$$

and $D^3\mathbf{f}(\mathbf{x})$ may be considered as a trilinear map having values in Y. In general $D^k\mathbf{f}(\mathbf{x})$ may be considered a k linear map. This means the function

$$(\mathbf{u}_1, \cdots, \mathbf{u}_k) \to D^k\mathbf{f}(\mathbf{x})(\mathbf{u}_1)\cdots(\mathbf{u}_k)$$

has the property

$$\mathbf{u}_j \to D^k\mathbf{f}(\mathbf{x})(\mathbf{u}_1)\cdots(\mathbf{u}_j)\cdots(\mathbf{u}_k)$$

is linear.

Also, instead of writing

$$D^2\mathbf{f}(\mathbf{x})(\mathbf{u})(\mathbf{v}),\ \text{or } D^3\mathbf{f}(\mathbf{x})(\mathbf{u})(\mathbf{v})(\mathbf{w})$$

the following notation is often used.

$$D^2\mathbf{f}(\mathbf{x})(\mathbf{u},\mathbf{v})\ \text{ or } D^3\mathbf{f}(\mathbf{x})(\mathbf{u},\mathbf{v},\mathbf{w})$$

with similar conventions for higher derivatives than 3. Another convention which is often used is the notation

$$D^k\mathbf{f}(\mathbf{x})\mathbf{v}^k$$

instead of

$$D^k\mathbf{f}(\mathbf{x})(\mathbf{v}, \cdots, \mathbf{v}).$$

Note that for every k, $D^k\mathbf{f}$ maps U to a normed vector space. As mentioned above, $D\mathbf{f}(\mathbf{x})$ has values in $\mathcal{L}(X,Y)$, $D^2\mathbf{f}(\mathbf{x})$ has values in $\mathcal{L}(X, \mathcal{L}(X,Y))$, etc. Thus it makes sense to consider whether $D^k\mathbf{f}$ is continuous. This is described in the following definition.

Definition 6.7.2 *Let U be an open subset of X, a normed vector space and let $\mathbf{f} : U \to Y$. Then $\mathbf{f}$ is $C^k(U)$ if $\mathbf{f}$ and its first k derivatives are all continuous. Also, $D^k\mathbf{f}(\mathbf{x})$ when it exists can be considered a Y valued multilinear function.*

6.8 C^k Functions

Recall that for a C^1 function, $\mathbf{f}$

$$\begin{aligned} D\mathbf{f}(\mathbf{x})\mathbf{v} &= \sum_j D_{\mathbf{v}_j}\mathbf{f}(\mathbf{x})\,a_j = \sum_{ij} D_{\mathbf{v}_j} f_i(\mathbf{x})\,\mathbf{w}_i a_j \\ &= \sum_{ij} D_{\mathbf{v}_j} f_i(\mathbf{x})\,\mathbf{w}_i\mathbf{v}_j\left(\sum_k a_k\mathbf{v}_k\right) = \sum_{ij} D_{\mathbf{v}_j} f_i(\mathbf{x})\,\mathbf{w}_i\mathbf{v}_j(\mathbf{v}) \end{aligned}$$

where $\sum_k a_k\mathbf{v}_k = \mathbf{v}$ and

$$\mathbf{f}(\mathbf{x}) = \sum_i f_i(\mathbf{x})\,\mathbf{w}_i. \tag{6.15}$$

This is because

$$\mathbf{w}_i\mathbf{v}_j\left(\sum_k a_k\mathbf{v}_k\right) \equiv \sum_k a_k\mathbf{w}_i\delta_{jk} = \mathbf{w}_i a_j.$$

Thus

$$D\mathbf{f}(\mathbf{x}) = \sum_{ij} D_{\mathbf{v}_j} f_i(\mathbf{x})\,\mathbf{w}_i\mathbf{v}_j$$

I propose to iterate this observation, starting with $\mathbf{f}$ and then going to $D\mathbf{f}$ and then $D^2\mathbf{f}$ and so forth. Hopefully it will yield a rational way to understand higher order derivatives in the same way that matrices can be used to understand linear transformations. Thus beginning with the derivative,

$$D\mathbf{f}(\mathbf{x}) = \sum_{ij_1} D_{\mathbf{v}_{j_1}} f_i(\mathbf{x})\,\mathbf{w}_i\mathbf{v}_{j_1}.$$

Then letting $\mathbf{w}_i\mathbf{v}_{j_1}$ play the role of $\mathbf{w}_i$ in 6.15,

$$\begin{aligned} D^2\mathbf{f}(\mathbf{x}) &= \sum_{ij_1j_2} D_{\mathbf{v}_{j_2}}\left(D_{\mathbf{v}_{j_1}} f_i\right)(\mathbf{x})\,\mathbf{w}_i\mathbf{v}_{j_1}\mathbf{v}_{j_2} \\ &\equiv \sum_{ij_1j_2} D_{\mathbf{v}_{j_1}\mathbf{v}_{j_2}} f_i(\mathbf{x})\,\mathbf{w}_i\mathbf{v}_{j_1}\mathbf{v}_{j_2} \end{aligned}$$

Then letting $\mathbf{w}_i\mathbf{v}_{j_1}\mathbf{v}_{j_2}$ play the role of $\mathbf{w}_i$ in 6.15,

$$\begin{aligned} D^3\mathbf{f}(\mathbf{x}) &= \sum_{ij_1j_2j_3} D_{\mathbf{v}_{j_3}}\left(D_{\mathbf{v}_{j_1}\mathbf{v}_{j_2}} f_i\right)(\mathbf{x})\,\mathbf{w}_i\mathbf{v}_{j_1}\mathbf{v}_{j_2}\mathbf{v}_{j_3} \\ &\equiv \sum_{ij_1j_2j_3} D_{\mathbf{v}_{j_1}\mathbf{v}_{j_2}\mathbf{v}_{j_3}} f_i(\mathbf{x})\,\mathbf{w}_i\mathbf{v}_{j_1}\mathbf{v}_{j_2}\mathbf{v}_{j_3} \end{aligned}$$

etc. In general, the notation,

$$\mathbf{w}_i\mathbf{v}_{j_1}\mathbf{v}_{j_2}\cdots\mathbf{v}_{j_n}$$

defines an appropriate linear transformation given by

$$\mathbf{w}_i\mathbf{v}_{j_1}\mathbf{v}_{j_2}\cdots\mathbf{v}_{j_n}(\mathbf{v}_k) = \mathbf{w}_i\mathbf{v}_{j_1}\mathbf{v}_{j_2}\cdots\mathbf{v}_{j_{n-1}}\delta_{kj_n}.$$

The following theorem is important.

Theorem 6.8.1 *The function $\mathbf{x}\to D^k\mathbf{f}(\mathbf{x})$ exists and is continuous for $k\le p$ if and only if there exists a basis for $X,\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ and a basis for $Y,\{\mathbf{w}_1,\cdots,\mathbf{w}_m\}$ such that for*

$$\mathbf{f}(\mathbf{x}) \equiv \sum_i f_i(\mathbf{x})\,\mathbf{w}_i,$$

it follows that for each $i = 1, 2, \cdots, m$ all Gateaux derivatives,

$$D_{\mathbf{v}_{j_1}\mathbf{v}_{j_2}\cdots\mathbf{v}_{j_k}} f_i(\mathbf{x})$$

for any choice of $\mathbf{v}_{j_1}\mathbf{v}_{j_2}\cdots\mathbf{v}_{j_k}$ and for any $k \leq p$ exist and are continuous.

Proof: This follows from a repeated application of Theorems 6.6.1 and 6.6.4 at each new differentiation. ■

Definition 6.8.2 *Let X, Y be finite dimensional normed vector spaces and let U be an open set in X and $\mathbf{f} : U \to Y$ be a function,*

$$\mathbf{f}(\mathbf{x}) = \sum_i f_i(\mathbf{x})\,\mathbf{w}_i$$

where $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ is a basis for Y. Then $\mathbf{f}$ is said to be a $C^n(U)$ function if for every $k \leq n$, $D^k\mathbf{f}(\mathbf{x})$ exists for all $\mathbf{x} \in U$ and is continuous. This is equivalent to the other condition which states that for each $i = 1, 2, \cdots, m$ all Gateaux derivatives,

$$D_{\mathbf{v}_{j_1}\mathbf{v}_{j_2}\cdots\mathbf{v}_{j_k}} f_i(\mathbf{x})$$

for any choice of $\mathbf{v}_{j_1}\mathbf{v}_{j_2}\cdots\mathbf{v}_{j_k}$ where $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis for X and for any $k \leq n$ exist and are continuous.

6.8.1 Some Standard Notation

In the case where $X = \mathbb{R}^n$ and the basis chosen is the standard basis, these Gateaux derivatives are just the partial derivatives. Recall the notation for partial derivatives in the following definition.

Definition 6.8.3 *Let $\mathbf{g} : U \to X$. Then*

$$\mathbf{g}_{x_k}(\mathbf{x}) \equiv \frac{\partial \mathbf{g}}{\partial x_k}(\mathbf{x}) \equiv \lim_{h\to 0} \frac{\mathbf{g}(\mathbf{x} + h\mathbf{e}_k) - \mathbf{g}(\mathbf{x})}{h}$$

Higher order partial derivatives are defined in the usual way.

$$\mathbf{g}_{x_k x_l}(\mathbf{x}) \equiv \frac{\partial^2 \mathbf{g}}{\partial x_l \partial x_k}(\mathbf{x})$$

and so forth.

A convenient notation which is often used which helps to make sense of higher order partial derivatives is presented in the following definition.

Definition 6.8.4 *$\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$ for $\alpha_1 \cdots \alpha_n$ positive integers is called a multi-index. For $\boldsymbol{\alpha}$ a multi-index, $|\boldsymbol{\alpha}| \equiv \alpha_1 + \cdots + \alpha_n$ and if $\mathbf{x} \in X$,*

$$\mathbf{x} = (x_1, \cdots, x_n),$$

and $\mathbf{f}$ a function, define

$$\mathbf{x}^{\boldsymbol{\alpha}} \equiv x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n},\ D^{\boldsymbol{\alpha}}\mathbf{f}(\mathbf{x}) \equiv \frac{\partial^{|\boldsymbol{\alpha}|}\mathbf{f}(\mathbf{x})}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_n^{\alpha_n}}.$$

Then in this special case, the following definition is equivalent to the above as a definition of what is meant by a C^k function.

Definition 6.8.5 *Let U be an open subset of $\mathbb{R}^n$ and let $\mathbf{f} : U \to Y$. Then for k a nonnegative integer, $\mathbf{f}$ is C^k if for every $|\boldsymbol{\alpha}| \leq k$, $D^{\boldsymbol{\alpha}}\mathbf{f}$ exists and is continuous.*

6.9 The Derivative And The Cartesian Product

There are theorems which can be used to get differentiability of a function based on existence and continuity of the partial derivatives. A generalization of this was given above. Here a function defined on a product space is considered. It is very much like what was presented above and could be obtained as a special case but to reinforce the ideas, I will do it from scratch because certain aspects of it are important in the statement of the implicit function theorem.

The following is an important abstract generalization of the concept of partial derivative presented above. Insead of taking the derivative with respect to one variable, it is taken with respect to several but not with respect to others. This vague notion is made precise in the following definition. First here is a lemma.

Lemma 6.9.1 *Suppose U is an open set in $X \times Y$. Then the set, $U_{\mathbf{y}}$ defined by*

$$U_{\mathbf{y}} \equiv \{\mathbf{x} \in X : (\mathbf{x}, \mathbf{y}) \in U\}$$

is an open set in X. Here $X \times Y$ is a finite dimensional vector space in which the vector space operations are defined componentwise. Thus for $a, b \in \mathbb{F}$,

$$a(\mathbf{x}_1, \mathbf{y}_1) + b(\mathbf{x}_2, \mathbf{y}_2) = (a\mathbf{x}_1 + b\mathbf{x}_2, a\mathbf{y}_1 + b\mathbf{y}_2)$$

and the norm can be taken to be

$$||(\mathbf{x}, \mathbf{y})|| \equiv \max(||\mathbf{x}||, ||\mathbf{y}||)$$

Proof: Recall by Theorem 5.8.4 it does not matter how this norm is defined and the definition above is convenient. It obviously satisfies most axioms of a norm. The only one which is not obvious is the triangle inequality. I will show this now.

$$\begin{aligned} ||(\mathbf{x}, \mathbf{y}) + (\mathbf{x}_1, \mathbf{y}_1)|| &= ||(\mathbf{x} + \mathbf{x}_1, \mathbf{y} + \mathbf{y}_1)|| \equiv \max(||\mathbf{x} + \mathbf{x}_1||, ||\mathbf{y} + \mathbf{y}_1||) \\ &\leq \max(||\mathbf{x}|| + ||\mathbf{x}_1||, ||\mathbf{y}|| + ||\mathbf{y}_1||) \end{aligned}$$

suppose then that $||\mathbf{x}|| + ||\mathbf{x}_1|| \geq ||\mathbf{y}|| + ||\mathbf{y}_1||$. Then the above equals

$$||\mathbf{x}|| + ||\mathbf{x}_1|| \leq \max(||\mathbf{x}||, ||\mathbf{y}||) + \max(||\mathbf{x}_1||, ||\mathbf{y}_1||) \equiv ||(\mathbf{x}, \mathbf{y})|| + ||(\mathbf{x}_1, \mathbf{y}_1)||$$

In case $||\mathbf{x}|| + ||\mathbf{x}_1|| < ||\mathbf{y}|| + ||\mathbf{y}_1||$, the argument is similar.

Let $\mathbf{x} \in U_{\mathbf{y}}$. Then $(\mathbf{x}, \mathbf{y}) \in U$ and so there exists $r > 0$ such that

$$B((\mathbf{x}, \mathbf{y}), r) \in U.$$

This says that if $(\mathbf{u}, \mathbf{v}) \in X \times Y$ such that $||(\mathbf{u}, \mathbf{v}) - (\mathbf{x}, \mathbf{y})|| < r$, then $(\mathbf{u}, \mathbf{v}) \in U$. Thus if

$$||(\mathbf{u}, \mathbf{y}) - (\mathbf{x}, \mathbf{y})|| = ||\mathbf{u} - \mathbf{x}|| < r,$$

then $(\mathbf{u}, \mathbf{y}) \in U$. This has just said that $B(\mathbf{x}, r)$, the ball taken in X is contained in $U_{\mathbf{y}}$. This proves the lemma. ■

Or course one could also consider

$$U_{\mathbf{x}} \equiv \{\mathbf{y} : (\mathbf{x}, \mathbf{y}) \in U\}$$

in the same way and conclude this set is open in Y. Also, the generalization to many factors yields the same conclusion. In this case, for $\mathbf{x} \in \prod_{i=1}^n X_i$, let

$$||\mathbf{x}|| \equiv \max\left(||\mathbf{x}_i||_{X_i} : \mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)\right)$$

Then a similar argument to the above shows this is a norm on $\prod_{i=1}^n X_i$.

Corollary 6.9.2 *Let $U \subseteq \prod_{i=1}^n X_i$ and let*

$$U_{(\mathbf{x}_1,\cdots,\mathbf{x}_{i-1},\mathbf{x}_{i+1},\cdots,\mathbf{x}_n)} \equiv \left\{\mathbf{x} \in \mathbb{F}^{r_i} : \left(\mathbf{x}_1,\cdots,\mathbf{x}_{i-1},\mathbf{x},\mathbf{x}_{i+1},\cdots,\mathbf{x}_n\right) \in U\right\}.$$

Then $U_{(\mathbf{x}_1,\cdots,\mathbf{x}_{i-1},\mathbf{x}_{i+1},\cdots,\mathbf{x}_n)}$ is an open set in $\mathbb{F}^{r_i}$.

The proof is similar to the above.

Definition 6.9.3 *Let $\mathbf{g} : U \subseteq \prod_{i=1}^n X_i \to Y$, where U is an open set. Then the map*

$$\mathbf{z} \to \mathbf{g}\left(\mathbf{x}_1,\cdots,\mathbf{x}_{i-1},\mathbf{z},\mathbf{x}_{i+1},\cdots,\mathbf{x}_n\right)$$

is a function from the open set in X_i,

$$\left\{\mathbf{z} : \mathbf{x} = \left(\mathbf{x}_1,\cdots,\mathbf{x}_{i-1},\mathbf{z},\mathbf{x}_{i+1},\cdots,\mathbf{x}_n\right) \in U\right\}$$

to Y. When this map is differentiable, its derivative is denoted by $D_i\mathbf{g}(\mathbf{x})$. To aid in the notation, for $\mathbf{v} \in X_i$, let $\theta_i\mathbf{v} \in \prod_{i=1}^n X_i$ be the vector $(\mathbf{0},\cdots,\mathbf{v},\cdots,\mathbf{0})$ where the $\mathbf{v}$ is in the i^{th} slot and for $\mathbf{v} \in \prod_{i=1}^n X_i$, let $\mathbf{v}_i$ denote the entry in the i^{th} slot of $\mathbf{v}$. Thus, by saying

$$\mathbf{z} \to \mathbf{g}\left(\mathbf{x}_1,\cdots,\mathbf{x}_{i-1},\mathbf{z},\mathbf{x}_{i+1},\cdots,\mathbf{x}_n\right)$$

is differentiable is meant that for $\mathbf{v} \in X_i$ sufficiently small,

$$\mathbf{g}(\mathbf{x}+\theta_i\mathbf{v}) - \mathbf{g}(\mathbf{x}) = D_i\mathbf{g}(\mathbf{x})\,\mathbf{v} + \mathbf{o}(\mathbf{v}).$$

Note $D_i\mathbf{g}(\mathbf{x}) \in \mathcal{L}(X_i, Y)$.

Definition 6.9.4 *Let $U \subseteq X$ be an open set. Then $\mathbf{f} : U \to Y$ is $C^1(U)$ if $\mathbf{f}$ is differentiable and the mapping*

$$\mathbf{x} \to D\mathbf{f}(\mathbf{x}),$$

is continuous as a function from U to $\mathcal{L}(X, Y)$.

With this definition of partial derivatives, here is the major theorem.

Theorem 6.9.5 *Let $\mathbf{g}, U, \prod_{i=1}^n X_i$, be given as in Definition 6.9.3. Then $\mathbf{g}$ is $C^1(U)$ if and only if $D_i\mathbf{g}$ exists and is continuous on U for each i. In this case, $\mathbf{g}$ is differentiable and*

$$D\mathbf{g}(\mathbf{x})(\mathbf{v}) = \sum_k D_k\mathbf{g}(\mathbf{x})\,\mathbf{v}_k \tag{6.16}$$

where $\mathbf{v} = (\mathbf{v}_1,\cdots,\mathbf{v}_n)$.

Proof: Suppose then that $D_i\mathbf{g}$ exists and is continuous for each i. Note that

$$\sum_{j=1}^k \theta_j\mathbf{v}_j = (\mathbf{v}_1,\cdots,\mathbf{v}_k,\mathbf{0},\cdots,\mathbf{0}).$$

Thus $\sum_{j=1}^n \theta_j\mathbf{v}_j = \mathbf{v}$ and define $\sum_{j=1}^0 \theta_j\mathbf{v}_j \equiv \mathbf{0}$. Therefore,

$$\mathbf{g}(\mathbf{x}+\mathbf{v}) - \mathbf{g}(\mathbf{x}) = \sum_{k=1}^n \left[\mathbf{g}\left(\mathbf{x} + \sum_{j=1}^k \theta_j\mathbf{v}_j\right) - \mathbf{g}\left(\mathbf{x} + \sum_{j=1}^{k-1} \theta_j\mathbf{v}_j\right)\right] \tag{6.17}$$

Consider the terms in this sum.

$$\mathbf{g}\left(\mathbf{x} + \sum_{j=1}^k \theta_j\mathbf{v}_j\right) - \mathbf{g}\left(\mathbf{x} + \sum_{j=1}^{k-1} \theta_j\mathbf{v}_j\right) = \mathbf{g}(\mathbf{x}+\theta_k\mathbf{v}_k) - \mathbf{g}(\mathbf{x}) + \tag{6.18}$$

$$\left(\mathbf{g}\left(\mathbf{x}+\sum_{j=1}^{k}\theta_j\mathbf{v}_j\right)-\mathbf{g}\left(\mathbf{x}+\theta_k\mathbf{v}_k\right)\right)-\left(\mathbf{g}\left(\mathbf{x}+\sum_{j=1}^{k-1}\theta_j\mathbf{v}_j\right)-\mathbf{g}\left(\mathbf{x}\right)\right) \tag{6.19}$$

and the expression in 6.19 is of the form $\mathbf{h}\left(\mathbf{v}_k\right)-\mathbf{h}\left(\mathbf{0}\right)$ where for small $\mathbf{w}\in X_k$,

$$\mathbf{h}\left(\mathbf{w}\right)\equiv\mathbf{g}\left(\mathbf{x}+\sum_{j=1}^{k-1}\theta_j\mathbf{v}_j+\theta_k\mathbf{w}\right)-\mathbf{g}\left(\mathbf{x}+\theta_k\mathbf{w}\right).$$

Therefore,

$$D\mathbf{h}\left(\mathbf{w}\right)=D_k\mathbf{g}\left(\mathbf{x}+\sum_{j=1}^{k-1}\theta_j\mathbf{v}_j+\theta_k\mathbf{w}\right)-D_k\mathbf{g}\left(\mathbf{x}+\theta_k\mathbf{w}\right)$$

and by continuity, $||D\mathbf{h}\left(\mathbf{w}\right)||<\varepsilon$ provided $||\mathbf{v}||$ is small enough. Therefore, by Theorem 6.5.2, whenever $||\mathbf{v}||$ is small enough,

$$||\mathbf{h}\left(\mathbf{v}_k\right)-\mathbf{h}\left(\mathbf{0}\right)||\leq\varepsilon\,||\mathbf{v}_k||\leq\varepsilon\,||\mathbf{v}||$$

which shows that since ε is arbitrary, the expression in 6.19 is $\mathbf{o}\left(\mathbf{v}\right)$. Now in 6.18

$$\mathbf{g}\left(\mathbf{x}+\theta_k\mathbf{v}_k\right)-\mathbf{g}\left(\mathbf{x}\right)=D_k\mathbf{g}\left(\mathbf{x}\right)\mathbf{v}_k+\mathbf{o}\left(\mathbf{v}_k\right)=D_k\mathbf{g}\left(\mathbf{x}\right)\mathbf{v}_k+\mathbf{o}\left(\mathbf{v}\right).$$

Therefore, referring to 6.17,

$$\mathbf{g}\left(\mathbf{x}+\mathbf{v}\right)-\mathbf{g}\left(\mathbf{x}\right)=\sum_{k=1}^{n}D_k\mathbf{g}\left(\mathbf{x}\right)\mathbf{v}_k+\mathbf{o}\left(\mathbf{v}\right)$$

which shows $D\mathbf{g}\left(\mathbf{x}\right)$ exists and equals the formula given in 6.16.

Next suppose $\mathbf{g}$ is C^1. I need to verify that $D_k\mathbf{g}\left(\mathbf{x}\right)$ exists and is continuous. Let $\mathbf{v}\in X_k$ sufficiently small. Then

$$\begin{aligned}\mathbf{g}\left(\mathbf{x}+\theta_k\mathbf{v}\right)-\mathbf{g}\left(\mathbf{x}\right) &= D\mathbf{g}\left(\mathbf{x}\right)\theta_k\mathbf{v}+\mathbf{o}\left(\theta_k\mathbf{v}\right)\\ &= D\mathbf{g}\left(\mathbf{x}\right)\theta_k\mathbf{v}+\mathbf{o}\left(\mathbf{v}\right)\end{aligned}$$

since $||\theta_k\mathbf{v}||=||\mathbf{v}||$. Then $D_k\mathbf{g}\left(\mathbf{x}\right)$ exists and equals

$$D\mathbf{g}\left(\mathbf{x}\right)\circ\theta_k$$

Now $\mathbf{x}\to D\mathbf{g}\left(\mathbf{x}\right)$ is continuous. Since θ_k is linear, it follows from Theorem 5.8.3 that $\theta_k:X_k\to\prod_{i=1}^{n}X_i$ is also continuous, This proves the theorem. ■

The way this is usually used is in the following corollary, a case of Theorem 6.9.5 obtained by letting $X_i=\mathbb{F}$ in the above theorem.

Corollary 6.9.6 *Let U be an open subset of $\mathbb{F}^n$ and let $\mathbf{f}:U\to\mathbb{F}^m$ be C^1 in the sense that all the partial derivatives of $\mathbf{f}$ exist and are continuous. Then $\mathbf{f}$ is differentiable and*

$$\mathbf{f}\left(\mathbf{x}+\mathbf{v}\right)=\mathbf{f}\left(\mathbf{x}\right)+\sum_{k=1}^{n}\frac{\partial\mathbf{f}}{\partial x_k}\left(\mathbf{x}\right)\mathbf{v}_k+\mathbf{o}\left(\mathbf{v}\right).$$

6.10 Mixed Partial Derivatives

Continuing with the special case where f is defined on an open set in $\mathbb{F}^n$, I will next consider an interesting result due to Euler in 1734 about mixed partial derivatives. It turns out that the mixed partial derivatives, if continuous will end up being equal. Recall the notation

$$f_x = \frac{\partial f}{\partial x} = D_{\mathbf{e}_1} f$$

and

$$f_{xy} = \frac{\partial^2 f}{\partial y \partial x} = D_{\mathbf{e}_1 \mathbf{e}_2} f.$$

Theorem 6.10.1 *Suppose $f : U \subseteq \mathbb{F}^2 \to \mathbb{R}$ where U is an open set on which f_x, f_y, f_{xy} and f_{yx} exist. Then if f_{xy} and f_{yx} are continuous at the point $(x, y) \in U$, it follows*

$$f_{xy}(x, y) = f_{yx}(x, y).$$

Proof: Since U is open, there exists $r > 0$ such that $B((x, y), r) \subseteq U$. Now let $|t|, |s| < r/2, t, s$ real numbers and consider

$$\Delta(s, t) \equiv \frac{1}{st}\{\overbrace{f(x + t, y + s) - f(x + t, y)}^{h(t)} - \overbrace{(f(x, y + s) - f(x, y))}^{h(0)}\}. \tag{6.20}$$

Note that $(x + t, y + s) \in U$ because

$$\begin{aligned} |(x + t, y + s) - (x, y)| &= |(t, s)| = \left(t^2 + s^2\right)^{1/2} \\ &\leq \left(\frac{r^2}{4} + \frac{r^2}{4}\right)^{1/2} = \frac{r}{\sqrt{2}} < r. \end{aligned}$$

As implied above, $h(t) \equiv f(x + t, y + s) - f(x + t, y)$. Therefore, by the mean value theorem and the (one variable) chain rule,

$$\begin{aligned} \Delta(s, t) &= \frac{1}{st}(h(t) - h(0)) = \frac{1}{st} h'(\alpha t)\, t \\ &= \frac{1}{s}(f_x(x + \alpha t, y + s) - f_x(x + \alpha t, y)) \end{aligned}$$

for some $\alpha \in (0, 1)$. Applying the mean value theorem again,

$$\Delta(s, t) = f_{xy}(x + \alpha t, y + \beta s)$$

where $\alpha, \beta \in (0, 1)$.

If the terms $f(x + t, y)$ and $f(x, y + s)$ are interchanged in 6.20, $\Delta(s, t)$ is unchanged and the above argument shows there exist $\gamma, \delta \in (0, 1)$ such that

$$\Delta(s, t) = f_{yx}(x + \gamma t, y + \delta s).$$

Letting $(s, t) \to (0, 0)$ and using the continuity of f_{xy} and f_{yx} at (x, y),

$$\lim_{(s,t) \to (0,0)} \Delta(s, t) = f_{xy}(x, y) = f_{yx}(x, y).$$

This proves the theorem. ■

The following is obtained from the above by simply fixing all the variables except for the two of interest.

Corollary 6.10.2 *Suppose U is an open subset of X and $f : U \to \mathbb{R}$ has the property that for two indices, k, l, f_{x_k}, $f_{x_l}, f_{x_l x_k}$, and $f_{x_k x_l}$ exist on U and $f_{x_k x_l}$ and $f_{x_l x_k}$ are both continuous at $\mathbf{x} \in U$. Then $f_{x_k x_l}(\mathbf{x}) = f_{x_l x_k}(\mathbf{x})$.*

By considering the real and imaginary parts of f in the case where f has values in $\mathbb{C}$ you obtain the following corollary.

Corollary 6.10.3 *Suppose U is an open subset of $\mathbb{F}^n$ and $f : U \to \mathbb{F}$ has the property that for two indices, k, l, f_{x_k}, $f_{x_l}, f_{x_l x_k}$, and $f_{x_k x_l}$ exist on U and $f_{x_k x_l}$ and $f_{x_l x_k}$ are both continuous at $\mathbf{x} \in U$. Then $f_{x_k x_l}(\mathbf{x}) = f_{x_l x_k}(\mathbf{x})$.*

Finally, by considering the components of $\mathbf{f}$ you get the following generalization.

Corollary 6.10.4 *Suppose U is an open subset of $\mathbb{F}^n$ and $\mathbf{f} : U \to \mathbb{F}^m$ has the property that for two indices, k, l, $\mathbf{f}_{x_k}$, $\mathbf{f}_{x_l}, \mathbf{f}_{x_l x_k}$, and $\mathbf{f}_{x_k x_l}$ exist on U and $\mathbf{f}_{x_k x_l}$ and $\mathbf{f}_{x_l x_k}$ are both continuous at $\mathbf{x} \in U$. Then $\mathbf{f}_{x_k x_l}(\mathbf{x}) = \mathbf{f}_{x_l x_k}(\mathbf{x})$.*

It is necessary to assume the mixed partial derivatives are continuous in order to assert they are equal. The following is a well known example [3].

Example 6.10.5 *Let*

$$f(x,y) = \begin{cases} \frac{xy(x^2-y^2)}{x^2+y^2} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0) \end{cases}$$

From the definition of partial derivatives it follows immediately that $f_x(0,0) = f_y(0,0) = 0$. Using the standard rules of differentiation, for $(x,y) \neq (0,0)$,

$$f_x = y\frac{x^4 - y^4 + 4x^2y^2}{(x^2+y^2)^2},\ f_y = x\frac{x^4 - y^4 - 4x^2y^2}{(x^2+y^2)^2}$$

Now

$$\begin{aligned} f_{xy}(0,0) &\equiv \lim_{y\to 0} \frac{f_x(0,y) - f_x(0,0)}{y} \\ &= \lim_{y\to 0} \frac{-y^4}{(y^2)^2} = -1 \end{aligned}$$

while

$$\begin{aligned} f_{yx}(0,0) &\equiv \lim_{x\to 0} \frac{f_y(x,0) - f_y(0,0)}{x} \\ &= \lim_{x\to 0} \frac{x^4}{(x^2)^2} = 1 \end{aligned}$$

showing that although the mixed partial derivatives do exist at $(0,0)$, they are not equal there.

6.11 Implicit Function Theorem

The following lemma is very useful.

Lemma 6.11.1 *Let $A \in \mathcal{L}(X,X)$ where X is a finite dimensional normed vector space and suppose $||A|| \leq r < 1$. Then*

$$(I-A)^{-1} \text{ exists} \tag{6.21}$$

and

$$\left|\left|(I-A)^{-1}\right|\right| \leq (1-r)^{-1}. \tag{6.22}$$

Furthermore, if

$$\mathcal{I} \equiv \left\{A \in \mathcal{L}(X,X) : A^{-1} \text{ exists}\right\}$$

the map $A \to A^{-1}$ is continuous on $\mathcal{I}$ and $\mathcal{I}$ is an open subset of $\mathcal{L}(X,X)$.

Proof: Let $||A|| \leq r < 1$. If $(I-A)\mathbf{x} = \mathbf{0}$, then $\mathbf{x} = A\mathbf{x}$ and so if $\mathbf{x} \neq \mathbf{0}$,

$$||\mathbf{x}|| = ||A\mathbf{x}|| \leq ||A||\,||\mathbf{x}|| < r\,||\mathbf{x}||$$

which is a contradiction. Therefore, $(I-A)$ is one to one. Hence it maps a basis of X to a basis of X and is therefore, onto. Here is why. Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis for X and suppose

$$\sum_{k=1}^{n} c_k (I-A)\mathbf{v}_k = \mathbf{0}.$$

Then

$$(I-A)\left(\sum_{k=1}^{n} c_k \mathbf{v}_k\right) = \mathbf{0}$$

and since $(I-A)$ is one to one, it follows

$$\sum_{k=1}^{n} c_k \mathbf{v}_k = \mathbf{0}$$

which requires each $c_k = 0$ because the $\{\mathbf{v}_k\}$ are independent. Hence $\{(I-A)\mathbf{v}_k\}_{k=1}^{n}$ is a basis for X because there are n of these vectors and every basis has the same size. Therefore, if $\mathbf{y} \in X$, there exist scalars, c_k such that

$$\mathbf{y} = \sum_{k=1}^{n} c_k (I-A)\mathbf{v}_k = (I-A)\left(\sum_{k=1}^{n} c_k \mathbf{v}_k\right)$$

so $(I-A)$ is onto as claimed. Thus $(I-A)^{-1} \in \mathcal{L}(X,X)$ and it remains to estimate its norm.

$$||\mathbf{x} - A\mathbf{x}|| \geq ||\mathbf{x}|| - ||A\mathbf{x}|| \geq ||\mathbf{x}|| - ||A||\,||\mathbf{x}|| \geq ||\mathbf{x}||\,(1-r)$$

Letting $\mathbf{y} = \mathbf{x} - A\mathbf{x}$ so $\mathbf{x} = (I-A)^{-1}\mathbf{y}$, this shows, since $(I-A)$ is onto that for all $\mathbf{y} \in X$,

$$||\mathbf{y}|| \geq \left|\left|(I-A)^{-1}\mathbf{y}\right|\right|(1-r)$$

and so $\left|\left|(I-A)^{-1}\right|\right| \leq (1-r)^{-1}$. This proves the first part.

To verify the continuity of the inverse map, let $A \in \mathcal{I}$. Then

$$B = A\left(I - A^{-1}(A-B)\right)$$

and so if $\left|\left|A^{-1}(A-B)\right|\right| < 1$ which, by Theorem 5.8.3, happens if

$$||A-B|| < 1/\left|\left|A^{-1}\right|\right|,$$

it follows from the first part of this proof that $\left(I - A^{-1}(A-B)\right)^{-1}$ exists and so

$$B^{-1} = \left(I - A^{-1}(A-B)\right)^{-1} A^{-1}$$

which shows $\mathcal{I}$ is open. Also, if

$$\left|\left|A^{-1}(A-B)\right|\right| \leq r < 1, \tag{6.23}$$

$$\left|\left|B^{-1}\right|\right| \leq \left|\left|A^{-1}\right|\right| (1-r)^{-1}$$

Now for such B this close to A such that 6.23 holds,

$$\begin{aligned} \left|\left|B^{-1} - A^{-1}\right|\right| &= \left|\left|B^{-1}(A-B)A^{-1}\right|\right| \leq ||A-B|| \left|\left|B^{-1}\right|\right| \left|\left|A^{-1}\right|\right| \\ &\leq ||A-B|| \left|\left|A^{-1}\right|\right|^2 (1-r)^{-1} \end{aligned}$$

which shows the map which takes a linear transformation in $\mathcal{I}$ to its inverse is continuous. This proves the lemma. ■

The next theorem is a very useful result in many areas. It will be used in this section to give a short proof of the implicit function theorem but it is also useful in studying differential equations and integral equations. It is sometimes called the uniform contraction principle.

Theorem 6.11.2 *Let X, Y be finite dimensional normed vector spaces. Also let E be a closed subset of X and F a closed subset of Y. Suppose for each $(\mathbf{x},\mathbf{y}) \in E \times F$, $\mathbf{T}(\mathbf{x},\mathbf{y}) \in E$ and satisfies*

$$||\mathbf{T}(\mathbf{x},\mathbf{y}) - \mathbf{T}(\mathbf{x}',\mathbf{y})|| \leq r\,||\mathbf{x}-\mathbf{x}'|| \tag{6.24}$$

where $0 < r < 1$ and also

$$||\mathbf{T}(\mathbf{x},\mathbf{y}) - \mathbf{T}(\mathbf{x},\mathbf{y}')|| \leq M\,||\mathbf{y}-\mathbf{y}'||. \tag{6.25}$$

Then for each $\mathbf{y} \in F$ there exists a unique "fixed point" for $\mathbf{T}(\cdot,\mathbf{y}), \mathbf{x} \in E$, satisfying

$$\mathbf{T}(\mathbf{x},\mathbf{y}) = \mathbf{x} \tag{6.26}$$

and also if $\mathbf{x}(\mathbf{y})$ is this fixed point,

$$||\mathbf{x}(\mathbf{y}) - \mathbf{x}(\mathbf{y}')|| \leq \frac{M}{1-r}\,||\mathbf{y}-\mathbf{y}'||. \tag{6.27}$$

Proof: First consider the claim there exists a fixed point for the mapping, $\mathbf{T}(\cdot,y)$. For a fixed $\mathbf{y}$, let $\mathbf{g}(\mathbf{x}) \equiv \mathbf{T}(\mathbf{x},\mathbf{y})$. Now pick any $\mathbf{x}_0 \in E$ and consider the sequence,

$$\mathbf{x}_1 = \mathbf{g}(\mathbf{x}_0),\ \mathbf{x}_{k+1} = \mathbf{g}(\mathbf{x}_k).$$

Then by 6.24,

$$||\mathbf{x}_{k+1} - \mathbf{x}_k|| = ||\mathbf{g}(\mathbf{x}_k) - \mathbf{g}(\mathbf{x}_{k-1})|| \leq r\,||\mathbf{x}_k - \mathbf{x}_{k-1}|| \leq$$

$$r^2\,||\mathbf{x}_{k-1} - \mathbf{x}_{k-2}|| \leq \cdots \leq r^k\,||\mathbf{g}(\mathbf{x}_0) - \mathbf{x}_0||.$$

Now by the triangle inequality,

$$||\mathbf{x}_{k+p} - \mathbf{x}_k|| \leq \sum_{i=1}^{p} ||\mathbf{x}_{k+i} - \mathbf{x}_{k+i-1}||$$

$$\leq \sum_{i=1}^{p} r^{k+i-1} \left\| \mathbf{g}\left(\mathbf{x}_0\right) - \mathbf{x}_0 \right\| \leq \frac{r^k \left\| \mathbf{g}\left(\mathbf{x}_0\right) - \mathbf{x}_0 \right\|}{1-r}.$$

Since $0 < r < 1$, this shows that $\{\mathbf{x}_k\}_{k=1}^{\infty}$ is a Cauchy sequence. Therefore, by completeness of E it converges to a point $\mathbf{x} \in E$. To see $\mathbf{x}$ is a fixed point, use the continuify of $\mathbf{g}$ to obtain

$$\mathbf{x} \equiv \lim_{k\to\infty} \mathbf{x}_k = \lim_{k\to\infty} \mathbf{x}_{k+1} = \lim_{k\to\infty} \mathbf{g}\left(\mathbf{x}_k\right) = \mathbf{g}\left(\mathbf{x}\right).$$

This proves 6.26. To verify 6.27,

$$\left\| \mathbf{x}\left(\mathbf{y}\right) - \mathbf{x}\left(\mathbf{y}'\right) \right\| = \left\| \mathbf{T}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}\right) - \mathbf{T}\left(\mathbf{x}\left(\mathbf{y}'\right), \mathbf{y}'\right) \right\| \leq$$

$$\left\| \mathbf{T}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}\right) - \mathbf{T}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}'\right) \right\| + \left\| \mathbf{T}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}'\right) - \mathbf{T}\left(\mathbf{x}\left(\mathbf{y}'\right), \mathbf{y}'\right) \right\|$$

$$\leq M \left\| \mathbf{y} - \mathbf{y}' \right\| + r \left\| \mathbf{x}\left(\mathbf{y}\right) - \mathbf{x}\left(\mathbf{y}'\right) \right\|.$$

Thus

$$\left(1-r\right) \left\| \mathbf{x}\left(\mathbf{y}\right) - \mathbf{x}\left(\mathbf{y}'\right) \right\| \leq M \left\| \mathbf{y} - \mathbf{y}' \right\|.$$

This also shows the fixed point for a given y is unique. This proves the theorem. ■

The implicit function theorem deals with the question of solving, $\mathbf{f}\left(\mathbf{x}, \mathbf{y}\right) = \mathbf{0}$ for $\mathbf{x}$ in terms of $\mathbf{y}$ and how smooth the solution is. It is one of the most important theorems in mathematics. The proof I will give holds with no change in the context of infinite dimensional complete normed vector spaces when suitable modifications are made on what is meant by $\mathcal{L}\left(X, Y\right)$. There are also even more general versions of this theorem than to normed vector spaces.

Recall that for X, Y normed vector spaces, the norm on $X \times Y$ is of the form

$$\left\| \left(\mathbf{x}, \mathbf{y}\right) \right\| = \max\left(\left\| \mathbf{x} \right\|, \left\| \mathbf{y} \right\| \right).$$

Theorem 6.11.3 *(implicit function theorem) Let X, Y, Z be finite dimensional normed vector spaces and suppose U is an open set in $X \times Y$. Let $\mathbf{f} : U \to Z$ be in $C^1\left(U\right)$ and suppose*

$$\mathbf{f}\left(\mathbf{x}_0, \mathbf{y}_0\right) = \mathbf{0}, \; D_1\mathbf{f}\left(\mathbf{x}_0, \mathbf{y}_0\right)^{-1} \in \mathcal{L}\left(Z, X\right). \tag{6.28}$$

Then there exist positive constants, δ, η, such that for every $\mathbf{y} \in B\left(\mathbf{y}_0, \eta\right)$ there exists a unique $\mathbf{x}\left(\mathbf{y}\right) \in B\left(\mathbf{x}_0, \delta\right)$ such that

$$\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}\right) = \mathbf{0}. \tag{6.29}$$

Furthermore, the mapping, $\mathbf{y} \to \mathbf{x}\left(\mathbf{y}\right)$ is in $C^1\left(B\left(\mathbf{y}_0, \eta\right)\right)$.

Proof: Let $\mathbf{T}\left(\mathbf{x}, \mathbf{y}\right) \equiv \mathbf{x} - D_1\mathbf{f}\left(\mathbf{x}_0, \mathbf{y}_0\right)^{-1} \mathbf{f}\left(\mathbf{x}, \mathbf{y}\right)$. Therefore,

$$D_1\mathbf{T}\left(\mathbf{x}, \mathbf{y}\right) = I - D_1\mathbf{f}\left(\mathbf{x}_0, \mathbf{y}_0\right)^{-1} D_1\mathbf{f}\left(\mathbf{x}, \mathbf{y}\right). \tag{6.30}$$

by continuity of the derivative which implies continuity of $D_1\mathbf{T}$, it follows there exists $\delta > 0$ such that if $\left\| \left(\mathbf{x} - \mathbf{x}_0, \mathbf{y} - \mathbf{y}_0\right) \right\| < \delta$, then

$$\left\| D_1\mathbf{T}\left(\mathbf{x}, \mathbf{y}\right) \right\| < \frac{1}{2}. \tag{6.31}$$

Also, it can be assumed δ is small enough that

$$\left\| D_1\mathbf{f}\left(\mathbf{x}_0, \mathbf{y}_0\right)^{-1} \right\| \left\| D_2\mathbf{f}\left(\mathbf{x}, \mathbf{y}\right) \right\| < M \tag{6.32}$$

where $M > \left|\left|D_1\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)^{-1}\right|\right| ||D_2\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)||$. By Theorem 6.5.2, whenever $\mathbf{x},\mathbf{x}' \in B(\mathbf{x}_0,\delta)$ and $\mathbf{y} \in B(\mathbf{y}_0,\delta)$,

$$||\mathbf{T}(\mathbf{x},\mathbf{y}) - \mathbf{T}(\mathbf{x}',\mathbf{y})|| \leq \frac{1}{2}||\mathbf{x}-\mathbf{x}'||. \tag{6.33}$$

Solving 6.30 for $D_1\mathbf{f}(\mathbf{x},\mathbf{y})$,

$$D_1\mathbf{f}(\mathbf{x},\mathbf{y}) = D_1\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)(I - D_1\mathbf{T}(\mathbf{x},\mathbf{y})).$$

By Lemma 6.11.1 and the assumption that $D_1\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)^{-1}$ exists, it follows, $D_1\mathbf{f}(\mathbf{x},\mathbf{y})^{-1}$ exists and equals

$$(I - D_1\mathbf{T}(\mathbf{x},\mathbf{y}))^{-1} D_1\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)^{-1}$$

By the estimate of Lemma 6.11.1 and 6.31,

$$\left|\left|D_1\mathbf{f}(\mathbf{x},\mathbf{y})^{-1}\right|\right| \leq 2\left|\left|D_1\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)^{-1}\right|\right|. \tag{6.34}$$

Next more restrictions are placed on $\mathbf{y}$ to make it even closer to $\mathbf{y}_0$. Let

$$0 < \eta < \min\left(\delta, \frac{\delta}{3M}\right).$$

Then suppose $\mathbf{x} \in \overline{B(\mathbf{x}_0,\delta)}$ and $\mathbf{y} \in \overline{B(\mathbf{y}_0,\eta)}$. Consider

$$\mathbf{x} - D_1\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)^{-1}\mathbf{f}(\mathbf{x},\mathbf{y}) - \mathbf{x}_0 = \mathbf{T}(\mathbf{x},\mathbf{y}) - \mathbf{x}_0 \equiv \mathbf{g}(\mathbf{x},\mathbf{y}).$$

$$D_1\mathbf{g}(\mathbf{x},\mathbf{y}) = I - D_1\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)^{-1} D_1\mathbf{f}(\mathbf{x},\mathbf{y}) = D_1\mathbf{T}(\mathbf{x},\mathbf{y}),$$

and

$$D_2\mathbf{g}(\mathbf{x},\mathbf{y}) = -D_1\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)^{-1} D_2\mathbf{f}(\mathbf{x},\mathbf{y}).$$

Also note that $\mathbf{T}(\mathbf{x},\mathbf{y}) = \mathbf{x}$ is the same as saying $\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0) = \mathbf{0}$ and also $\mathbf{g}(\mathbf{x}_0,\mathbf{y}_0) = \mathbf{0}$. Thus by 6.32 and Theorem 6.5.2, it follows that for such $(\mathbf{x},\mathbf{y}) \in \overline{B(\mathbf{x}_0,\delta)} \times \overline{B(\mathbf{y}_0,\eta)}$,

$$\begin{aligned} ||\mathbf{T}(\mathbf{x},\mathbf{y}) - \mathbf{x}_0|| &= ||\mathbf{g}(\mathbf{x},\mathbf{y})|| = ||\mathbf{g}(\mathbf{x},\mathbf{y}) - \mathbf{g}(\mathbf{x}_0,\mathbf{y}_0)|| \\ &\leq ||\mathbf{g}(\mathbf{x},\mathbf{y}) - \mathbf{g}(\mathbf{x},\mathbf{y}_0)|| + ||\mathbf{g}(\mathbf{x},\mathbf{y}_0) - \mathbf{g}(\mathbf{x}_0,\mathbf{y}_0)|| \\ &\leq M||\mathbf{y}-\mathbf{y}_0|| + \frac{1}{2}||\mathbf{x}-\mathbf{x}_0|| < \frac{\delta}{2} + \frac{\delta}{3} = \frac{5\delta}{6} < \delta. \end{aligned} \tag{6.35}$$

Also for such $(\mathbf{x},\mathbf{y}_i)$, $i = 1,2$, Theorem 6.5.2 and 6.32 implies

$$\begin{aligned} ||\mathbf{T}(\mathbf{x},\mathbf{y}_1) - \mathbf{T}(\mathbf{x},\mathbf{y}_2)|| &= \left|\left|D_1\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)^{-1}(\mathbf{f}(\mathbf{x},\mathbf{y}_2) - \mathbf{f}(\mathbf{x},\mathbf{y}_1))\right|\right| \\ &\leq M||\mathbf{y}_2 - \mathbf{y}_1||. \end{aligned} \tag{6.36}$$

From now on assume $||\mathbf{x}-\mathbf{x}_0|| < \delta$ and $||\mathbf{y}-\mathbf{y}_0|| < \eta$ so that 6.36, 6.34, 6.35, 6.33, and 6.32 all hold. By 6.36, 6.33, 6.35, and the uniform contraction principle, Theorem 6.11.2 applied to $E \equiv \overline{B\left(\mathbf{x}_0, \frac{5\delta}{6}\right)}$ and $F \equiv \overline{B(\mathbf{y}_0,\eta)}$ implies that for each $\mathbf{y} \in B(\mathbf{y}_0,\eta)$, there exists a unique $\mathbf{x}(\mathbf{y}) \in B(\mathbf{x}_0,\delta)$ (actually in $\overline{B\left(\mathbf{x}_0, \frac{5\delta}{6}\right)}$) such that $\mathbf{T}(\mathbf{x}(\mathbf{y}),\mathbf{y}) = \mathbf{x}(\mathbf{y})$ which is equivalent to

$$\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y}) = \mathbf{0}.$$

Furthermore,

$$||\mathbf{x}(\mathbf{y}) - \mathbf{x}(\mathbf{y}')|| \leq 2M||\mathbf{y}-\mathbf{y}'||. \tag{6.37}$$

This proves the implicit function theorem except for the verification that $\mathbf{y} \to \mathbf{x}(\mathbf{y})$ is C^1. This is shown next. Letting $\mathbf{v}$ be sufficiently small, Theorem 6.9.5 and Theorem 6.5.2 imply

$$\mathbf{0} = \mathbf{f}(\mathbf{x}(\mathbf{y}+\mathbf{v}), \mathbf{y}+\mathbf{v}) - \mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y}) =$$

$$D_1\mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y})(\mathbf{x}(\mathbf{y}+\mathbf{v}) - \mathbf{x}(\mathbf{y})) +$$

$$+D_2\mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y})\mathbf{v} + o((\mathbf{x}(\mathbf{y}+\mathbf{v}) - \mathbf{x}(\mathbf{y}), \mathbf{v})).$$

The last term in the above is $o(\mathbf{v})$ because of 6.37. Therefore, using 6.34, solve the above equation for $\mathbf{x}(\mathbf{y}+\mathbf{v}) - \mathbf{x}(\mathbf{y})$ and obtain

$$\mathbf{x}(\mathbf{y}+\mathbf{v}) - \mathbf{x}(\mathbf{y}) = -D_1(\mathbf{x}(\mathbf{y}), \mathbf{y})^{-1} D_2\mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y})\mathbf{v} + o(\mathbf{v})$$

Which shows that $\mathbf{y} \to \mathbf{x}(\mathbf{y})$ is differentiable on $B(\mathbf{y}_0, \eta)$ and

$$D\mathbf{x}(\mathbf{y}) = -D_1\mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y})^{-1} D_2\mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y}). \tag{6.38}$$

Now it follows from the continuity of $D_2\mathbf{f}$, $D_1\mathbf{f}$, the inverse map, 6.37, and this formula for $D\mathbf{x}(\mathbf{y})$that $\mathbf{x}(\cdot)$ is $C^1(B(\mathbf{y}_0, \eta))$. This proves the theorem. ■

The next theorem is a very important special case of the implicit function theorem known as the inverse function theorem. Actually one can also obtain the implicit function theorem from the inverse function theorem. It is done this way in [28] and in [2].

Theorem 6.11.4 *(inverse function theorem) Let $\mathbf{x}_0 \in U$, an open set in X , and let $\mathbf{f} : U \to Y$ where X, Y are finite dimensional normed vector spaces. Suppose*

$$\mathbf{f} \text{ is } C^1(U) \ , \quad \text{and} \quad D\mathbf{f}(\mathbf{x}_0)^{-1} \in \mathcal{L}(Y, X). \tag{6.39}$$

Then there exist open sets W, and V such that

$$\mathbf{x}_0 \in W \subseteq U, \tag{6.40}$$

$$\mathbf{f} : W \to V \text{ is one to one and onto}, \tag{6.41}$$

$$\mathbf{f}^{-1} \text{ is } C^1, \tag{6.42}$$

Proof: Apply the implicit function theorem to the function

$$\mathbf{F}(\mathbf{x}, \mathbf{y}) \equiv \mathbf{f}(\mathbf{x}) - \mathbf{y}$$

where $\mathbf{y}_0 \equiv \mathbf{f}(\mathbf{x}_0)$. Thus the function $\mathbf{y} \to \mathbf{x}(\mathbf{y})$ defined in that theorem is $\mathbf{f}^{-1}$. Now let

$$W \equiv B(\mathbf{x}_0, \delta) \cap \mathbf{f}^{-1}(B(\mathbf{y}_0, \eta))$$

and

$$V \equiv B(\mathbf{y}_0, \eta).$$

This proves the theorem. ■

6.11.1 More Derivatives

In the implicit function theorem, suppose $\mathbf{f}$ is C^k. Will the implicitly defined function also be C^k? It was shown above that this is the case if $k = 1$. In fact it holds for any positive integer k.

First of all, consider $D_2\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y}) \in \mathcal{L}(Y,Z)$. Let $\{\mathbf{w}_1,\cdots,\mathbf{w}_m\}$ be a basis for Y and let $\{\mathbf{z}_1,\cdots,\mathbf{z}_n\}$ be a basis for Z. Then $D_2\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})$ has a matrix with respect to these bases. Thus conserving on notation, denote this matrix by $\left(D_2\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})_{ij}\right)$. Thus

$$D_2\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y}) = \sum_{ij} D_2\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})_{ij}\,\mathbf{z}_i\mathbf{w}_j$$

The scalar valued entries of the matrix of $D_2\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})$ have the same differentiability as the function $\mathbf{y} \to D_2\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})$. This is because the linear projection map, π_{ij} mapping $\mathcal{L}(Y,Z)$ to $\mathbb{F}$ given by $\pi_{ij}L \equiv L_{ij}$, the ij^{th} entry of the matrix of L with respect to the given bases is continuous thanks to Theorem 5.8.3. Similar considerations apply to $D_1\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})$ and the entries of its matrix, $D_1\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})_{ij}$ taken with respect to suitable bases. From the formula for the inverse of a matrix, Theorem 3.5.14, the ij^{th} entries of the matrix of $D_1\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})^{-1}$, $D_1\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})^{-1}_{ij}$ also have the same differentiability as $\mathbf{y} \to D_1\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})$.

Now consider the formula for the derivative of the implicitly defined function in 6.38,

$$D\mathbf{x}(\mathbf{y}) = -D_1\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})^{-1} D_2\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y}). \tag{6.43}$$

The above derivative is in $\mathcal{L}(Y,X)$. Let $\{\mathbf{w}_1,\cdots,\mathbf{w}_m\}$ be a basis for Y and let $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ be a basis for X. Letting x_i be the i^{th} component of $\mathbf{x}$ with respect to the basis for X, it follows from Theorem 6.8.1, $\mathbf{y} \to \mathbf{x}(\mathbf{y})$ will be C^k if all such Gateaux derivatives, $D_{\mathbf{w}_{j_1}\mathbf{w}_{j_2}\cdots\mathbf{w}_{j_r}} x_i(\mathbf{y})$ exist and are continuous for $r \le k$ and for any i. Consider what is required for this to happen. By 6.43,

$$\begin{aligned} D_{\mathbf{w}_j} x_i(\mathbf{y}) &= \sum_k \left(-D_1\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})^{-1}\right)_{ik} (D_2\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y}))_{kj} \\ &\equiv G_1(\mathbf{x}(\mathbf{y}),\mathbf{y}) \end{aligned} \tag{6.44}$$

where $(\mathbf{x},\mathbf{y}) \to G_1(\mathbf{x},\mathbf{y})$ is C^{k-1} because it is assumed $\mathbf{f}$ is C^k and one derivative has been taken to write the above. If $k \ge 2$, then another Gateaux derivative can be taken.

$$\begin{aligned} D_{\mathbf{w}_j\mathbf{w}_k} x_i(\mathbf{y}) &\equiv \lim_{t\to 0} \frac{G_1(\mathbf{x}(\mathbf{y}+t\mathbf{w}_k),\mathbf{y}+t\mathbf{w}_k) - G_1(\mathbf{x}(\mathbf{y}),\mathbf{y})}{t} \\ &= D_1G_1(\mathbf{x}(\mathbf{y}),\mathbf{y})\,D\mathbf{x}(\mathbf{y})\,\mathbf{w}_k + D_2G_1(\mathbf{x}(\mathbf{y}),\mathbf{y}) \\ &\equiv G_2(\mathbf{x}(\mathbf{y}),\mathbf{y},D\mathbf{x}(\mathbf{y})) \end{aligned}$$

Since a similar result holds for all i and any choice of $\mathbf{w}_j, \mathbf{w}_k$, this shows $\mathbf{x}$ is at least C^2. If $k \ge 3$, then another Gateaux derivative can be taken because then $(\mathbf{x},\mathbf{y},\mathbf{z}) \to G_2(\mathbf{x},\mathbf{y},\mathbf{z})$ is C^1 and it has been established $D\mathbf{x}$ is C^1. Continuing this way shows $D_{\mathbf{w}_{j_1}\mathbf{w}_{j_2}\cdots\mathbf{w}_{j_r}} x_i(\mathbf{y})$ exists and is continuous for $r \le k$. This proves the following corollary to the implicit and inverse function theorems.

Corollary 6.11.5 *In the implicit and inverse function theorems, you can replace C^1 with C^k in the statements of the theorems for any $k \in \mathbb{N}$.*

6.11.2 The Case Of $\mathbb{R}^n$

In many applications of the implicit function theorem,

$$\mathbf{f} : U \subseteq \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$$

and $\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}$ while $\mathbf{f}$ is C^1. How can you recognize the condition of the implicit function theorem which says $D_1\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0)^{-1}$ exists? This is really not hard. You recall the matrix of the transformation $D_1\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0)$ with respect to the usual basis vectors is

$$\begin{pmatrix} f_{1,x_1}(\mathbf{x}_0, \mathbf{y}_0) & \cdots & f_{1,x_n}(\mathbf{x}_0, \mathbf{y}_0) \\ \vdots & & \vdots \\ f_{n,x_1}(\mathbf{x}_0, \mathbf{y}_0) & \cdots & f_{n,x_n}(\mathbf{x}_0, \mathbf{y}_0) \end{pmatrix}$$

and so $D_1\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0)^{-1}$ exists exactly when the determinant of the above matrix is nonzero. This is the condition to check. In the general case, you just need to verify $D_1\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0)$ is one to one and this can also be accomplished by looking at the matrix of the transformation with respect to some bases on X and Z.

6.12 Taylor's Formula

First recall the Taylor formula with the Lagrange form of the remainder. It will only be needed on $[0,1]$ so that is what I will show.

Theorem 6.12.1 *Let $h : [0,1] \to \mathbb{R}$ have $m+1$ derivatives. Then there exists $t \in (0,1)$ such that*

$$h(1) = h(0) + \sum_{k=1}^{m} \frac{h^{(k)}(0)}{k!} + \frac{h^{(m+1)}(t)}{(m+1)!}.$$

Proof: Let K be a number chosen such that

$$h(1) - \left(h(0) + \sum_{k=1}^{m} \frac{h^{(k)}(0)}{k!} + K \right) = 0$$

Now the idea is to find K. To do this, let

$$F(t) = h(1) - \left(h(t) + \sum_{k=1}^{m} \frac{h^{(k)}(t)}{k!} (1-t)^k + K(1-t)^{m+1} \right)$$

Then $F(1) = F(0) = 0$. Therefore, by Rolle's theorem there exists t between 0 and 1 such that $F'(t) = 0$. Thus,

$$\begin{aligned} 0 &= -F'(t) = h'(t) + \sum_{k=1}^{m} \frac{h^{(k+1)}(t)}{k!} (1-t)^k \\ &\quad - \sum_{k=1}^{m} \frac{h^{(k)}(t)}{k!} k(1-t)^{k-1} - K(m+1)(1-t)^m \end{aligned}$$

And so

$$\begin{aligned} &= h'(t) + \sum_{k=1}^{m} \frac{h^{(k+1)}(t)}{k!} (1-t)^k - \sum_{k=0}^{m-1} \frac{h^{(k+1)}(t)}{k!} (1-t)^k \\ &\quad -K(m+1)(1-t)^m \end{aligned}$$

$$= h'(t) + \frac{h^{(m+1)}(t)}{m!} (1-t)^m - h'(t) - K(m+1)(1-t)^m$$

and so

$$K = \frac{h^{(m+1)}(t)}{(m+1)!}.$$

This proves the theorem. ■

Now let $f : U \to \mathbb{R}$ where $U \subseteq X$ a normed vector space and suppose $f \in C^m(U)$. Let $\mathbf{x} \in U$ and let $r > 0$ be such that

$$B(\mathbf{x},r) \subseteq U.$$

Then for $||\mathbf{v}|| < r$ consider

$$f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x}) \equiv h(t)$$

for $t \in [0,1]$. Then by the chain rule,

$$h'(t) = Df(\mathbf{x}+t\mathbf{v})(\mathbf{v}),\ h''(t) = D^2 f(\mathbf{x}+t\mathbf{v})(\mathbf{v})(\mathbf{v})$$

and continuing in this way,

$$h^{(k)}(t) = D^{(k)} f(\mathbf{x}+t\mathbf{v})(\mathbf{v})(\mathbf{v})\cdots(\mathbf{v}) \equiv D^{(k)} f(\mathbf{x}+t\mathbf{v})\,\mathbf{v}^k.$$

It follows from Taylor's formula for a function of one variable given above that

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \sum_{k=1}^{m} \frac{D^{(k)} f(\mathbf{x})\,\mathbf{v}^k}{k!} + \frac{D^{(m+1)} f(\mathbf{x}+t\mathbf{v})\,\mathbf{v}^{m+1}}{(m+1)!}. \tag{6.45}$$

This proves the following theorem.

Theorem 6.12.2 *Let $f : U \to \mathbb{R}$ and let $f \in C^{m+1}(U)$. Then if*

$$B(\mathbf{x},r) \subseteq U,$$

and $||\mathbf{v}|| < r$, there exists $t \in (0,1)$ such that 6.45 holds.

6.12.1 Second Derivative Test

Now consider the case where $U \subseteq \mathbb{R}^n$ and $f : U \to \mathbb{R}$ is $C^2(U)$. Then from Taylor's theorem, if $\mathbf{v}$ is small enough, there exists $t \in (0,1)$ such that

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + Df(\mathbf{x})\,\mathbf{v} + \frac{D^2 f(\mathbf{x}+t\mathbf{v})\,\mathbf{v}^2}{2}. \tag{6.46}$$

Consider

$$\begin{aligned} D^2 f(\mathbf{x}+t\mathbf{v})(\mathbf{e}_i)(\mathbf{e}_j) &\equiv D\left(D\left(f(\mathbf{x}+t\mathbf{v})\right)\mathbf{e}_i\right)\mathbf{e}_j \\ &= D\left(\frac{\partial f(\mathbf{x}+t\mathbf{v})}{\partial x_i}\right)\mathbf{e}_j \\ &= \frac{\partial^2 f(\mathbf{x}+t\mathbf{v})}{\partial x_j \partial x_i} \end{aligned}$$

where $\mathbf{e}_i$ are the usual basis vectors. Letting

$$\mathbf{v} = \sum_{i=1}^{n} v_i \mathbf{e}_i,$$

the second derivative term in 6.46 reduces to

$$\frac{1}{2}\sum_{i,j} D^2 f(\mathbf{x}+t\mathbf{v})(\mathbf{e}_i)(\mathbf{e}_j)\,v_i v_j = \frac{1}{2}\sum_{i,j} H_{ij}(\mathbf{x}+t\mathbf{v})\,v_i v_j$$

where

$$H_{ij}(\mathbf{x}+t\mathbf{v}) = D^2 f(\mathbf{x}+t\mathbf{v})(\mathbf{e}_i)(\mathbf{e}_j) = \frac{\partial^2 f(\mathbf{x}+t\mathbf{v})}{\partial x_j \partial x_i}.$$

Definition 6.12.3 *The matrix whose ij^{th} entry is $\frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_i}$ is called the Hessian matrix, denoted as $\mathbf{H}(\mathbf{x})$.*

From Theorem 6.10.1, this is a symmetric real matrix, thus self adjoint. By the continuity of the second partial derivative,

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + Df(\mathbf{x})\mathbf{v} + \frac{1}{2}\mathbf{v}^T H(\mathbf{x})\mathbf{v} + \frac{1}{2}\left(\mathbf{v}^T\left(H(\mathbf{x}+t\mathbf{v}) - H(\mathbf{x})\right)\mathbf{v}\right). \tag{6.47}$$

where the last two terms involve ordinary matrix multiplication and

$$\mathbf{v}^T = (v_1 \cdots v_n)$$

for v_i the components of $\mathbf{v}$ relative to the standard basis.

Definition 6.12.4 *Let $f : D \to \mathbb{R}$ where D is a subset of some normed vector space. Then f has a local minimum at $\mathbf{x} \in D$ if there exists $\delta > 0$ such that for all $\mathbf{y} \in B(\mathbf{x}, \delta)$*

$$f(\mathbf{y}) \geq f(\mathbf{x}).$$

f has a local maximum at $\mathbf{x} \in D$ if there exists $\delta > 0$ such that for all $\mathbf{y} \in B(\mathbf{x}, \delta)$

$$f(\mathbf{y}) \leq f(\mathbf{x}).$$

Theorem 6.12.5 *If $f : U \to \mathbb{R}$ where U is an open subset of $\mathbb{R}^n$ and f is C^2, suppose $Df(\mathbf{x}) = 0$. Then if $H(\mathbf{x})$ has all positive eigenvalues, $\mathbf{x}$ is a local minimum. If the Hessian matrix $H(\mathbf{x})$ has all negative eigenvalues, then $\mathbf{x}$ is a local maximum. If $H(\mathbf{x})$ has a positive eigenvalue, then there exists a direction in which f has a local minimum at $\mathbf{x}$, while if $H(\mathbf{x})$ has a negative eigenvalue, there exists a direction in which $H(\mathbf{x})$ has a local maximum at $\mathbf{x}$.*

Proof: Since $Df(\mathbf{x}) = 0$, formula 6.47 holds and by continuity of the second derivative, $H(\mathbf{x})$ is a symmetric matrix. Thus $H(\mathbf{x})$ has all real eigenvalues. Suppose first that $H(\mathbf{x})$ has all positive eigenvalues and that all are larger than $\delta^2 > 0$. Then by Theorem 3.8.23, $H(\mathbf{x})$ has an orthonormal basis of eigenvectors, $\{\mathbf{v}_i\}_{i=1}^n$ and if $\mathbf{u}$ is an arbitrary vector, such that $\mathbf{u} = \sum_{j=1}^n u_j \mathbf{v}_j$ where $u_j = \mathbf{u} \cdot \mathbf{v}_j$, then

$$\mathbf{u}^T H(\mathbf{x})\mathbf{u} = \sum_{j=1}^n u_j \mathbf{v}_j^T H(\mathbf{x}) \sum_{j=1}^n u_j \mathbf{v}_j$$

$$= \sum_{j=1}^n u_j^2 \lambda_j \geq \delta^2 \sum_{j=1}^n u_j^2 = \delta^2 |\mathbf{u}|^2.$$

From 6.47 and the continuity of H, if $\mathbf{v}$ is small enough,

$$f(\mathbf{x}+\mathbf{v}) \geq f(\mathbf{x}) + \frac{1}{2}\delta^2 |\mathbf{v}|^2 - \frac{1}{4}\delta^2 |\mathbf{v}|^2 = f(\mathbf{x}) + \frac{\delta^2}{4}|\mathbf{v}|^2.$$

This shows the first claim of the theorem. The second claim follows from similar reasoning. Suppose $H(\mathbf{x})$ has a positive eigenvalue λ^2. Then let $\mathbf{v}$ be an eigenvector for this eigenvalue. Then from 6.47,

$$f(\mathbf{x}+t\mathbf{v}) = f(\mathbf{x}) + \frac{1}{2}t^2 \mathbf{v}^T H(\mathbf{x})\mathbf{v} +$$

$$\frac{1}{2}t^2\left(\mathbf{v}^T\left(H\left(\mathbf{x}+t\mathbf{v}\right)-H\left(\mathbf{x}\right)\right)\mathbf{v}\right)$$

which implies

$$\begin{aligned} f\left(\mathbf{x}+t\mathbf{v}\right) &= f\left(\mathbf{x}\right)+\frac{1}{2}t^2\lambda^2\left|\mathbf{v}\right|^2+\frac{1}{2}t^2\left(\mathbf{v}^T\left(H\left(\mathbf{x}+t\mathbf{v}\right)-H\left(\mathbf{x}\right)\right)\mathbf{v}\right) \\ &\geq f\left(\mathbf{x}\right)+\frac{1}{4}t^2\lambda^2\left|\mathbf{v}\right|^2 \end{aligned}$$

whenever t is small enough. Thus in the direction $\mathbf{v}$ the function has a local minimum at $\mathbf{x}$. The assertion about the local maximum in some direction follows similarly. This proves the theorem. ■

This theorem is an analogue of the second derivative test for higher dimensions. As in one dimension, when there is a zero eigenvalue, it may be impossible to determine from the Hessian matrix what the local qualitative behavior of the function is. For example, consider

$$f_1\left(x,y\right)=x^4+y^2,\ f_2\left(x,y\right)=-x^4+y^2.$$

Then $Df_i\left(0,0\right)=\mathbf{0}$ and for both functions, the Hessian matrix evaluated at $(0,0)$ equals

$$\begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}$$

but the behavior of the two functions is very different near the origin. The second has a saddle point while the first has a minimum there.

6.13 The Method Of Lagrange Multipliers

As an application of the implicit function theorem, consider the method of Lagrange multipliers from calculus. Recall the problem is to maximize or minimize a function subject to equality constraints. Let $f:U\to\mathbb{R}$ be a C^1 function where $U\subseteq\mathbb{R}^n$ and let

$$g_i\left(\mathbf{x}\right)=0,\ i=1,\cdots,m \tag{6.48}$$

be a collection of equality constraints with $m<n$. Now consider the system of nonlinear equations

$$\begin{aligned} f\left(\mathbf{x}\right) &= a \\ g_i\left(\mathbf{x}\right) &= 0,\ i=1,\cdots,m. \end{aligned}$$

$\mathbf{x}_0$ is a local maximum if $f\left(\mathbf{x}_0\right)\geq f\left(\mathbf{x}\right)$ for all $\mathbf{x}$ near $\mathbf{x}_0$ which also satisfies the constraints 6.48. A local minimum is defined similarly. Let $\mathbf{F}:U\times\mathbb{R}\to\mathbb{R}^{m+1}$ be defined by

$$\mathbf{F}\left(\mathbf{x},a\right)\equiv\begin{pmatrix} f\left(\mathbf{x}\right)-a \\ g_1\left(\mathbf{x}\right) \\ \vdots \\ g_m\left(\mathbf{x}\right) \end{pmatrix}. \tag{6.49}$$

Now consider the $m+1\times n$ Jacobian matrix, the matrix of the linear transformation, $D_1\mathbf{F}\left(\mathbf{x},a\right)$ with respect to the usual basis for $\mathbb{R}^n$ and $\mathbb{R}^{m+1}$.

$$\begin{pmatrix} f_{x_1}\left(\mathbf{x}_0\right) & \cdots & f_{x_n}\left(\mathbf{x}_0\right) \\ g_{1x_1}\left(\mathbf{x}_0\right) & \cdots & g_{1x_n}\left(\mathbf{x}_0\right) \\ \vdots & & \vdots \\ g_{mx_1}\left(\mathbf{x}_0\right) & \cdots & g_{mx_n}\left(\mathbf{x}_0\right) \end{pmatrix}.$$

If this matrix has rank $m+1$ then some $m+1 \times m+1$ submatrix has nonzero determinant. It follows from the implicit function theorem that there exist $m+1$ variables, $x_{i_1}, \cdots, x_{i_{m+1}}$ such that the system

$$\mathbf{F}(\mathbf{x},a) = \mathbf{0} \tag{6.50}$$

specifies these $m+1$ variables as a function of the remaining $n-(m+1)$ variables and a in an open set of $\mathbb{R}^{n-m}$. Thus there is a solution $(\mathbf{x},a)$ to 6.50 for some $\mathbf{x}$ close to $\mathbf{x}_0$ whenever a is in some open interval. Therefore, $\mathbf{x}_0$ cannot be either a local minimum or a local maximum. It follows that if $\mathbf{x}_0$ is either a local maximum or a local minimum, then the above matrix must have rank less than $m+1$ which, by Corollary 3.5.20, requires the rows to be linearly dependent. Thus, there exist m scalars,

$$\lambda_1, \cdots, \lambda_m,$$

and a scalar μ, not all zero such that

$$\mu \begin{pmatrix} f_{x_1}(\mathbf{x}_0) \\ \vdots \\ f_{x_n}(\mathbf{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\mathbf{x}_0) \\ \vdots \\ g_{1x_n}(\mathbf{x}_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(\mathbf{x}_0) \\ \vdots \\ g_{mx_n}(\mathbf{x}_0) \end{pmatrix}. \tag{6.51}$$

If the column vectors

$$\begin{pmatrix} g_{1x_1}(\mathbf{x}_0) \\ \vdots \\ g_{1x_n}(\mathbf{x}_0) \end{pmatrix}, \cdots \begin{pmatrix} g_{mx_1}(\mathbf{x}_0) \\ \vdots \\ g_{mx_n}(\mathbf{x}_0) \end{pmatrix} \tag{6.52}$$

are linearly independent, then, $\mu \neq 0$ and dividing by μ yields an expression of the form

$$\begin{pmatrix} f_{x_1}(\mathbf{x}_0) \\ \vdots \\ f_{x_n}(\mathbf{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\mathbf{x}_0) \\ \vdots \\ g_{1x_n}(\mathbf{x}_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(\mathbf{x}_0) \\ \vdots \\ g_{mx_n}(\mathbf{x}_0) \end{pmatrix} \tag{6.53}$$

at every point $\mathbf{x}_0$ which is either a local maximum or a local minimum. This proves the following theorem.

Theorem 6.13.1 *Let U be an open subset of $\mathbb{R}^n$ and let $f : U \to \mathbb{R}$ be a C^1 function. Then if $\mathbf{x}_0 \in U$ is either a local maximum or local minimum of f subject to the constraints 6.48, then 6.51 must hold for some scalars $\mu, \lambda_1, \cdots, \lambda_m$ not all equal to zero. If the vectors in 6.52 are linearly independent, it follows that an equation of the form 6.53 holds.*

6.14 Exercises

1. Suppose $L \in \mathcal{L}(X,Y)$ and suppose L is one to one. Show there exists $r > 0$ such that for all $\mathbf{x} \in X$,
$$||L\mathbf{x}|| \geq r\,||\mathbf{x}||.$$
Hint: You might argue that $|||\mathbf{x}||| \equiv ||L\mathbf{x}||$ is a norm.

2. Show every polynomial, $\sum_{|\boldsymbol{\alpha}| \leq k} d_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$ is C^k for every k.

3. If $f : U \to \mathbb{R}$ where U is an open set in X and f is C^2, show the mixed Gateaux derivatives, $D_{\mathbf{v}_1\mathbf{v}_2} f(\mathbf{x})$ and $D_{\mathbf{v}_2\mathbf{v}_1} f(\mathbf{x})$ are equal.

4. Give an example of a function which is differentiable everywhere but at some point it fails to have continuous partial derivatives. Thus this function will be an example of a differentiable function which is not C^1.

5. The existence of partial derivatives does not imply continuity as was shown in an example. However, much more can be said than this. Consider

$$f(x,y) = \begin{cases} \frac{\left(x^2-y^4\right)^2}{(x^2+y^4)^2} \text{ if } (x,y) \neq (0,0), \\ 1 \text{ if } (x,y) = (0,0). \end{cases}$$

Show each Gateaux derivative, $D_{\mathbf{v}} f(\mathbf{0})$ exists and equals 0 for every $\mathbf{v}$. Also show each Gateaux derivative exists at every other point in $\mathbb{R}^2$. Now consider the curve $x^2 = y^4$ and the curve $y = 0$ to verify the function fails to be continuous at $(0,0)$. This is an example of an everywhere Gateaux differentiable function which is not differentiable and not continuous.

6. Let f be a real valued function defined on $\mathbb{R}^2$ by

$$f(x,y) \equiv \begin{cases} \frac{x^3-y^3}{x^2+y^2} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0) \end{cases}$$

Determine whether f is continuous at $(0,0)$. Find $f_x(0,0)$ and $f_y(0,0)$. Are the partial derivatives of f continuous at $(0,0)$? Find $D_{(u,v)} f((0,0)), \lim_{t\to 0} \frac{f(t(u,v))}{t}$. Is the mapping $(u,v) \to D_{(u,v)} f((0,0))$ linear? Is f differentiable at $(0,0)$?

7. Let $f : V \to \mathbb{R}$ where V is a finite dimensional normed vector space. Suppose f is convex which means

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t) f(\mathbf{y})$$

whenever $t \in [0,1]$. Suppose also that f is differentiable. Show then that for every $\mathbf{x}, \mathbf{y} \in V$,

$$(Df(\mathbf{x}) - Df(\mathbf{y}))(\mathbf{x} - \mathbf{y}) \geq 0.$$

8. Suppose $f : U \subseteq V \to \mathbb{F}$ where U is an open subset of V, a finite dimensional inner product space with the inner product denoted by $(\cdot,\cdot)$. Suppose f is differentiable. Show there exists a unique vector $\mathbf{v}(\mathbf{x}) \in V$ such that

$$(\mathbf{u} \cdot \mathbf{v}(\mathbf{x})) = Df(\mathbf{x})\mathbf{u}.$$

This special vector is called the gradient and is usually denoted by $\nabla f(\mathbf{x})$. **Hint:** You might review the Riesz representation theorem presented earlier.

9. Suppose $\mathbf{f} : U \to Y$ where U is an open subset of X, a finite dimensional normed vector space. Suppose that for all $\mathbf{v} \in X, D_{\mathbf{v}}\mathbf{f}(\mathbf{x})$ exists. Show that whenever $a \in \mathbb{F}$ $D_{a\mathbf{v}}\mathbf{f}(\mathbf{x}) = aD_{\mathbf{v}}\mathbf{f}(\mathbf{x})$. Explain why if $\mathbf{x} \to D_{\mathbf{v}}\mathbf{f}(\mathbf{x})$ is continuous then $\mathbf{v} \to D_{\mathbf{v}}\mathbf{f}(\mathbf{x})$ is linear. Show that if $\mathbf{f}$ is differentiable at $\mathbf{x}$, then $D_{\mathbf{v}}\mathbf{f}(\mathbf{x}) = D\mathbf{f}(\mathbf{x})\mathbf{v}$.

10. Suppose B is an open ball in X and $\mathbf{f} : B \to Y$ is differentiable. Suppose also there exists $L \in \mathcal{L}(X,Y)$ such that

$$||D\mathbf{f}(\mathbf{x}) - L|| < k$$

for all $\mathbf{x} \in B$. Show that if $\mathbf{x}_1, \mathbf{x}_2 \in B$,

$$|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2) - L(\mathbf{x}_1 - \mathbf{x}_2)| \leq k|\mathbf{x}_1 - \mathbf{x}_2|.$$

Hint: Consider $T\mathbf{x} = \mathbf{f}(\mathbf{x}) - L\mathbf{x}$ and argue $||DT(\mathbf{x})|| < k$. Then consider Theorem 6.5.2.

11. Let U be an open subset of X, $\mathbf{f} : U \to Y$ where X, Y are finite dimensional normed vector spaces and suppose $\mathbf{f} \in C^1(U)$ and $D\mathbf{f}(\mathbf{x}_0)$ is one to one. Then show $\mathbf{f}$ is one to one near $\mathbf{x}_0$. **Hint:** Show using the assumption that $\mathbf{f}$ is C^1 that there exists $\delta > 0$ such that if

$$\mathbf{x}_1, \mathbf{x}_2 \in B(\mathbf{x}_0, \delta),$$

then

$$|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2) - D\mathbf{f}(\mathbf{x}_0)(\mathbf{x}_1 - \mathbf{x}_2)| \leq \frac{r}{2} |\mathbf{x}_1 - \mathbf{x}_2| \tag{6.54}$$

then use Problem 1.

12. Suppose $M \in \mathcal{L}(X, Y)$ and suppose M is onto. Show there exists $L \in \mathcal{L}(Y, X)$ such that

$$LM\mathbf{x} = P\mathbf{x}$$

where $P \in \mathcal{L}(X, X)$, and $P^2 = P$. Also show L is one to one and onto. **Hint:** Let $\{\mathbf{y}_1, \cdots, \mathbf{y}_m\}$ be a basis of Y and let $M\mathbf{x}_i = \mathbf{y}_i$. Then define

$$L\mathbf{y} = \sum_{i=1}^{m} \alpha_i \mathbf{x}_i \text{ where } \mathbf{y} = \sum_{i=1}^{m} \alpha_i \mathbf{y}_i.$$

Show $\{\mathbf{x}_1, \cdots, \mathbf{x}_m\}$ is a linearly independent set and show you can obtain $\{\mathbf{x}_1, \cdots, \mathbf{x}_m, \cdots, \mathbf{x}_n\}$, a basis for X in which $M\mathbf{x}_j = \mathbf{0}$ for $j > m$. Then let

$$P\mathbf{x} \equiv \sum_{i=1}^{m} \alpha_i \mathbf{x}_i$$

where

$$\mathbf{x} = \sum_{i=1}^{m} \alpha_i \mathbf{x}_i.$$

13. This problem depends on the result of Problem 12. Let $\mathbf{f} : U \subseteq X \to Y$, $\mathbf{f}$ is C^1, and $D\mathbf{f}(\mathbf{x})$ is onto for each $\mathbf{x} \in U$. Then show $\mathbf{f}$ maps open subsets of U onto open sets in Y. **Hint:** Let $P = LD\mathbf{f}(\mathbf{x})$ as in Problem 12. Argue L maps open sets from Y to open sets of the vector space $X_1 \equiv PX$ and L^{-1} maps open sets from X_1 to open sets of Y. Then $L\mathbf{f}(\mathbf{x} + \mathbf{v}) = L\mathbf{f}(\mathbf{x}) + LD\mathbf{f}(\mathbf{x})\mathbf{v} + \mathbf{o}(\mathbf{v})$. Now for $\mathbf{z} \in X_1$, let $\mathbf{h}(\mathbf{z}) = L\mathbf{f}(\mathbf{x} + \mathbf{z}) - L\mathbf{f}(\mathbf{x})$. Then $\mathbf{h}$ is C^1 on some small open subset of X_1 containing $\mathbf{0}$ and $D\mathbf{h}(\mathbf{0}) = LD\mathbf{f}(\mathbf{x})$ which is seen to be one to one and onto and in $\mathcal{L}(X_1, X_1)$. Therefore, if r is small enough, $\mathbf{h}(B(\mathbf{0},r))$ equals an open set in X_1, V. This is by the inverse function theorem. Hence $L(\mathbf{f}(\mathbf{x} + B(\mathbf{0},r)) - \mathbf{f}(\mathbf{x})) = V$ and so $\mathbf{f}(\mathbf{x} + B(\mathbf{0},r)) - \mathbf{f}(\mathbf{x}) = L^{-1}(V)$, an open set in Y.

14. Suppose $U \subseteq \mathbb{R}^2$ is an open set and $\mathbf{f} : U \to \mathbb{R}^3$ is C^1. Suppose $D\mathbf{f}(s_0, t_0)$ has rank two and

$$\mathbf{f}(s_0, t_0) = \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix}.$$

Show that for (s, t) near (s_0, t_0), the points $\mathbf{f}(s, t)$ may be realized in one of the following forms.

$$\{(x, y, \phi(x, y)) : (x, y) \text{ near } (x_0, y_0)\},$$

$$\{(\phi(y, z), y, z) : (y, z) \text{ near } (y_0, z_0)\},$$

or

$$\{(x,\phi(x,z),z,):(x,z)\ \text{ near }\ (x_0,z_0)\}.$$

This shows that parametrically defined surfaces can be obtained locally in a particularly simple form.

15. Let $\mathbf{f}:U\to Y$, $D\mathbf{f}(\mathbf{x})$ exists for all $\mathbf{x}\in U$, $B(\mathbf{x}_0,\delta)\subseteq U$, and there exists $L\in\mathcal{L}(X,Y)$, such that $L^{-1}\in\mathcal{L}(Y,X)$, and for all $\mathbf{x}\in B(\mathbf{x}_0,\delta)$

$$||D\mathbf{f}(\mathbf{x})-L||<\frac{r}{||L^{-1}||},\ r<1.$$

Show that there exists $\varepsilon>0$ and an open subset of $B(\mathbf{x}_0,\delta),V$, such that $\mathbf{f}:V\to B(\mathbf{f}(\mathbf{x}_0),\varepsilon)$ is one to one and onto. Also $D\mathbf{f}^{-1}(\mathbf{y})$ exists for each $\mathbf{y}\in B(f(\mathbf{x}_0),\varepsilon)$ and is given by the formula

$$D\mathbf{f}^{-1}(\mathbf{y})=\left[D\mathbf{f}\left(\mathbf{f}^{-1}(\mathbf{y})\right)\right]^{-1}.$$

Hint: Let

$$T_{\mathbf{y}}(\mathbf{x})\equiv T(\mathbf{x},\mathbf{y})\equiv\mathbf{x}-L^{-1}(\mathbf{f}(\mathbf{x})-\mathbf{y})$$

for $|\mathbf{y}-\mathbf{f}(\mathbf{x}_0)|<\frac{(1-r)\delta}{2||L^{-1}||}$, consider $\{T_{\mathbf{y}}^n(\mathbf{x}_0)\}$. This is a version of the inverse function theorem for $\mathbf{f}$ only differentiable, not C^1.

16. Recall the n^{th} derivative can be considered a multilinear function defined on X^n with values in some normed vector space. Now define a function denoted as $\mathbf{w}_i\mathbf{v}_{j_1}\cdots\mathbf{v}_{j_n}$ which maps $X^n\to Y$ in the following way

$$\mathbf{w}_i\mathbf{v}_{j_1}\cdots\mathbf{v}_{j_n}(\mathbf{v}_{k_1},\cdots,\mathbf{v}_{k_n})\equiv\mathbf{w}_i\delta_{j_1k_1}\delta_{j_2k_2}\cdots\delta_{j_nk_n}\tag{6.55}$$

and $\mathbf{w}_i\mathbf{v}_{j_1}\cdots\mathbf{v}_{j_n}$ is to be linear in each variable. Thus, for

$$\left(\sum_{k_1=1}^n a_{k_1}\mathbf{v}_{k_1},\cdots,\sum_{k_n=1}^n a_{k_n}\mathbf{v}_{k_n}\right)\in X^n,$$

$$\begin{aligned}&\mathbf{w}_i\mathbf{v}_{j_1}\cdots\mathbf{v}_{j_n}\left(\sum_{k_1=1}^n a_{k_1}\mathbf{v}_{k_1},\cdots,\sum_{k_n=1}^n a_{k_n}\mathbf{v}_{k_n}\right)\\ \equiv&\sum_{k_1k_2\cdots k_n}\mathbf{w}_i(a_{k_1}a_{k_2}\cdots a_{k_n})\delta_{j_1k_1}\delta_{j_2k_2}\cdots\delta_{j_nk_n}\\ =&\ \mathbf{w}_ia_{j_1}a_{j_2}\cdots a_{j_n}\end{aligned}\tag{6.56}$$

Show each $\mathbf{w}_i\mathbf{v}_{j_1}\cdots\mathbf{v}_{j_n}$ is an n linear Y valued function. Next show the set of n linear Y valued functions is a vector space and these special functions, $\mathbf{w}_i\mathbf{v}_{j_1}\cdots\mathbf{v}_{j_n}$ for all choices of i and the j_k is a basis of this vector space. Find the dimension of the vector space.

17. Minimize $\sum_{j=1}^n x_j$ subject to the constraint $\sum_{j=1}^n x_j^2=a^2$. Your answer should be some function of a which you may assume is a positive number.

18. Find the point, (x,y,z) on the level surface, $4x^2+y^2-z^2=1$which is closest to $(0,0,0)$.

19. A curve is formed from the intersection of the plane, $2x+3y+z=3$ and the cylinder $x^2+y^2=4$. Find the point on this curve which is closest to $(0,0,0)$.

20. A curve is formed from the intersection of the plane, $2x + 3y + z = 3$ and the sphere $x^2 + y^2 + z^2 = 16$. Find the point on this curve which is closest to $(0, 0, 0)$.

21. Find the point on the plane, $2x + 3y + z = 4$ which is closest to the point $(1, 2, 3)$.

22. Let $A = (A_{ij})$ be an $n \times n$ matrix which is symmetric. Thus $A_{ij} = A_{ji}$ and recall $(A\mathbf{x})_i = A_{ij}x_j$ where as usual sum over the repeated index. Show $\frac{\partial}{\partial x_i}(A_{ij}x_jx_i) = 2A_{ij}x_j$. Show that when you use the method of Lagrange multipliers to maximize the function, $A_{ij}x_jx_i$ subject to the constraint, $\sum_{j=1}^n x_j^2 = 1$, the value of λ which corresponds to the maximum value of this functions is such that $A_{ij}x_j = \lambda x_i$. Thus $A\mathbf{x} = \lambda\mathbf{x}$. Thus λ is an eigenvalue of the matrix, A.

23. Let $x_1, \cdots, x_5$ be 5 positive numbers. Maximize their product subject to the constraint that
$$x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 = 300.$$

24. Let $f(x_1, \cdots, x_n) = x_1^n x_2^{n-1} \cdots x_n^1$. Then f achieves a maximum on the set,
$$S \equiv \left\{ \mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^n i x_i = 1 \text{ and each } x_i \geq 0 \right\}.$$
If $\mathbf{x} \in S$ is the point where this maximum is achieved, find x_1/x_n.

25. Let (x, y) be a point on the ellipse, $x^2/a^2 + y^2/b^2 = 1$ which is in the first quadrant. Extend the tangent line through (x, y) till it intersects the x and y axes and let $A(x, y)$ denote the area of the triangle formed by this line and the two coordinate axes. Find the minimum value of the area of this triangle as a function of a and b.

26. Maximize $\prod_{i=1}^n x_i^2$ $(\equiv x_1^2 \times x_2^2 \times x_3^2 \times \cdots \times x_n^2)$ subject to the constraint, $\sum_{i=1}^n x_i^2 = r^2$. Show the maximum is $\left(r^2/n\right)^n$. Now show from this that
$$\left(\prod_{i=1}^n x_i^2 \right)^{1/n} \leq \frac{1}{n} \sum_{i=1}^n x_i^2$$
and finally, conclude that if each number $x_i \geq 0$, then
$$\left(\prod_{i=1}^n x_i \right)^{1/n} \leq \frac{1}{n} \sum_{i=1}^n x_i$$
and there exist values of the x_i for which equality holds. This says the "geometric mean" is always smaller than the arithmetic mean.

27. Maximize x^2y^2 subject to the constraint
$$\frac{x^{2p}}{p} + \frac{y^{2q}}{q} = r^2$$
where p, q are real numbers larger than 1 which have the property that
$$\frac{1}{p} + \frac{1}{q} = 1.$$
show the maximum is achieved when $x^{2p} = y^{2q}$ and equals r^2. Now conclude that if $x, y > 0$, then
$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}$$
and there are values of x and y where this inequality is an equation.

Chapter 7

Measures And Measurable Functions

The integral to be discussed next is the Lebesgue integral. This integral is more general than the Riemann integral of beginning calculus. It is not as easy to define as this integral but is vastly superior in every application. In fact, the Riemann integral has been obsolete for over 100 years. There exist convergence theorems for the Lebesgue integral which are not available for the Riemann integral and unlike the Riemann integral, the Lebesgue integral generalizes readily to abstract settings used in probability theory. Much of the analysis done in the last 100 years applies to the Lebesgue integral. For these reasons, and because it is very easy to generalize the Lebesgue integral to functions of many variables I will present the Lebesgue integral here. First it is convenient to discuss outer measures, measures, and measurable function in a general setting.

7.1 Compact Sets

This is a good place to put an important theorem about compact sets. The definition of what is meant by a compact set follows.

Definition 7.1.1 *Let $\mathcal{U}$ denote a collection of open sets in a normed vector space. Then $\mathcal{U}$ is said to be an open cover of a set K if $K \subseteq \cup\mathcal{U}$. Let K be a subset of a normed vector space. Then K is compact if whenever $\mathcal{U}$ is an open cover of K there exist finitely many sets of $\mathcal{U}$, $\{U_1, \cdots, U_m\}$ such that*

$$K \subseteq \cup_{k=1}^{m} U_k.$$

In words, every open cover admits a finite subcover.

It was shown earlier that in any finite dimensional normed vector space the closed and bounded sets are those which are sequentially compact. The next theorem says that in any normed vector space, sequentially compact and compact are the same.[1] First here is a very interesting lemma about the existence of something called a Lebesgue number, the number r in the next lemma.

Lemma 7.1.2 *Let K be a sequentially compact set in a normed vector space and let $\mathcal{U}$ be an open cover of K. Then there exists $r > 0$ such that if $\mathbf{x} \in K$, then $B(\mathbf{x}, r)$ is a subset of some set of $\mathcal{U}$.*

[1] Actually, this is true more generally than for normed vector spaces. It is also true for metric spaces, those on which there is a distance defined.

Proof: Suppose no such r exists. Then in particular, $1/n$ does not work for each $n \in \mathbb{N}$. Therefore, there exists $\mathbf{x}_n \in K$ such that $B(\mathbf{x}_n, r)$ is not a subset of any of the sets of $\mathcal{U}$. Since K is sequentially compact, there exists a subsequence, $\{\mathbf{x}_{n_k}\}$ converging to a point $\mathbf{x}$ of K. Then there exists $r > 0$ such that $B(\mathbf{x}, r) \subseteq U \in \mathcal{U}$ because $\mathcal{U}$ is an open cover. Also $\mathbf{x}_{n_k} \in B(\mathbf{x}, r/2)$ for all k large enough and also for all k large enough, $1/n_k < r/2$. Therefore, there exists $\mathbf{x}_{n_k} \in B(\mathbf{x}, r/2)$ and $1/n_k < r/2$. But this is a contradiction because

$$B(\mathbf{x}_{n_k}, 1/n_k) \subseteq B(\mathbf{x}, r) \subseteq U$$

contrary to the choice of $\mathbf{x}_{n_k}$ which required $B(\mathbf{x}_{n_k}, 1/n_k)$ is not contained in any set of $\mathcal{U}$. ■

Theorem 7.1.3 *Let K be a set in a normed vector space. Then K is compact if and only if K is sequentially compact. In particular if K is a closed and bounded subset of a finite dimensional normed vector space, then K is compact.*

Proof: Suppose first K is sequentially compact and let $\mathcal{U}$ be an open cover. Let r be a Lebesgue number as described in Lemma 7.1.2. Pick $\mathbf{x}_1 \in K$. Then $B(\mathbf{x}_1, r) \subseteq U_1$ for some $U_1 \in \mathcal{U}$. Suppose $\{B(\mathbf{x}_i, r)\}_{i=1}^m$ have been chosen such that

$$B(\mathbf{x}_i, r) \subseteq U_i \in \mathcal{U}.$$

If their union contains K then $\{U_i\}_{i=1}^m$ is a finite subcover of $\mathcal{U}$. If $\{B(\mathbf{x}_i, r)\}_{i=1}^m$ does not cover K, then there exists $\mathbf{x}_{m+1} \notin \cup_{i=1}^m B(\mathbf{x}_i, r)$ and so $B(\mathbf{x}_{m+1}, r) \subseteq U_{m+1} \in \mathcal{U}$. This process must stop after finitely many choices of $B(\mathbf{x}_i, r)$ because if not, $\{\mathbf{x}_k\}_{k=1}^\infty$ would have a subsequence which converges to a point of K which cannot occur because whenever $i \neq j$,

$$||\mathbf{x}_i - \mathbf{x}_j|| > r$$

Therefore, eventually

$$K \subseteq \cup_{k=1}^m B(\mathbf{x}_k, r) \subseteq \cup_{k=1}^m U_k.$$

this proves one half of the theorem.

Now suppose K is compact. I need to show it is sequentially compact. Suppose it is not. Then there exists a sequence, $\{\mathbf{x}_k\}$ which has no convergent subsequence. This requires that $\{\mathbf{x}_k\}$ have no limit point for if it did have a limit point $\mathbf{x}$, then $B(\mathbf{x}, 1/n)$ would contain infinitely many distinct points of $\{\mathbf{x}_k\}$ and so a subsequence of $\{\mathbf{x}_k\}$ converging to $\mathbf{x}$ could be obtained. Also no $\mathbf{x}_k$ is repeated infinitely often because if there were such, a convergent subsequence could be obtained. Hence $\cup_{k=m}^\infty \{\mathbf{x}_k\} \equiv C_m$ is a closed set, closed because it contains all its limit points. (It has no limit points so it contains them all.) Then letting $U_m = C_m^C$, it follows $\{U_m\}$ is an open cover of K which has no finite subcover. Thus K must be sequentially compact after all.

If K is a closed and bounded set in a finite dimensional normed vector space, then K is sequentially compact by Theorem 5.8.4. Therefore, by the first part of this theorem, it is sequentially compact. This proves the theorem. ■

Summarizing the above theorem along with Theorem 5.8.4 yields the following corollary which is often called the Heine Borel theorem.

Corollary 7.1.4 *Let X be a finite dimensional normed vector space and let $K \subseteq X$. Then the following are equivalent.*

1. *K is closed and bounded.*
2. *K is sequentially compact.*
3. *K is compact.*

7.2 An Outer Measure On $\mathcal{P}(\mathbb{R})$

A measure on $\mathbb{R}$ is like length. I will present something more general because it is no trouble to do so and the generalization is useful in many areas of mathematics such as probability. Recall that $\mathcal{P}(S)$ denotes the set of all subsets of S.

Theorem 7.2.1 *Let F be an increasing function defined on $\mathbb{R}$, an integrator function. There exists a function $\mu : \mathcal{P}(\mathbb{R}) \to [0,\infty]$ which satisfies the following properties.*

1. *If $A \subseteq B$, then $0 \leq \mu(A) \leq \mu(B), \mu(\emptyset) = 0$.*
2. *$\mu(\cup_{k=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$*
3. *$\mu([a,b]) = F(b+) - F(a-)$,*
4. *$\mu((a,b)) = F(b-) - F(a+)$*
5. *$\mu((a,b]) = F(b+) - F(a+)$*
6. *$\mu([a,b)) = F(b-) - F(a-)$ where*

$$F(b+) \equiv \lim_{t \to b+} F(t), F(b-) \equiv \lim_{t \to a-} F(t).$$

Proof: First it is necessary to define the function, μ. This is contained in the following definition.

Definition 7.2.2 *For $A \subseteq \mathbb{R}$,*

$$\mu(A) = \inf\left\{\sum_{j=1}^{\infty} (F(b_i-) - F(a_i+)) : A \subseteq \cup_{i=1}^{\infty}(a_i, b_i)\right\}$$

In words, you look at all coverings of A with open intervals. For each of these open coverings, you add the "lengths" of the individual open intervals and you take the infimum of all such numbers obtained.

Then 1.) is obvious because if a countable collection of open intervals covers B then it also covers A. Thus the set of numbers obtained for B is smaller than the set of numbers for A. Why is $\mu(\emptyset) = 0$? Pick a point of continuity of F. Such points exist because F is increasing and so it has only countably many points of discontinuity. Let a be this point. Then $\emptyset \subseteq (a-\delta, a+\delta)$ and so $\mu(\emptyset) \leq 2\delta$ for every $\delta > 0$.

Consider 2.). If any $\mu(A_i) = \infty$, there is nothing to prove. The assertion simply is $\infty \leq \infty$. Assume then that $\mu(A_i) < \infty$ for all i. Then for each $m \in \mathbb{N}$ there exists a countable set of open intervals, $\{(a_i^m, b_i^m)\}_{i=1}^{\infty}$ such that

$$\mu(A_m) + \frac{\varepsilon}{2^m} > \sum_{i=1}^{\infty} (F(b_i^m-) - F(a_i^m+)).$$

Then using Theorem 2.3.4 on Page 23,

$$\begin{aligned}
\mu(\cup_{m=1}^{\infty} A_m) &\leq \sum_{im} (F(b_i^m-) - F(a_i^m+)) \\
&= \sum_{m=1}^{\infty}\sum_{i=1}^{\infty} (F(b_i^m-) - F(a_i^m+)) \\
&\leq \sum_{m=1}^{\infty} \mu(A_m) + \frac{\varepsilon}{2^m} \\
&= \sum_{m=1}^{\infty} \mu(A_m) + \varepsilon
\end{aligned}$$

and since ε is arbitrary, this establishes 2.).

Next consider 3.). By definition, there exists a sequence of open intervals, $\{(a_i, b_i)\}_{i=1}^{\infty}$ whose union contains $[a, b]$ such that

$$\mu([a,b]) + \varepsilon \geq \sum_{i=1}^{\infty} (F(b_i-) - F(a_i+))$$

By Theorem 7.1.3, finitely many of these intervals also cover $[a, b]$. It follows there exists finitely many of these intervals, $\{(a_i, b_i)\}_{i=1}^{n}$ which overlap such that $a \in (a_1, b_1), b_1 \in (a_2, b_2), \cdots, b \in (a_n, b_n)$. Therefore,

$$\mu([a,b]) \leq \sum_{i=1}^{n} (F(b_i-) - F(a_i+))$$

It follows

$$\begin{aligned} \sum_{i=1}^{n} (F(b_i-) - F(a_i+)) &\geq \mu([a,b]) \\ &\geq \sum_{i=1}^{n} (F(b_i-) - F(a_i+)) - \varepsilon \\ &\geq F(b+) - F(a-) - \varepsilon \end{aligned}$$

Since ε is arbitrary, this shows

$$\mu([a,b]) \geq F(b+) - F(a-)$$

but also, from the definition, the following inequality holds for all $\delta > 0$.

$$\mu([a,b]) \leq F((b+\delta)-) - F((a-\delta)+) \leq F(b+\delta) - F(a-\delta)$$

Therefore, letting $\delta \to 0$ yields

$$\mu([a,b]) \leq F(b+) - F(a-)$$

This establishes 3.).

Consider 4.). For small $\delta > 0$,

$$\mu([a+\delta, b-\delta]) \leq \mu((a,b)) \leq \mu([a,b]).$$

Therefore, from 3.) and the definition of μ,

$$\begin{aligned} F((b-\delta)) - F((a+\delta)) &\leq F((b-\delta)+) - F((a+\delta)-) \\ &= \mu([a+\delta, b-\delta]) \leq \mu((a,b)) \leq F(b-) - F(a+) \end{aligned}$$

Now letting δ decrease to 0 it follows

$$F(b-) - F(a+) \leq \mu((a,b)) \leq F(b-) - F(a+)$$

This shows 4.)

Consider 5.). From 3.) and 4.), for small $\delta > 0$,

$$\begin{aligned} & F(b+) - F((a+\delta)) \\ \leq\ & F(b+) - F((a+\delta)-) \\ =\ & \mu([a+\delta, b]) \leq \mu((a,b]) \\ \leq\ & \mu((a, b+\delta)) = F((b+\delta)-) - F(a+) \\ \leq\ & F(b+\delta) - F(a+). \end{aligned}$$

Now let δ converge to 0 from above to obtain

$$F(b+) - F(a+) = \mu((a,b]) = F(b+) - F(a+).$$

This establishes 5.) and 6.) is entirely similar to 5.). This proves the theorem. ■

Definition 7.2.3 *Let Ω be a nonempty set. A function mapping $\mathcal{P}(\Omega) \to [0,\infty]$ is called an outer measure if it satisfies the conditions 1.) and 2.) in Theorem 7.2.1.*

7.3 General Outer Measures And Measures

First the general concept of a measure will be presented. Then it will be shown how to get a measure from any outer measure. Using the outer measure just obtained, this yields Lebesgue Stieltjes measure on $\mathbb{R}$. Then an abstract Lebesgue integral and its properties will be presented. After this the theory is specialized to the situation of $\mathbb{R}$ and the outer measure in Theorem 7.2.1. This will yield the Lebesgue Stieltjes integral on $\mathbb{R}$ along with spectacular theorems about its properties. The generalization to Lebesgue integration on $\mathbb{R}^n$ turns out to be very easy.

7.3.1 Measures And Measure Spaces

First here is a definition of a measure.

Definition 7.3.1 *$\mathcal{S} \subseteq \mathcal{P}(\Omega)$ is called a σ algebra , pronounced "sigma algebra", if*

$$\emptyset, \Omega \in \mathcal{S},$$

$$\text{If } E \in \mathcal{S} \text{ then } E^C \in \mathcal{S}$$

and

$$\text{If } E_i \in \mathcal{S}, \text{ for } i = 1, 2, \cdots, \text{ then } \cup_{i=1}^{\infty} E_i \in \mathcal{S}.$$

A function $\mu : \mathcal{S} \to [0,\infty]$ where $\mathcal{S}$ is a σ algebra is called a measure if whenever $\{E_i\}_{i=1}^{\infty} \subseteq \mathcal{S}$ and the E_i are disjoint, then it follows

$$\mu\left(\cup_{j=1}^{\infty} E_j\right) = \sum_{j=1}^{\infty} \mu(E_j).$$

The triple $(\Omega, \mathcal{S}, \mu)$ is often called a measure space. Sometimes people refer to $(\Omega, \mathcal{S})$ as a measurable space, making no reference to the measure. Sometimes $(\Omega, \mathcal{S})$ may also be called a measure space.

Theorem 7.3.2 *Let $\{E_m\}_{m=1}^{\infty}$ be a sequence of measurable sets in a measure space $(\Omega, \mathcal{F}, \mu)$. Then if $\cdots E_n \subseteq E_{n+1} \subseteq E_{n+2} \subseteq \cdots$,*

$$\mu(\cup_{i=1}^{\infty} E_i) = \lim_{n\to\infty} \mu(E_n) \tag{7.1}$$

and if $\cdots E_n \supseteq E_{n+1} \supseteq E_{n+2} \supseteq \cdots$ and $\mu(E_1) < \infty$, then

$$\mu(\cap_{i=1}^{\infty} E_i) = \lim_{n\to\infty} \mu(E_n). \tag{7.2}$$

Stated more succinctly, $E_k \uparrow E$ implies $\mu(E_k) \uparrow \mu(E)$ and $E_k \downarrow E$ with $\mu(E_1) < \infty$ implies $\mu(E_k) \downarrow \mu(E)$.

Proof: First note that $\cap_{i=1}^{\infty} E_i = (\cup_{i=1}^{\infty} E_i^C)^C \in \mathcal{F}$ so $\cap_{i=1}^{\infty} E_i$ is measurable. Also note that for A and B sets of $\mathcal{F}$, $A \setminus B \equiv \left(A^C \cup B\right)^C \in \mathcal{F}$. To show 7.1, note that 7.1 is obviously true if $\mu(E_k) = \infty$ for any k. Therefore, assume $\mu(E_k) < \infty$ for all k. Thus

$$\mu(E_{k+1} \setminus E_k) + \mu(E_k) = \mu(E_{k+1})$$

and so

$$\mu(E_{k+1} \setminus E_k) = \mu(E_{k+1}) - \mu(E_k).$$

Also,

$$\bigcup_{k=1}^{\infty} E_k = E_1 \cup \bigcup_{k=1}^{\infty} (E_{k+1} \setminus E_k)$$

and the sets in the above union are disjoint. Hence

$$\mu(\cup_{i=1}^{\infty} E_i) = \mu(E_1) + \sum_{k=1}^{\infty} \mu(E_{k+1} \setminus E_k) = \mu(E_1)$$

$$+ \sum_{k=1}^{\infty} \mu(E_{k+1}) - \mu(E_k)$$

$$= \mu(E_1) + \lim_{n\to\infty} \sum_{k=1}^{n} \mu(E_{k+1}) - \mu(E_k) = \lim_{n\to\infty} \mu(E_{n+1}).$$

This shows part 7.1.

To verify 7.2,

$$\mu(E_1) = \mu(\cap_{i=1}^{\infty} E_i) + \mu(E_1 \setminus \cap_{i=1}^{\infty} E_i)$$

since $\mu(E_1) < \infty$, it follows $\mu(\cap_{i=1}^{\infty} E_i) < \infty$. Also, $E_1 \setminus \cap_{i=1}^{n} E_i \uparrow E_1 \setminus \cap_{i=1}^{\infty} E_i$ and so by 7.1,

$$\mu(E_1) - \mu(\cap_{i=1}^{\infty} E_i) = \mu(E_1 \setminus \cap_{i=1}^{\infty} E_i) = \lim_{n\to\infty} \mu(E_1 \setminus \cap_{i=1}^{n} E_i)$$

$$= \mu(E_1) - \lim_{n\to\infty} \mu(\cap_{i=1}^{n} E_i) = \mu(E_1) - \lim_{n\to\infty} \mu(E_n),$$

Hence, subtracting $\mu(E_1)$ from both sides,

$$\lim_{n\to\infty} \mu(E_n) = \mu(\cap_{i=1}^{\infty} E_i).$$

This proves the theorem. ■

The following definition is important.

Definition 7.3.3 *If something happens except for on a set of measure zero, then it is said to happen a.e. "almost everywhere". For example, $\{f_k(x)\}$ is said to converge to $f(x)$ a.e. if there is a set of measure zero, N such that if $x \in N$, then $f_k(x) \to f(x)$.*

7.4 The Borel Sets, Regular Measures

7.4.1 Definition of Regular Measures

It is important to consider the interaction between measures and open and compact sets. This involves the concept of a regular measure.

Definition 7.4.1 *Let Y be a closed subset of X a finite dimensional normed vector space. The closed sets in Y are the intersections of closed sets in X with Y. The open sets in Y are intersections of open sets of X with Y. Now let $\mathcal{F}$ be a σ algebra of sets of Y and let μ be a measure defined on $\mathcal{F}$. Then μ is said to be a regular measure if the following two conditions hold.*

1. *For every* $F \in \mathcal{F}$

$$\mu(F) = \sup\{\mu(K) : K \subseteq F \text{ and } K \text{ is compact}\} \tag{7.3}$$

2. *For every* $F \in \mathcal{F}$

$$\mu(F) = \inf\{\mu(V) : V \supseteq F \text{ and } V \text{ is open in } Y\} \tag{7.4}$$

The first of the above conditions is called inner regularity and the second is called outer regularity.

Proposition 7.4.2 *In the above situation, a set,* $K \subseteq Y$ *is compact in* Y *if and only if it is compact in* X*.*

Proof: If K is compact in X and $K \subseteq Y$, let $\mathcal{U}$ be an open cover of K of sets open in Y. This means $\mathcal{U} = \{Y \cap V : V \in \mathcal{V}\}$ where $\mathcal{V}$ is an open cover of K consisting of sets open in X. Therefore, $\mathcal{V}$ admits a finite subcover, $\{V_1, \cdots, V_m\}$ and consequently, $\{Y \cap V_1, \cdots, Y \cap V_m\}$ is a finite subcover from $\mathcal{U}$. Thus K is compact in Y.

Now suppose K is compact in Y. This means that if $\mathcal{U}$ is an open cover of sets open in Y it admitts a finite subcover. Now let $\mathcal{V}$ be any open cover of K, consisting of sets open in X. Then $\mathcal{U} \equiv \{V \cap Y : V \in \mathcal{V}\}$ is a cover consisting of sets open in Y and by definition, this admitts a finite subcover, $\{Y \cap V_1, \cdots, Y \cap V_m\}$ but this implies $\{V_1, \cdots, V_m\}$ is also a finite subcover consisting of sets of $\mathcal{V}$. This proves the proposition. ■

7.4.2 The Borel Sets

If Y is a closed subset of X, a normed vector space, denote by $\mathcal{B}(Y)$ the smallest σ algebra of subsets of Y which contains all the open sets of Y. To see such a smallest σ algebra exists, let $\mathfrak{H}$ denote the set of all σ algebras which contain the open sets $\mathcal{P}(Y)$, the set of all subsets of Y is one such σ algebra. Define $\mathcal{B}(Y) \equiv \cap\mathfrak{H}$. Then $\mathcal{B}(Y)$ is a σ algebra because $\emptyset, Y$ are both open sets in Y and so they are in each σ algebra of $\mathfrak{H}$. If $F \in \mathcal{B}(Y)$, then F is a set of every σ algebra of $\mathfrak{H}$ and so F^C is also a set of every σ algebra of $\mathfrak{H}$. Thus $F^C \in \mathcal{B}(Y)$. If $\{F_i\}$ is a sequence of sets of $\mathcal{B}(Y)$, then $\{F_i\}$ is a sequence of sets of every σ algebra of $\mathfrak{H}$ and so $\cup_i F_i$ is a set in every σ algebra of $\mathfrak{H}$ which implies $\cup_i F_i \in \mathcal{B}(Y)$ so $\mathcal{B}(Y)$ is a σ algebra as claimed. From its definition, it is the smallest σ algebra which contains the open sets.

7.4.3 Borel Sets And Regularity

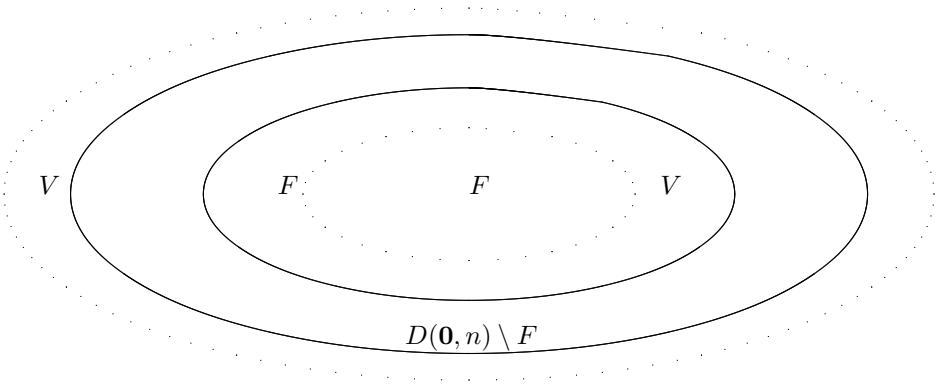
To illustrate how nice the Borel sets are, here are some interesting results about regularity. The first Lemma holds for any σ algebra, not just the Borel sets. Here is some notation which will be used. Let

$$\begin{aligned} S(\mathbf{0},r) &\equiv \{\mathbf{x} \in Y : ||\mathbf{x}|| = r\} \\ D(\mathbf{0},r) &\equiv \{\mathbf{x} \in Y : ||\mathbf{x}|| \leq r\} \\ B(\mathbf{0},r) &\equiv \{\mathbf{x} \in Y : ||\mathbf{x}|| < r\} \end{aligned}$$

Thus $S(\mathbf{0},r)$ is a closed set as is $D(\mathbf{0},r)$ while $B(\mathbf{0},r)$ is an open set. These are closed or open as stated in Y. Since $S(\mathbf{0},r)$ and $D(\mathbf{0},r)$ are intersections of closed sets Y and a closed set in X, these are also closed in X. Of course $B(\mathbf{0},r)$ might not be open in X. This would happen if Y has empty interior in X for example. However, $S(\mathbf{0},r)$ and $D(\mathbf{0},r)$ are compact.

Lemma 7.4.3 *Let Y be a closed subset of X a finite dimensional normed vector space and let $\mathcal{S}$ be a σ algebra of sets of Y containing the open sets of Y. Suppose μ is a measure defined on $\mathcal{S}$ and suppose also $\mu(K) < \infty$ whenever K is compact. Then if 7.4 holds, so does 7.3.*

Proof: It is desired to show that in this setting outer regularity implies inner regularity. First suppose $F \subseteq D(\mathbf{0}, n)$ where $n \in \mathbb{N}$ and $F \in \mathcal{S}$. The following diagram will help to follow the technicalities. In this picture, V is the material between the two dotted curves, F is the inside of the solid curve and $D(\mathbf{0},n)$ is inside the larger solid curve.



The idea is to use outer regularity on $D(\mathbf{0}, n) \setminus F$ to come up with V approximating this set as suggested in the picture. Then $V^C \cap D(\mathbf{0}, n)$ is a compact set contained in F which approximates F. On the picture, the error is represented by the material between the small dotted curve and the smaller solid curve which is less than the error between V and $D(\mathbf{0}, n) \setminus F$ as indicated by the picture. If you need the details, they follow. Otherwise the rest of the proof starts at

Taking complements with respect to Y

$$D(\mathbf{0}, n) \setminus F = D(\mathbf{0}, n) \cap F^C = \left(D(\mathbf{0}, n)^C \cup F \right)^C \in \mathcal{S}$$

because it is given that $\mathcal{S}$ contains the open sets. By 7.4 there exists an open set, $V \supseteq D(\mathbf{0}, n) \setminus F$ such that

$$\mu(D(\mathbf{0}, n) \setminus F) + \varepsilon > \mu(V). \tag{7.5}$$

Since μ is a measure,

$$\mu(V \setminus (D(\mathbf{0}, n) \setminus F)) + \mu(D(\mathbf{0}, n) \setminus F) = \mu(V)$$

and so from 7.5

$$\mu(V \setminus (D(\mathbf{0}, n) \setminus F)) < \varepsilon \tag{7.6}$$

Note

$$V \setminus (D(\mathbf{0}, n) \setminus F) = V \cap \left(D(\mathbf{0}, n) \cap F^C \right)^C = \left(V \cap D(\mathbf{0}, n)^C \right) \cup (V \cap F)$$

and by 7.6,

$$\mu(V \setminus (D(\mathbf{0}, n) \setminus F)) < \varepsilon$$

so in particular,

$$\mu(V \cap F) < \varepsilon.$$

Now

$$V \supseteq D(\mathbf{0}, n) \cap F^C$$

and so

$$V^C \subseteq D(\mathbf{0}, n)^C \cup F$$

which implies

$$V^C \cap D(\mathbf{0}, n) \subseteq F \cap D(\mathbf{0}, n) = F$$

Since $F \subseteq D(\mathbf{0}, n)$,

$$\begin{aligned} \mu\left(F \setminus \left(V^C \cap D(\mathbf{0}, n)\right)\right) &= \mu\left(F \cap \left(V^C \cap D(\mathbf{0}, n)\right)^C\right) \\ &= \mu\left((F \cap V) \cup \left(F \cap D(\mathbf{0}, n)^C\right)\right) \\ &= \mu(F \cap V) < \varepsilon \end{aligned}$$

showing the compact set, $V^C \cap D(\mathbf{0}, n)$ is contained in F and

$$\mu\left(V^C \cap D(\mathbf{0}, n)\right) + \varepsilon > \mu(F).$$

In the general case where F is only given to be in $\mathcal{S}$, let $F_n = B(\mathbf{0}, n) \cap F$. Then by 7.1, if $l < \mu(F)$ is given, then for all ε sufficiently small,

$$l + \varepsilon < \mu(F_n)$$

provided n is large enough. Now it was just shown there exists K a compact subset of F_n such that $\mu(F_n) < \mu(K) + \varepsilon$. Then $K \subseteq F$ and

$$l + \varepsilon < \mu(F_n) < \mu(K) + \varepsilon$$

and so whenever $l < \mu(F)$, it follows there exists K a compact subset of F such that

$$l < \mu(K)$$

and This proves the lemma. ■

The following is a useful result which will be used in what follows.

Lemma 7.4.4 *Let X be a normed vector space and let S be any nonempty subset of X. Define*

$$\operatorname{dist}(\mathbf{x}, S) \equiv \inf\{||\mathbf{x} - \mathbf{y}|| : \mathbf{y} \in S\}$$

Then

$$|\operatorname{dist}(\mathbf{x}_1, S) - \operatorname{dist}(\mathbf{x}_2, S)| \leq ||\mathbf{x}_1 - \mathbf{x}_2||.$$

Proof: Suppose $\operatorname{dist}(\mathbf{x}_1, S) \geq \operatorname{dist}(\mathbf{x}_2, S)$. Then let $\mathbf{y} \in S$ such that

$$\operatorname{dist}(\mathbf{x}_2, S) + \varepsilon > ||\mathbf{x}_2 - \mathbf{y}||$$

Then

$$\begin{aligned} |\operatorname{dist}(\mathbf{x}_1, S) - \operatorname{dist}(\mathbf{x}_2, S)| &= \operatorname{dist}(\mathbf{x}_1, S) - \operatorname{dist}(\mathbf{x}_2, S) \\ &\leq \operatorname{dist}(\mathbf{x}_1, S) - (||\mathbf{x}_2 - \mathbf{y}|| - \varepsilon) \\ &\leq ||\mathbf{x}_1 - \mathbf{y}|| - ||\mathbf{x}_2 - \mathbf{y}|| + \varepsilon \\ &\leq |||\mathbf{x}_1 - \mathbf{y}|| - ||\mathbf{x}_2 - \mathbf{y}||| + \varepsilon \\ &\leq ||\mathbf{x}_1 - \mathbf{x}_2|| + \varepsilon. \end{aligned}$$

Since ε is arbitrary, this proves the lemma in case $\text{dist}(\mathbf{x}_1, S) \geq \text{dist}(\mathbf{x}_2, S)$. The case where $\text{dist}(\mathbf{x}_2, S) \geq \text{dist}(\mathbf{x}_1, S)$ is entirely similar. This proves the lemma. ■

The next lemma says that regularity comes free for finite measures defined on the Borel sets. Actually, it only almost says this. The following theorem will say it. This lemma deals with closed in place of compact.

Lemma 7.4.5 *Let μ be a finite measure defined on $\mathcal{B}(Y)$ where Y is a closed subset of X, a finite dimensional normed vector space. Then for every $F \in \mathcal{B}(Y)$,*

$$\mu(F) = \sup\{\mu(K) : K \subseteq F,\ K \text{ is closed }\}$$

$$\mu(F) = \inf\{\mu(V) : V \supseteq F,\ V \text{ is open}\}$$

Proof: For convenience, I will call a measure which satisfies the above two conditions "almost regular". It would be regular if closed were replaced with compact. First note every open set is the countable union of closed sets and every closed set is the countable intersection of open sets. Here is why. Let V be an open set and let

$$K_k \equiv \left\{\mathbf{x} \in V : \text{dist}\left(\mathbf{x}, V^C\right) \geq 1/k\right\}.$$

Then clearly the union of the K_k equals V and each is closed because $x \to \text{dist}(\mathbf{x}, S)$ is always a continuous function whenever S is any nonempty set. Next, for K closed let

$$V_k \equiv \{\mathbf{x} \in Y : \text{dist}(\mathbf{x}, K) < 1/k\}.$$

Clearly the intersection of the V_k equals K because if $\mathbf{x} \notin K$, then since K is closed, $B(\mathbf{x}, r)$ has empty intersection with K and so for k large enough that $1/k < r$, V_k excludes $\mathbf{x}$. Thus the only points in the intersection of the V_k are those in K and in addition each point of K is in this intersection.

Therefore from what was just shown, letting V denote an open set and K a closed set, it follows from Theorem 7.3.2 that

$$\begin{aligned} \mu(V) &= \sup\{\mu(K) : K \subseteq V \text{ and } K \text{ is closed}\} \\ \mu(K) &= \inf\{\mu(V) : V \supseteq K \text{ and } V \text{ is open}\}. \end{aligned}$$

Also since V is open and K is closed,

$$\begin{aligned} \mu(V) &= \inf\{\mu(U) : U \supseteq V \text{ and } V \text{ is open}\} \\ \mu(K) &= \sup\{\mu(L) : L \subseteq K \text{ and } L \text{ is closed}\} \end{aligned}$$

In words, μ is almost regular on open and closed sets. Let

$$\mathcal{F} \equiv \{F \in \mathcal{B}(Y) \text{ such that } \mu \text{ is almost regular on } F\}.$$

Then $\mathcal{F}$ contains the open sets. I want to show $\mathcal{F}$ is a σ algebra and then it will follow $\mathcal{F} = \mathcal{B}(Y)$.

First I will show $\mathcal{F}$ is closed with respect to complements. Let $F \in \mathcal{F}$. Then since μ is finite and F is inner regular, there exists $K \subseteq F$ such that

$$\mu(F \setminus K) = \mu(F) - \mu(K) < \varepsilon.$$

But $K^C \setminus F^C = F \setminus K$ and so

$$\mu\left(K^C \setminus F^C\right) = \mu\left(K^C\right) - \mu\left(F^C\right) < \varepsilon$$

showing that μ is outer regular on F^C. I have just approximated the measure of F^C with the measure of K^C, an open set containing F^C. A similar argument works to show F^C

is inner regular. You start with $V \supseteq F$ such that $\mu(V \setminus F) < \varepsilon$, note $F^C \setminus V^C = V \setminus F$, and then conclude $\mu\left(F^C \setminus V^C\right) < \varepsilon$, thus approximating F^C with the closed subset, V^C.

Next I will show $\mathcal{F}$ is closed with respect to taking countable unions. Let $\{F_k\}$ be a sequence of sets in $\mathcal{F}$. Then since $F_k \in \mathcal{F}$, there exist $\{K_k\}$ such that $K_k \subseteq F_k$ and $\mu(F_k \setminus K_k) < \varepsilon/2^{k+1}$. First choose m large enough that

$$\mu\left(\left(\cup_{k=1}^{\infty} F_k\right) \setminus \left(\cup_{k=1}^{m} F_k\right)\right) < \frac{\varepsilon}{2}.$$

Then

$$\begin{aligned} \mu\left(\left(\cup_{k=1}^{m} F_k\right) \setminus \left(\cup_{k=1}^{m} K_k\right)\right) &\leq \mu\left(\cup_{k=1}^{m} \left(F_k \setminus K_k\right)\right) \\ &\leq \sum_{k=1}^{m} \frac{\varepsilon}{2^{k+1}} < \frac{\varepsilon}{2} \end{aligned}$$

and so

$$\begin{aligned} \mu\left(\left(\cup_{k=1}^{\infty} F_k\right) \setminus \left(\cup_{k=1}^{m} K_k\right)\right) &\leq \mu\left(\left(\cup_{k=1}^{\infty} F_k\right) \setminus \left(\cup_{k=1}^{m} F_k\right)\right) \\ &\quad + \mu\left(\left(\cup_{k=1}^{m} F_k\right) \setminus \left(\cup_{k=1}^{m} K_k\right)\right) \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

Since μ is outer regular on F_k, there exists V_k such that $\mu(V_k \setminus F_k) < \varepsilon/2^k$. Then

$$\begin{aligned} \mu\left(\left(\cup_{k=1}^{\infty} V_k\right) \setminus \left(\cup_{k=1}^{\infty} F_k\right)\right) &\leq \mu\left(\cup_{k=1}^{\infty} \left(V_k \setminus F_k\right)\right) \\ &\leq \sum_{k=1}^{\infty} \mu\left(V_k \setminus F_k\right) \\ &< \sum_{k=1}^{\infty} \frac{\varepsilon}{2^k} = \varepsilon \end{aligned}$$

and this completes the demonstration that $\mathcal{F}$ is a σ algebra. This proves the lemma. ■

The next theorem is the main result. It shows regularity is automatic if $\mu(K) < \infty$ for all compact K.

Theorem 7.4.6 *Let μ be a finite measure defined on $\mathcal{B}(Y)$ where Y is a closed subset of X, a finite dimensional normed vector space. Then μ is regular. If μ is not necessarily finite but is finite on compact sets, then μ is regular.*

Proof: From Lemma 7.4.5 μ is outer regular. Now let $F \in \mathcal{B}(Y)$. Then since μ is finite, it follows from Lemma 7.4.5 there exists $H \subseteq F$ such that H is closed, $H \subseteq F$, and

$$\mu(F) < \mu(H) + \varepsilon.$$

Then let $K_k \equiv H \cap \overline{B(\mathbf{0}, k)}$. Thus K_k is a closed and bounded, hence compact set and $\cup_{k=1}^{\infty} K_k = H$. Therefore by Theorem 7.3.2, for all k large enough,

$$\begin{aligned} & \mu(F) \\ < \; & \mu(K_k) + \varepsilon \\ < \; & \sup\{\mu(K) : K \subseteq F \text{ and } K \text{ compact}\} + \varepsilon \\ \leq \; & \mu(F) + \varepsilon \end{aligned}$$

Since ε was arbitrary, it follows

$$\sup\{\mu(K) : K \subseteq F \text{ and } K \text{ compact}\} = \mu(F).$$

This establishes μ is regular if μ is finite.

Now suppose it is only known that μ is finite on compact sets. Consider outer regularity. There are at most finitely many $r \in [0,R]$ such that $\mu(S(\mathbf{0},r)) > \delta > 0$. If this were not so, then $\mu(D(\mathbf{0},R)) = \infty$ contrary to the assumption that μ is finite on compact sets. Therefore, there are at most countably many $r \in [0,R]$ such that $\mu(S(\mathbf{0},r)) > 0$. Here is why. Let S_k denote those values of $r \in [0,R]$ such that $\mu(S(\mathbf{0},r)) > 1/k$. Then the values of r such that $\mu(S(\mathbf{0},r)) > 0$ equals $\cup_{m=1}^{\infty} S_m$, a countable union of finite sets which is at most countable.

It follows there are at most countably many $r \in (0,\infty)$ such that $\mu(S(\mathbf{0},r)) > 0$. Therefore, there exists an increasing sequence $\{r_k\}$ such that $\lim_{k\to\infty} r_k = \infty$ and $\mu(S(\mathbf{0},r_k)) = 0$. This is easy to see by noting that $(n,n+1]$ contains uncountably many points and so it contains at least one r such that $\mu(S(\mathbf{0},r)) = 0$.

$$S(\mathbf{0},r) = \cap_{k=1}^{\infty} (B(\mathbf{0},r+1/k) - D(\mathbf{0},r-1/k))$$

a countable intersection of open sets which are decreasing as $k \to \infty$. Since $\mu(B(\mathbf{0},r)) < \infty$ by assumption, it follows from Theorem 7.3.2 that for each r_k there exists an open set, $U_k \supseteq S(\mathbf{0},r_k)$ such that

$$\mu(U_k) < \varepsilon/2^{k+1}.$$

Let $\mu(F) < \infty$. There is nothing to show if $\mu(F) = \infty$. Define finite measures, μ_k as follows.

$$\begin{aligned} \mu_1(A) &\equiv \mu(B(\mathbf{0},1) \cap A), \\ \mu_2(A) &\equiv \mu((B(\mathbf{0},2) \setminus D(\mathbf{0},1)) \cap A), \\ \mu_3(A) &\equiv \mu((B(\mathbf{0},3) \setminus D(\mathbf{0},2)) \cap A) \end{aligned}$$

etc. Thus

$$\mu(A) = \sum_{k=1}^{\infty} \mu_k(A)$$

and each μ_k is a finite measure. By the first part there exists an open set V_k such that

$$V_k \supseteq F \cap (B(\mathbf{0},k) \setminus D(\mathbf{0},k-1))$$

and

$$\mu_k(V_k) < \mu_k(F) + \varepsilon/2^{k+1}$$

Without loss of generality $V_k \subseteq (B(\mathbf{0},k) \setminus D(\mathbf{0},k-1))$ since you can take the intersection of V_k with this open set. Thus

$$\mu_k(V_k) = \mu((B(\mathbf{0},k) \setminus D(\mathbf{0},k-1)) \cap V_k) = \mu(V_k)$$

and the V_k are disjoint. Then let $V = \cup_{k=1}^{\infty} V_k$ and $U = \cup_{k=1}^{\infty} U_k$. It follows $V \cup U$ is an open set containing F and

$$\begin{aligned} \mu(F) &= \sum_{k=1}^{\infty} \mu_k(F) > \sum_{k=1}^{\infty} \mu_k(V_k) - \frac{\varepsilon}{2^{k+1}} = \sum_{k=1}^{\infty} \mu(V_k) - \frac{\varepsilon}{2} \\ &= \mu(V) - \frac{\varepsilon}{2} \geq \mu(V) + \mu(U) - \frac{\varepsilon}{2} - \frac{\varepsilon}{2} \geq \mu(V \cup U) - \varepsilon \end{aligned}$$

which shows μ is outer regular. Inner regularity can be obtained from Lemma 7.4.3. Alternatively, you can use the above construction to get it right away. It is easier than the outer regularity.

First assume $\mu(F) < \infty$. By the first part, there exists a compact set,

$$K_k \subseteq F \cap (B(\mathbf{0},k) \setminus D(\mathbf{0},k-1))$$

such that

$$\begin{aligned}\mu_k(K_k)+\varepsilon/2^{k+1} &> \mu_k(F\cap(B(\mathbf{0},k)\setminus D(\mathbf{0},k-1)))\\ &= \mu_k(F)=\mu(F\cap(B(\mathbf{0},k)\setminus D(\mathbf{0},k-1)))\,.\end{aligned}$$

Since K_k is a subset of $F\cap(B(\mathbf{0},k)\setminus D(\mathbf{0},k-1))$ it follows $\mu_k(K_k)=\mu(K_k)$. Therefore,

$$\begin{aligned}\mu(F) &= \sum_{k=1}^{\infty}\mu_k(F)<\sum_{k=1}^{\infty}\mu_k(K_k)+\varepsilon/2^k\\ &< \left(\sum_{k=1}^{\infty}\mu_k(K_k)\right)+\varepsilon/2<\sum_{k=1}^{N}\mu(K_k)+\varepsilon\end{aligned}$$

provided N is large enough. The K_k are disjoint and so letting $K=\cup_{k=1}^{N}K_k$, this says $K\subseteq F$ and

$$\mu(F)<\mu(K)+\varepsilon.$$

Now consider the case where $\mu(F)=\infty$. If $l<\infty$, it follows from Theorem 7.3.2

$$\mu(F\cap B(\mathbf{0},m))>l$$

whenever m is large enough. Therefore, letting $\mu_m(A)\equiv\mu(A\cap B(\mathbf{0},m))$, there exists a compact set, $K\subseteq F\cap B(\mathbf{0},m)$ such that

$$\mu(K)=\mu_m(K)>\mu_m(F\cap B(\mathbf{0},m))=\mu(F\cap B(\mathbf{0},m))>l$$

This proves the theorem. ■

7.5 Measures And Outer Measures

7.5.1 Measures From Outer Measures

Earlier an outer measure on $\mathcal{P}(\mathbb{R})$ was constructed. This can be used to obtain a measure defined on $\mathbb{R}$. However, the procedure for doing so is a special case of a general approach due to Caratheodory in about 1918.

Definition 7.5.1 *Let Ω be a nonempty set and let $\mu:\mathcal{P}(\Omega)\to[0,\infty]$ be an outer measure. For $E\subseteq\Omega$, E is μ measurable if for all $S\subseteq\Omega$,*

$$\mu(S)=\mu(S\setminus E)+\mu(S\cap E). \tag{7.7}$$

To help in remembering 7.7, think of a measurable set, E, as a process which divides a given set into two pieces, the part in E and the part not in E as in 7.7. In the Bible, there are several incidents recorded in which a process of division resulted in more stuff than was originally present.[2] Measurable sets are exactly those which are incapable of such a miracle. You might think of the measurable sets as the nonmiraculous sets. The idea is to show that they form a σ algebra on which the outer measure, μ is a measure.

First here is a definition and a lemma.

[2] 1 Kings 17, 2 Kings 4, Mathew 14, and Mathew 15 all contain such descriptions. The stuff involved was either oil, bread, flour or fish. In mathematics such things have also been done with sets. In the book by Bruckner Bruckner and Thompson there is an interesting discussion of the Banach Tarski paradox which says it is possible to divide a ball in $\mathbb{R}^3$ into five disjoint pieces and assemble the pieces to form two disjoint balls of the same size as the first. The details can be found in: The Banach Tarski Paradox by Wagon, Cambridge University press. 1985. It is known that all such examples must involve the axiom of choice.

Definition 7.5.2 *$(\mu\lfloor S)(A) \equiv \mu(S \cap A)$ for all $A \subseteq \Omega$. Thus $\mu\lfloor S$ is the name of a new outer measure, called μ restricted to S.*

The next lemma indicates that the property of measurability is not lost by considering this restricted measure.

Lemma 7.5.3 *If A is μ measurable, then A is $\mu\lfloor S$ measurable.*

Proof: Suppose A is μ measurable. It is desired to to show that for all $T \subseteq \Omega$,

$$(\mu\lfloor S)(T) = (\mu\lfloor S)(T \cap A) + (\mu\lfloor S)(T \setminus A).$$

Thus it is desired to show

$$\mu(S \cap T) = \mu(T \cap A \cap S) + \mu(T \cap S \cap A^C). \tag{7.8}$$

But 7.8 holds because A is μ measurable. Apply Definition 7.5.1 to $S \cap T$ instead of S. ■

If A is $\mu\lfloor S$ measurable, it does not follow that A is μ measurable. Indeed, if you believe in the existence of non measurable sets, you could let $A = S$ for such a μ non measurable set and verify that S is $\mu\lfloor S$ measurable. In fact there do exist nonmeasurable sets but this is a topic for a more advanced course in analysis and will not be needed in this book.

The next theorem is the main result on outer measures which shows that starting with an outer measure you can obtain a measure.

Theorem 7.5.4 *Let Ω be a set and let μ be an outer measure on $\mathcal{P}(\Omega)$. The collection of μ measurable sets $\mathcal{S}$, forms a σ algebra and*

$$\text{If } F_i \in \mathcal{S},\ F_i \cap F_j = \emptyset,\ \text{then } \mu(\cup_{i=1}^{\infty} F_i) = \sum_{i=1}^{\infty} \mu(F_i). \tag{7.9}$$

If $\cdots F_n \subseteq F_{n+1} \subseteq \cdots$, then if $F = \cup_{n=1}^{\infty} F_n$ and $F_n \in \mathcal{S}$, it follows that

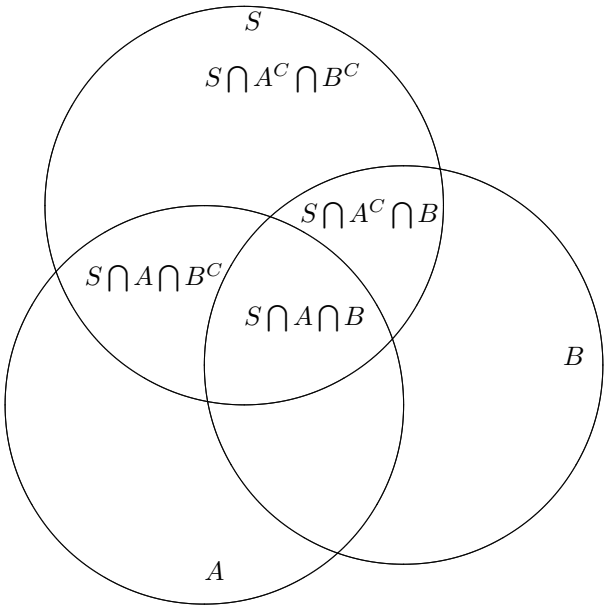
$$\mu(F) = \lim_{n\to\infty} \mu(F_n). \tag{7.10}$$

If $\cdots F_n \supseteq F_{n+1} \supseteq \cdots$, and if $F = \cap_{n=1}^{\infty} F_n$ for $F_n \in \mathcal{S}$ then if $\mu(F_1) < \infty$,

$$\mu(F) = \lim_{n\to\infty} \mu(F_n). \tag{7.11}$$

This measure space is also complete which means that if $\mu(F) = 0$ for some $F \in \mathcal{S}$ then if $G \subseteq F$, it follows $G \in \mathcal{S}$ also.

Proof: First note that $\emptyset$ and Ω are obviously in $\mathcal{S}$. Now suppose $A, B \in \mathcal{S}$. I will show $A \setminus B \equiv A \cap B^C$ is in $\mathcal{S}$. To do so, consider the following picture.



Since μ is subadditive,

$$\mu(S) \leq \mu\left(S \cap A \cap B^C\right) + \mu(A \cap B \cap S) + \mu\left(S \cap B \cap A^C\right) + \mu\left(S \cap A^C \cap B^C\right).$$

Now using $A, B \in \mathcal{S}$,

$$\begin{aligned} \mu(S) &\leq \mu\left(S \cap A \cap B^C\right) + \mu(S \cap A \cap B) + \mu\left(S \cap B \cap A^C\right) + \mu\left(S \cap A^C \cap B^C\right) \\ &= \mu(S \cap A) + \mu\left(S \cap A^C\right) = \mu(S) \end{aligned}$$

It follows equality holds in the above. Now observe, using the picture if you like, that

$$(A \cap B \cap S) \cup \left(S \cap B \cap A^C\right) \cup \left(S \cap A^C \cap B^C\right) = S \setminus (A \setminus B)$$

and therefore,

$$\begin{aligned} \mu(S) &= \mu\left(S \cap A \cap B^C\right) + \mu(A \cap B \cap S) + \mu\left(S \cap B \cap A^C\right) + \mu\left(S \cap A^C \cap B^C\right) \\ &\geq \mu(S \cap (A \setminus B)) + \mu(S \setminus (A \setminus B)). \end{aligned}$$

Therefore, since S is arbitrary, this shows $A \setminus B \in \mathcal{S}$.

Since $\Omega \in \mathcal{S}$, this shows that $A \in \mathcal{S}$ if and only if $A^C \in \mathcal{S}$. Now if $A, B \in \mathcal{S}$, $A \cup B = (A^C \cap B^C)^C = (A^C \setminus B)^C \in \mathcal{S}$. By induction, if $A_1, \cdots, A_n \in \mathcal{S}$, then so is $\cup_{i=1}^n A_i$. If $A, B \in \mathcal{S}$, with $A \cap B = \emptyset$,

$$\mu(A \cup B) = \mu((A \cup B) \cap A) + \mu((A \cup B) \setminus A) = \mu(A) + \mu(B).$$

By induction, if $A_i \cap A_j = \emptyset$ and $A_i \in \mathcal{S}$,

$$\mu(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i). \tag{7.12}$$

Now let $A = \cup_{i=1}^\infty A_i$ where $A_i \cap A_j = \emptyset$ for $i \neq j$.

$$\sum_{i=1}^\infty \mu(A_i) \geq \mu(A) \geq \mu(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i).$$

Since this holds for all n, you can take the limit as $n \to \infty$ and conclude,

$$\sum_{i=1}^{\infty} \mu(A_i) = \mu(A)$$

which establishes 7.9.

Consider part 7.10. Without loss of generality $\mu(F_k) < \infty$ for all k since otherwise there is nothing to show. Suppose $\{F_k\}$ is an increasing sequence of sets of $\mathcal{S}$. Then letting $F_0 \equiv \emptyset$, $\{F_{k+1} \setminus F_k\}_{k=0}^{\infty}$ is a sequence of disjoint sets of $\mathcal{S}$ since it was shown above that the difference of two sets of $\mathcal{S}$ is in $\mathcal{S}$. Also note that from 7.12

$$\mu(F_{k+1} \setminus F_k) + \mu(F_k) = \mu(F_{k+1})$$

and so if $\mu(F_k) < \infty$, then

$$\mu(F_{k+1} \setminus F_k) = \mu(F_{k+1}) - \mu(F_k).$$

Therefore, letting

$$F \equiv \cup_{k=1}^{\infty} F_k$$

which also equals

$$\cup_{k=1}^{\infty} (F_{k+1} \setminus F_k),$$

it follows from part 7.9 just shown that

$$\begin{aligned} \mu(F) &= \sum_{k=1}^{\infty} \mu(F_{k+1} \setminus F_k) = \lim_{n\to\infty} \sum_{k=1}^{n} \mu(F_{k+1} \setminus F_k) \\ &= \lim_{n\to\infty} \sum_{k=1}^{n} \mu(F_{k+1}) - \mu(F_k) = \lim_{n\to\infty} \mu(F_{n+1}). \end{aligned}$$

In order to establish 7.11, let the F_n be as given there. Then, since $(F_1 \setminus F_n)$ increases to $(F_1 \setminus F)$, 7.10 implies

$$\lim_{n\to\infty} (\mu(F_1) - \mu(F_n)) = \mu(F_1 \setminus F).$$

Now $\mu(F_1 \setminus F) + \mu(F) \geq \mu(F_1)$ and so $\mu(F_1 \setminus F) \geq \mu(F_1) - \mu(F)$. Hence

$$\lim_{n\to\infty} (\mu(F_1) - \mu(F_n)) = \mu(F_1 \setminus F) \geq \mu(F_1) - \mu(F)$$

which implies

$$\lim_{n\to\infty} \mu(F_n) \leq \mu(F).$$

But since $F \subseteq F_n$,

$$\mu(F) \leq \lim_{n\to\infty} \mu(F_n)$$

and this establishes 7.11. Note that it was assumed $\mu(F_1) < \infty$ because $\mu(F_1)$ was subtracted from both sides.

It remains to show $\mathcal{S}$ is closed under countable unions. Recall that if $A \in \mathcal{S}$, then $A^C \in \mathcal{S}$ and $\mathcal{S}$ is closed under finite unions. Let $A_i \in \mathcal{S}$, $A = \cup_{i=1}^{\infty} A_i$, $B_n = \cup_{i=1}^{n} A_i$. Then

$$\begin{aligned} \mu(S) &= \mu(S \cap B_n) + \mu(S \setminus B_n) \\ &= (\mu \lfloor S)(B_n) + (\mu \lfloor S)(B_n^C). \end{aligned} \tag{7.13}$$

By Lemma 7.5.3 B_n is $(\mu\lfloor S)$ measurable and so is B_n^C. I want to show $\mu(S) \geq \mu(S \setminus A) + \mu(S \cap A)$. If $\mu(S) = \infty$, there is nothing to prove. Assume $\mu(S) < \infty$. Then apply Parts 7.11 and 7.10 to the outer measure, $\mu\lfloor S$ in 7.13 and let $n \to \infty$. Thus

$$B_n \uparrow A, \ B_n^C \downarrow A^C$$

and this yields $\mu(S) = (\mu\lfloor S)(A) + (\mu\lfloor S)(A^C) = \mu(S \cap A) + \mu(S \setminus A)$.

Therefore $A \in \mathcal{S}$ and this proves Parts 7.9, 7.10, and 7.11.

It only remains to verify the assertion about completeness. Letting G and F be as described above, let $S \subseteq \Omega$. I need to verify

$$\mu(S) \geq \mu(S \cap G) + \mu(S \setminus G)$$

However,

$$\begin{aligned} \mu(S \cap G) + \mu(S \setminus G) &\leq \mu(S \cap F) + \mu(S \setminus F) + \mu(F \setminus G) \\ &= \mu(S \cap F) + \mu(S \setminus F) = \mu(S) \end{aligned}$$

because by assumption, $\mu(F \setminus G) \leq \mu(F) = 0$.This proves the theorem. ■

7.5.2 Completion Of Measure Spaces

Suppose $(\Omega, \mathcal{F}, \mu)$ is a measure space. Then it is always possible to enlarge the σ algebra and define a new measure $\overline{\mu}$ on this larger σ algebra such that $\left(\Omega, \overline{\mathcal{F}}, \overline{\mu}\right)$ is a complete measure space. Recall this means that if $N \subseteq N' \in \overline{\mathcal{F}}$ and $\overline{\mu}(N') = 0$, then $N \in \overline{\mathcal{F}}$. The following theorem is the main result. The new measure space is called the completion of the measure space.

Definition 7.5.5 *A measure space, $(\Omega, \mathcal{F}, \mu)$ is called σ finite if there exists a sequence $\{\Omega_n\} \subseteq \mathcal{F}$ such that $\cup_n \Omega_n = \Omega$ and $\mu(\Omega_n) < \infty$.*

For example, if X is a finite dimensional normed vector space and μ is a measure defined on $\mathcal{B}(X)$ which is finite on compact sets, then you could take $\Omega_n = B(\mathbf{0}, n)$.

Theorem 7.5.6 *Let $(\Omega, \mathcal{F}, \mu)$ be a σ finite measure space. Then there exists a unique measure space, $\left(\Omega, \overline{\mathcal{F}}, \overline{\mu}\right)$ satisfying*

1. *$\left(\Omega, \overline{\mathcal{F}}, \overline{\mu}\right)$ is a complete measure space.*
2. *$\overline{\mu} = \mu$ on $\mathcal{F}$*
3. *$\overline{\mathcal{F}} \supseteq \mathcal{F}$*
4. *For every $E \in \overline{\mathcal{F}}$ there exists $G \in \mathcal{F}$ such that $G \supseteq E$ and $\mu(G) = \overline{\mu}(E)$.*

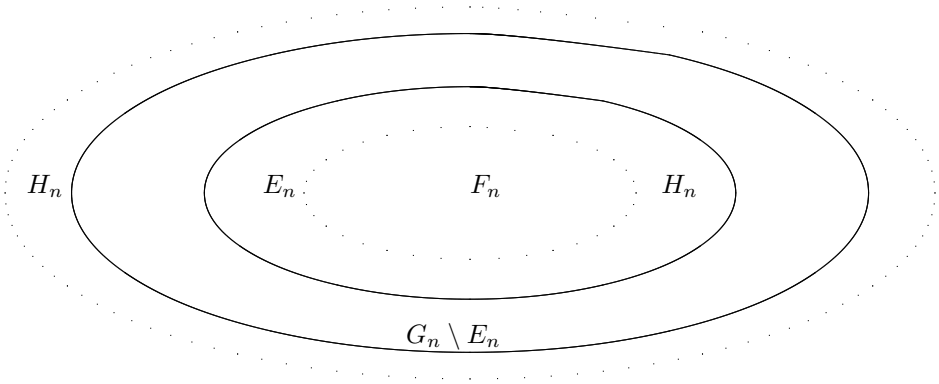
   In addition to this,

5. *For every $E \in \overline{\mathcal{F}}$ there exists $F \in \mathcal{F}$ such that $F \subseteq E$ and $\mu(F) = \overline{\mu}(E)$.*

Also for every $E \in \overline{\mathcal{F}}$ there exist sets $G, F \in \mathcal{F}$ such that $G \supseteq E \supseteq F$ and

$$\mu(G \setminus F) = \overline{\mu}(G \setminus F) = 0 \tag{7.14}$$

Proof: First consider the claim about uniqueness. Suppose $(\Omega, \mathcal{F}_1, \nu_1)$ and $(\Omega, \mathcal{F}_2, \nu_2)$ both satisfy 1.) - 4.) and let $E \in \mathcal{F}_1$. Also let $\mu(\Omega_n) < \infty$, $\cdots \Omega_n \subseteq \Omega_{n+1} \cdots$, and $\cup_{n=1}^{\infty} \Omega_n = \Omega$. Define $E_n \equiv E \cap \Omega_n$. Then there exists $G_n \supseteq E_n$ such that $\mu(G_n) = \nu_1(E_n), G_n \in \mathcal{F}$ and $G_n \subseteq \Omega_n$. I claim there exists $F_n \in \mathcal{F}$ such that $G_n \supseteq E_n \supseteq F_n$ and $\mu(G_n \setminus F_n) = 0$. To see this, look at the following diagram.



In this diagram, there exists $H_n \in \mathcal{F}$ containing $G_n \setminus E_n$, represented in the picture as the set between the dotted lines, such that $\mu(H_n) = \overline{\mu}(G_n \setminus E_n)$. Then define $F_n \equiv H_n^C \cap G_n$. This set is in $\mathcal{F}$, is contained in E_n and as shown in the diagram,

$$\overline{\mu}(E_n) - \mu(F_n) \leq \mu(H_n) - \overline{\mu}(G_n \setminus E_n) = 0.$$

Therefore, since $\overline{\mu}$ is a measure,

$$\begin{aligned} \mu(G_n \setminus F_n) &= \overline{\mu}(G_n \setminus E_n) + \overline{\mu}(E_n \setminus F_n) \\ &= \mu(G_n) - \overline{\mu}(E_n) + \overline{\mu}(E_n) - \mu(F_n) = 0 \end{aligned}$$

Then letting $G = \cup_n G_n, F \equiv \cup_n F_n$, it follows $G \supseteq E \supseteq F$ and

$$\begin{aligned} \mu(G \setminus F) &\leq \mu(\cup_n (G_n \setminus F_n)) \\ &\leq \sum_n \mu(G_n \setminus F_n) = 0. \end{aligned}$$

Thus $\nu_i(G \setminus F) = 0$ for $i = 1, 2$. Now $E \setminus F \subseteq G \setminus F$ and since $(\Omega, \mathcal{F}_2, \nu_2)$ is complete, it follows $E \setminus F \in \mathcal{F}_2$. Since $F \in \mathcal{F}_2$, it follows $E = (E \setminus F) \cup F \in \mathcal{F}_2$. Thus $\mathcal{F}_1 \subseteq \mathcal{F}_2$. Similarly $\mathcal{F}_2 \subseteq \mathcal{F}_1$.

Now it only remains to verify $\nu_1 = \nu_2$. Thus let $E \in \mathcal{F}_1 = \mathcal{F}_2$ and let G and F be as just described. Since $\nu_i = \mu$ on $\mathcal{F}$,

$$\begin{aligned} \mu(F) &\leq \nu_1(E) \\ &= \nu_1(E \setminus F) + \nu_1(F) \\ &\leq \nu_1(G \setminus F) + \nu_1(F) \\ &= \nu_1(F) = \mu(F) \end{aligned}$$

Similarly $\nu_2(E) = \mu(F)$. This proves uniqueness. The construction has also verified 7.14.

Next define an outer measure, $\overline{\mu}$ on $\mathcal{P}(\Omega)$ as follows. For $S \subseteq \Omega$,

$$\overline{\mu}(S) \equiv \inf \{\mu(E) : E \in \mathcal{F}\}.$$

Then it is clear $\overline{\mu}$ is increasing. It only remains to verify $\overline{\mu}$ is subadditive. Then let $S = \cup_{i=1}^{\infty} S_i$. If any $\overline{\mu}(S_i) = \infty$, there is nothing to prove so suppose $\overline{\mu}(S_i) < \infty$ for each i. Then there exist $E_i \in \mathcal{F}$ such that $E_i \supseteq S_i$ and

$$\overline{\mu}(S_i) + \varepsilon/2^i > \mu(E_i).$$

Then

$$\begin{aligned} \overline{\mu}(S) &= \overline{\mu}(\cup_i S_i) \\ &\leq \mu(\cup_i E_i) \leq \sum_i \mu(E_i) \\ &\leq \sum_i \left(\overline{\mu}(S_i) + \varepsilon/2^i\right) = \sum_i \overline{\mu}(S_i) + \varepsilon. \end{aligned}$$

Since ε is arbitrary, this verifies $\overline{\mu}$ is subadditive and is an outer measure as claimed.

Denote by $\overline{\mathcal{F}}$ the σ algebra of measurable sets in the sense of Caratheodory. Then it follows from the Caratheodory procedure, Theorem 7.5.4, that $\left(\Omega, \overline{\mathcal{F}}, \overline{\mu}\right)$ is a complete measure space. This verifies 1.

Now let $E \in \mathcal{F}$. Then from the definition of $\overline{\mu}$, it follows

$$\overline{\mu}(E) \equiv \inf\{\mu(F) : F \in \mathcal{F} \text{ and } F \supseteq E\} \leq \mu(E).$$

If $F \supseteq E$ and $F \in \mathcal{F}$, then $\mu(F) \geq \mu(E)$ and so $\mu(E)$ is a lower bound for all such $\mu(F)$ which shows that

$$\overline{\mu}(E) \equiv \inf\{\mu(F) : F \in \mathcal{F} \text{ and } F \supseteq E\} \geq \mu(E).$$

This verifies 2.

Next consider 3. Let $E \in \mathcal{F}$ and let S be a set. I must show

$$\overline{\mu}(S) \geq \overline{\mu}(S \setminus E) + \overline{\mu}(S \cap E).$$

If $\overline{\mu}(S) = \infty$ there is nothing to show. Therefore, suppose $\overline{\mu}(S) < \infty$. Then from the definition of $\overline{\mu}$ there exists $G \supseteq S$ such that $G \in \mathcal{F}$ and $\mu(G) = \overline{\mu}(S)$. Then from the definition of $\overline{\mu}$,

$$\begin{aligned} \overline{\mu}(S) &\leq \overline{\mu}(S \setminus E) + \overline{\mu}(S \cap E) \\ &\leq \mu(G \setminus E) + \mu(G \cap E) \\ &= \mu(G) = \overline{\mu}(S) \end{aligned}$$

This verifies 3.

Claim 4 comes by the definition of $\overline{\mu}$ as used above. The other case is when $\overline{\mu}(S) = \infty$. However, in this case, you can let $G = \Omega$.

It only remains to verify 5. Let the Ω_n be as described above and let $E \in \overline{\mathcal{F}}$ such that $E \subseteq \Omega_n$. By 4 there exists $H \in \mathcal{F}$ such that $H \subseteq \Omega_n$, $H \supseteq \Omega_n \setminus E$, and

$$\mu(H) = \overline{\mu}(\Omega_n \setminus E). \tag{7.15}$$

Then let $F \equiv \Omega_n \cap H^C$. It follows $F \subseteq E$ and

$$\begin{aligned} E \setminus F &= E \cap F^C = E \cap \left(H \cup \Omega_n^C\right) \\ &= E \cap H = H \setminus (\Omega_n \setminus E) \end{aligned}$$

Hence from 7.15

$$\overline{\mu}(E \setminus F) = \overline{\mu}(H \setminus (\Omega_n \setminus E)) = 0.$$

It follows

$$\overline{\mu}(E) = \overline{\mu}(F) = \mu(F).$$

In the case where $E \in \overline{\mathcal{F}}$ is arbitrary, not necessarily contained in some Ω_n, it follows from what was just shown that there exists $F_n \in \mathcal{F}$ such that $F_n \subseteq E \cap \Omega_n$ and

$$\mu(F_n) = \overline{\mu}(E \cap \Omega_n).$$

Letting $F \equiv \cup_n F_n$

$$\overline{\mu}(E \setminus F) \leq \overline{\mu}(\cup_n (E \cap \Omega_n \setminus F_n)) \leq \sum_n \overline{\mu}(E \cap \Omega_n \setminus F_n) = 0.$$

Therefore, $\overline{\mu}(E) = \mu(F)$ and this proves 5. This proves the theorem. ■

Here is another observation about regularity which follows from the above theorem.

Theorem 7.5.7 *Suppose μ is a regular measure defined on $\mathcal{B}(X)$ where X is a finite dimensional normed vector space. Then denoting by $\left(X, \overline{\mathcal{B}(X)}, \overline{\mu}\right)$ the completion of $(X, \mathcal{B}(X), \mu)$, it follows $\overline{\mu}$ is also regular. Furthermore, if a σ algebra, $\mathcal{F} \supseteq \mathcal{B}(X)$ and $(X, \mathcal{F}, \mu)$ is a complete measure space such that for every $F \in \mathcal{F}$ there exists $G \in \mathcal{B}(X)$ such that $\mu(F) = \mu(G)$ and $G \supseteq F$, then $\mathcal{F} = \overline{\mathcal{B}(X)}$ and $\mu = \overline{\mu}$.*

Proof: Let $F \in \overline{\mathcal{B}(X)}$ with $\overline{\mu}(F) < \infty$. By Theorem 7.5.6 there exists $G \in \mathcal{B}(X)$ such that

$$\overline{\mu}(G) = \mu(G) = \mu(F).$$

Now by regularity of μ there exists an open set, $V \supseteq G \supseteq F$ such that

$$\overline{\mu}(F) + \varepsilon = \mu(G) + \varepsilon > \mu(V) = \overline{\mu}(V)$$

Therefore, $\overline{\mu}$ is outer regular. If $\overline{\mu}(F) = \infty$, there is nothing to show.

Now take $F \in \overline{\mathcal{B}(X)}$. By Theorem 7.5.6 there exists $H \subseteq F$ with $H \in \mathcal{B}(X)$ and $\mu(H) = \overline{\mu}(F)$. If $l < \overline{\mu}(F) = \mu(H)$, it follows from regularity of μ there exists K a compact subset of H such that

$$l < \mu(K) = \overline{\mu}(K)$$

Thus $\overline{\mu}$ is also inner regular. The last assertion follows from the uniqueness part of Theorem 7.5.6 and This proves the theorem. ■

A repeat of the above argument yields the following corollary.

Corollary 7.5.8 *The conclusion of the above theorem holds for X replaced with Y where Y is a closed subset of X.*

7.6 One Dimensional Lebesgue Stieltjes Measure

Now with these major results about measures, it is time to specialize to the outer measure of Theorem 7.2.1. The next theorem gives Lebesgue Stieltjes measure on $\mathbb{R}$.

Theorem 7.6.1 *Let $\mathcal{S}$ denote the σ algebra of Theorem 7.5.4 applied to the outer measure μ in Theorem 7.2.1 on which μ is a measure. Then every open interval is in $\mathcal{S}$. So are all open and closed sets. Furthermore, if E is any set in $\mathcal{S}$*

$$\mu(E) = \sup\{\mu(K) : K \text{ is a closed and bounded set, } K \subseteq E\} \tag{7.16}$$

$$\mu(E) = \inf\{\mu(V) : V \text{ is an open set, } V \supseteq E\} \tag{7.17}$$

Proof: The first task is to show $(a,b) \in \mathcal{S}$. I need to show that for every $S \subseteq \mathbb{R}$,

$$\mu(S) \geq \mu(S \cap (a,b)) + \mu\left(S \cap (a,b)^C\right) \tag{7.18}$$

Suppose first S is an open interval, (c,d). If (c,d) has empty intersection with (a,b) or is contained in (a,b) there is nothing to prove. The above expression reduces to nothing more than $\mu(S) = \mu(S)$. Suppose next that $(c,d) \supseteq (a,b)$. In this case, the right side of the above reduces to

$$\begin{aligned} & \mu((a,b)) + \mu((c,a] \cup [b,d)) \\ \leq\ & F(b-) - F(a+) + F(a+) - F(c+) + F(d-) - F(b-) \\ =\ & F(d-) - F(c+) = \mu((c,d)) \end{aligned}$$

The only other cases are $c \leq a < d \leq b$ or $a \leq c < d \leq b$. Consider the first of these cases. Then the right side of 7.18 for $S = (c,d)$ is

$$\begin{aligned} \mu((a,d)) + \mu((c,a]) &= F(d-) - F(a+) + F(a+) - F(c+) \\ &= F(d-) - F(c+) = \mu((c,d)) \end{aligned}$$

The last case is entirely similar. Thus 7.18 holds whenever S is an open interval. Now it is clear 7.18 also holds if $\mu(S) = \infty$. Suppose then that $\mu(S) < \infty$ and let

$$S \subseteq \cup_{k=1}^{\infty} (a_k, b_k)$$

such that

$$\mu(S) + \varepsilon > \sum_{k=1}^{\infty} (F(b_k-) - F(a_k+)) = \sum_{k=1}^{\infty} \mu((a_k,b_k)).$$

Then since μ is an outer measure, and using what was just shown,

$$\begin{aligned} & \mu(S \cap (a,b)) + \mu\left(S \cap (a,b)^C\right) \\ \leq\ & \mu(\cup_{k=1}^{\infty}(a_k,b_k) \cap (a,b)) + \mu\left(\cup_{k=1}^{\infty}(a_k,b_k) \cap (a,b)^C\right) \\ \leq\ & \sum_{k=1}^{\infty} \mu((a_k,b_k) \cap (a,b)) + \mu\left((a_k,b_k) \cap (a,b)^C\right) \\ \leq\ & \sum_{k=1}^{\infty} \mu((a_k,b_k)) \leq \mu(S) + \varepsilon. \end{aligned}$$

Since ε is arbitrary, this shows 7.18 holds for any S and so any open interval is in S.

It follows any open set is in $\mathcal{S}$. This follows from Theorem 5.3.10 which implies that if U is open, it is the countable union of disjoint open intervals. Since each of these open intervals is in $\mathcal{S}$ and $\mathcal{S}$ is a σ algebra, their union is also in $\mathcal{S}$. It follows every closed set is in $\mathcal{S}$ also. This is because $\mathcal{S}$ is a σ algebra and if a set is in $\mathcal{S}$ then so is its complement. The closed sets are those which are complements of open sets.

Thus the σ algebra of μ measurable sets $\mathcal{F}$ includes $\mathcal{B}(\mathbb{R})$. Consider the completion of the measure space, $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$, $\left(\mathbb{R}, \overline{\mathcal{B}(\mathbb{R})}, \overline{\mu}\right)$. By the uniqueness assertion in Theorem 7.5.6 and the fact that $(\mathbb{R}, \mathcal{F}, \mu)$ is complete, this coincides with $(\mathbb{R}, \mathcal{F}, \mu)$ because the construction of μ implies μ is outer regular and for every $F \in \mathcal{F}$, there exists $G \in \mathcal{B}(\mathbb{R})$ containing F such that $\mu(F) = \mu(G)$. In fact, you can take G to equal a countable intersection of open sets. By Theorem 7.4.6 μ is regular on every set of $\mathcal{B}(\mathbb{R})$, this because μ is finite on compact sets. Therefore, by Theorem 7.5.7 $\mu = \overline{\mu}$ is regular on $\mathcal{F}$ which verifies the last two claims. This proves the theorem. ■

7.7 Measurable Functions

The integral will be defined on measurable functions which is the next topic considered. It is sometimes convenient to allow functions to take the value $+\infty$. You should think of $+\infty$, usually referred to as ∞ as something out at the right end of the real line and its only importance is the notion of sequences converging to it. $x_n \to \infty$ exactly when for all $l \in \mathbb{R}$, there exists N such that if $n \geq N$, then

$$x_n > l.$$

This is what it means for a sequence to converge to ∞. Don't think of ∞ as a number. It is just a convenient symbol which allows the consideration of some limit operations more simply. Similar considerations apply to $-\infty$ but this value is not of very great interest. In fact the set of most interest for the values of a function, f is the complex numbers or more generally some normed vector space.

Recall the notation,

$$f^{-1}(A) \equiv \{x : f(x) \in A\} \equiv [f(x) \in A]$$

in whatever context the notation occurs.

Lemma 7.7.1 *Let $f : \Omega \to (-\infty, \infty]$ where $\mathcal{F}$ is a σ algebra of subsets of Ω. Then the following are equivalent.*

$$f^{-1}((d,\infty]) \in \mathcal{F} \text{ for all finite } d,$$

$$f^{-1}((-\infty,d)) \in \mathcal{F} \text{ for all finite } d,$$

$$f^{-1}([d,\infty]) \in \mathcal{F} \text{ for all finite } d,$$

$$f^{-1}((-\infty,d]) \in \mathcal{F} \text{ for all finite } d,$$

$$f^{-1}((a,b)) \in \mathcal{F} \text{ for all } a < b, -\infty < a < b < \infty.$$

Proof: First note that the first and the third are equivalent. To see this, observe

$$f^{-1}([d,\infty]) = \cap_{n=1}^{\infty} f^{-1}((d-1/n,\infty]),$$

and so if the first condition holds, then so does the third.

$$f^{-1}((d,\infty]) = \cup_{n=1}^{\infty} f^{-1}([d+1/n,\infty]),$$

and so if the third condition holds, so does the first.

Similarly, the second and fourth conditions are equivalent. Now

$$f^{-1}((-\infty,d]) = (f^{-1}((d,\infty]))^C$$

so the first and fourth conditions are equivalent. Thus the first four conditions are equivalent and if any of them hold, then for $-\infty < a < b < \infty$,

$$f^{-1}((a,b)) = f^{-1}((-\infty,b)) \cap f^{-1}((a,\infty]) \in \mathcal{F}.$$

Finally, if the last condition holds,

$$f^{-1}([d,\infty]) = \left(\cup_{k=1}^{\infty} f^{-1}((-k+d,d))\right)^C \in \mathcal{F}$$

and so the third condition holds. Therefore, all five conditions are equivalent. This proves the lemma. ■

This lemma allows for the following definition of a measurable function having values in $(-\infty, \infty]$.

Definition 7.7.2 *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $f : \Omega \to (-\infty, \infty]$. Then f is said to be $\mathcal{F}$ measurable if any of the equivalent conditions of Lemma 7.7.1 hold.*

Theorem 7.7.3 *Let f_n and f be functions mapping Ω to $(-\infty, \infty]$ where $\mathcal{F}$ is a σ algebra of measurable sets of Ω. Then if f_n is measurable, and $f(\omega) = \lim_{n\to\infty} f_n(\omega)$, it follows that f is also measurable. (Pointwise limits of measurable functions are measurable.)*

Proof: The idea is to show $f^{-1}((a,b)) \in \mathcal{F}$. Let $V_m \equiv \left(a + \frac{1}{m}, b - \frac{1}{m}\right)$ and $\overline{V}_m = \left[a + \frac{1}{m}, b - \frac{1}{m}\right]$. Then for all m, $V_m \subseteq (a,b)$ and

$$(a,b) = \cup_{m=1}^{\infty} V_m = \cup_{m=1}^{\infty} \overline{V}_m.$$

Note that $V_m \neq \emptyset$ for all m large enough. Since f is the pointwise limit of f_n,

$$f^{-1}(V_m) \subseteq \{\omega : f_k(\omega) \in V_m \text{ for all } k \text{ large enough}\} \subseteq f^{-1}(\overline{V}_m).$$

You should note that the expression in the middle is of the form

$$\cup_{n=1}^{\infty} \cap_{k=n}^{\infty} f_k^{-1}(V_m).$$

Therefore,

$$\begin{aligned} f^{-1}((a,b)) &= \cup_{m=1}^{\infty} f^{-1}(V_m) \subseteq \cup_{m=1}^{\infty} \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} f_k^{-1}(V_m) \\ &\subseteq \cup_{m=1}^{\infty} f^{-1}(\overline{V}_m) = f^{-1}((a,b)). \end{aligned}$$

It follows $f^{-1}((a,b)) \in \mathcal{F}$ because it equals the expression in the middle which is measurable. This shows f is measurable.

Proposition 7.7.4 *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $f : \Omega \to (-\infty, \infty]$. Then f is $\mathcal{F}$ measurable if and only if $f^{-1}(U) \in \mathcal{F}$ whenever U is an open set in $\mathbb{R}$.*

Proof: If $f^{-1}(U) \in \mathcal{F}$ whenever U is an open set in $\mathbb{R}$ then it follows from the last condition of Lemma 7.7.1 that f is measurable. Next suppose f is measurable so this last condition of Lemma 7.7.1 holds. Then by Theorem 5.3.10 if U is any open set in $\mathbb{R}$, it is the countable union of open intervals, $U = \cup_{k=1}^{\infty}(a_k, b_k)$. Hence

$$f^{-1}(U) = \cup_{k=1}^{\infty} f^{-1}((a_k, b_k)) \in \mathcal{F}$$

because $\mathcal{F}$ is a σ algebra.

From this proposition, it follows one can generalize the definition of a measurable function to those which have values in any normed vector space as follows.

Definition 7.7.5 *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $f : \Omega \to X$ where X is a normed vector space. Then f is measurable means $f^{-1}(U) \in \mathcal{F}$ whenever U is an open set in X.*

Now here is an important theorem which shows that you can do lots of things to measurable functions and still have a measurable function.

Theorem 7.7.6 *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let X, Y be normed vector spaces and $\mathbf{g} : X \to Y$ continuous. Then if $\mathbf{f} : \Omega \to X$ is $\mathcal{F}$ measurable, it follows $\mathbf{g} \circ \mathbf{f}$ is also $\mathcal{F}$ measurable.*

Proof: From the definition, it suffices to show $(\mathbf{g}\circ\mathbf{f})^{-1}(U)\in\mathcal{F}$ whenever U is an open set in Y. However, since $\mathbf{g}$ is continuous, it follows $\mathbf{g}^{-1}(U)$ is open and so

$$(\mathbf{g}\circ\mathbf{f})^{-1}(U)=\mathbf{f}^{-1}\left(\mathbf{g}^{-1}(U)\right)=\mathbf{f}^{-1}(\text{an open set})\in\mathcal{F}.$$

This proves the theorem. ■

This theorem implies for example that if $\mathbf{f}$ is a measurable X valued function, then $||\mathbf{f}||$ is a measurable $\mathbb{R}$ valued function. It also implies that if $\mathbf{f}$ is an X valued function, then if $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ is a basis for X and π_k is the projection onto the k^{th} component, then $\pi_k\circ\mathbf{f}$ is a measurable $\mathbb{F}$ valued function. Does it go the other way? That is, if it is known that $\pi_k\circ\mathbf{f}$ is measurable for each k, does it follow $\mathbf{f}$ is measurable? The following technical lemma is interesting for its own sake.

Lemma 7.7.7 *Let $||\mathbf{x}||\equiv\max\{|x_i|,i=1,2,\cdots,n\}$ for $\mathbf{x}\in\mathbb{F}^n$. Then every set U which is open in $\mathbb{F}^n$ is the countable union of balls of the form $B(\mathbf{x},r)$ where the open ball is defined in terms of the above norm.*

Proof: By Theorem 5.8.3 if you consider the two normed vector spaces $(\mathbb{F}^n,|\cdot|)$ and $(\mathbb{F}^n,||\cdot||)$, the identity map is continuous in both directions. Therefore, if a set, U is open with respect to $|\cdot|$ it follows it is open with respect to $||\cdot||$ and the other way around. The other thing to notice is that there exists a countable dense subset of $\mathbb{F}$. The rationals will work if $\mathbb{F}=\mathbb{R}$ and if $\mathbb{F}=\mathbb{C}$, then you use $\mathbb{Q}+i\mathbb{Q}$. Letting D be a countable dense subset of $\mathbb{F}$, D^n is a countable dense subset of $\mathbb{F}^n$. It is countable because it is a finite Cartesian product of countable sets and you can use Theorem 2.1.7 of Page 15 repeatedly. It is dense because if $\mathbf{x}\in\mathbb{F}^n$, then by density of D, there exists $d_j\in D$ such that

$$|d_j-x_j|<\varepsilon$$

then $\mathbf{d}\equiv(d_1,\cdots,d_n)$ is such that $||\mathbf{d}-\mathbf{x}||<\varepsilon$.

Now consider the set of open balls,

$$\mathcal{B}\equiv\{B(\mathbf{d},r):\mathbf{d}\in D^n,r\in\mathbb{Q}\}.$$

This collection of open balls is countable by Theorem 2.1.7 of Page 15. I claim every open set is the union of balls from $\mathcal{B}$. Let U be an open set in $\mathbb{F}^n$ and $\mathbf{x}\in U$. Then there exists $\delta>0$ such that $B(\mathbf{x},\delta)\subseteq U$. There exists $\mathbf{d}\in D^n\cap B(\mathbf{x},\delta/5)$. Then pick rational number $\delta/5<r<2\delta/5$. Consider the set of $\mathcal{B}$, $B(\mathbf{d},r)$. Then $\mathbf{x}\in B(\mathbf{d},r)$ because $r>\delta/5$. However, it is also the case that $B(\mathbf{d},r)\subseteq B(\mathbf{x},\delta)$ because if $\mathbf{y}\in B(\mathbf{d},r)$ then

$$\begin{aligned}||\mathbf{y}-\mathbf{x}||&\leq||\mathbf{y}-\mathbf{d}||+||\mathbf{d}-\mathbf{x}||\\&<\frac{2\delta}{5}+\frac{\delta}{5}<\delta.\end{aligned}$$

This proves the lemma. ■

Corollary 7.7.8 *Let $(\Omega,\mathcal{F},\mu)$ be a measure space and let X be a normed vector space with basis $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$. Let π_k be the k^{th} projection map onto the k^{th} component. Thus*

$$\pi_k\mathbf{x}\equiv x_k\text{ where }\mathbf{x}=\sum_{i=1}^n x_i\mathbf{v}_i.$$

Then each $\pi_k\circ\mathbf{f}$ is a measurable $\mathbb{F}$ valued function if and only if $\mathbf{f}$ is a measurable X valued function.

Proof: The if part has already been noted. Suppose that each $\pi_k \circ \mathbf{f}$ is an $\mathbb{F}$ valued measurable function. Let $\mathbf{g} : X \to \mathbb{F}^n$ be given by

$$\mathbf{g}(\mathbf{x}) \equiv (\pi_1 \mathbf{x}, \cdots, \pi_n \mathbf{x}).$$

Thus $\mathbf{g}$ is linear, one to one, and onto. By Theorem 5.8.3 both $\mathbf{g}$ and $\mathbf{g}^{-1}$ are continuous. Therefore, every open set in X is of the form $\mathbf{g}^{-1}(U)$ where U is an open set in $\mathbb{F}^n$. To see this, start with V open set in X. Since $\mathbf{g}^{-1}$ is continuous, $\mathbf{g}(V)$ is open in $\mathbb{F}^n$ and so $V = \mathbf{g}^{-1}(\mathbf{g}(V))$. Therefore, it suffices to show that for every U an open set in $\mathbb{F}^n$,

$$\mathbf{f}^{-1}\left(\mathbf{g}^{-1}(U)\right) = (\mathbf{g} \circ \mathbf{f})^{-1}(U) \in \mathcal{F}.$$

By Lemma 7.7.7 there are countably many open balls of the form $B(\mathbf{x}_j, r_j)$ such that U is equal to the union of these balls. Thus

$$\begin{aligned} (\mathbf{g} \circ \mathbf{f})^{-1}(U) &= (\mathbf{g} \circ \mathbf{f})^{-1}\left(\cup_{k=1}^{\infty} B(\mathbf{x}_k, r_k)\right) \\ &= \cup_{k=1}^{\infty} (\mathbf{g} \circ \mathbf{f})^{-1}(B(\mathbf{x}_k, r_k)) \end{aligned} \tag{7.19}$$

Now from the definition of the norm,

$$B(\mathbf{x}_k, r_k) = \prod_{j=1}^{n} (x_{kj} - \delta, x_{kj} + \delta)$$

and so

$$(\mathbf{g} \circ \mathbf{f})^{-1}(B(\mathbf{x}_k, r_k)) = \cap_{j=1}^{n} (\pi_j \circ \mathbf{f})^{-1}((x_{kj} - \delta, x_{kj} + \delta)) \in \mathcal{F}.$$

It follows 7.19 is the countable union of sets in $\mathcal{F}$ and so it is also in $\mathcal{F}$. This proves the corollary. ■

Note that if $\{f_i\}_{i=1}^n$ are measurable functions defined on $(\Omega, \mathcal{F}, \mu)$ having values in $\mathbb{F}$ then letting $\mathbf{f} \equiv (f_1, \cdots, f_n)$, it follows $\mathbf{f}$ is a measurable $\mathbb{F}^n$ valued function. Now let $\Sigma : \mathbb{F}^n \to \mathbb{F}$ be given by $\Sigma(\mathbf{x}) \equiv \sum_{k=1}^n a_k x_k$. Then Σ is linear and so by Theorem 5.8.3 it follows Σ is continuous. Hence by Theorem 7.7.6, $\Sigma(\mathbf{f})$ is an $\mathbb{F}$ valued measurable function. Thus linear combinations of measurable functions are measurable. By similar reasoning, products of measurable functions are measurable. In general, it seems like you can start with a collection of measurable functions and do almost anything you like with them and the result, if it is a function will be measurable. This is in stark contrast to the functions which are generalized Riemann integrable.

The following theorem considers the case of functions which have values in a normed vector space.

Theorem 7.7.9 *Let $\{\mathbf{f}_n\}$ be a sequence of measurable functions mapping Ω to X where X is a normed vector space and $(\Omega, \mathcal{F})$ is a measure space. Suppose also that $\mathbf{f}(\omega) = \lim_{n\to\infty} \mathbf{f}_n(\omega)$ for all $\omega \in \Omega$. Then $\mathbf{f}$ is also a measurable function.*

Proof: It is required to show $\mathbf{f}^{-1}(U)$ is measurable for all U open. Let

$$V_m \equiv \left\{\mathbf{x} \in U : \text{dist}\left(\mathbf{x}, U^C\right) > \frac{1}{m}\right\}.$$

Thus

$$V_m \subseteq \left\{\mathbf{x} \in U : \text{dist}\left(\mathbf{x}, U^C\right) \geq \frac{1}{m}\right\}$$

and $V_m \subseteq \overline{V_m} \subseteq V_{m+1}$ and $\cup_m V_m = U$. Then since V_m is open, it follows that if $\mathbf{f}(\omega) \in V_m$ then for all sufficiently large k, it must be the case $\mathbf{f}_k(\omega) \in V_m$ also. That is, $\omega \in \mathbf{f}_k^{-1}(V_m)$ for all sufficiently large k. Thus

$$\mathbf{f}^{-1}(V_m) = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} \mathbf{f}_k^{-1}(V_m)$$

and so

$$\begin{aligned}\mathbf{f}^{-1}(U) &= \cup_{m=1}^{\infty}\mathbf{f}^{-1}(V_m)\\ &= \cup_{m=1}^{\infty}\cup_{n=1}^{\infty}\cap_{k=n}^{\infty}\mathbf{f}_k^{-1}(V_m)\\ &\subseteq \cup_{m=1}^{\infty}\mathbf{f}^{-1}\left(\overline{V_m}\right)=\mathbf{f}^{-1}(U)\end{aligned}$$

which shows $\mathbf{f}^{-1}(U)$ is measurable. The step from the second to the last line follows because if $\omega \in \cup_{n=1}^{\infty}\cap_{k=n}^{\infty}\mathbf{f}_k^{-1}(V_m)$, this says $\mathbf{f}_k(\omega) \in V_m$ for all k large enough. Therefore, the point of X to which the sequence $\{\mathbf{f}_k(\omega)\}$ converges must be in $\overline{V_m}$ which equals $V_m \cup V_m'$, the limit points of V_m. This proves the theorem. ■

Now here is a simple observation involving something called simple functions. It uses the following notation.

Notation 7.7.10 *For E a set let $\mathcal{X}_E(\omega)$ be defined by*

$$\mathcal{X}_E(x)=\begin{cases}1 \text{ if } \omega \in E\\ 0 \text{ if } \omega \notin E\end{cases}$$

Theorem 7.7.11 *Let $\mathbf{f}:\Omega \to X$ where X is some normed vector space. Suppose*

$$\mathbf{f}(\omega)=\sum_{k=1}^{m}\mathbf{x}_k\mathcal{X}_{A_k}(\omega)$$

where each $\mathbf{x}_k \in X$ and the A_k are disjoint measurable sets. (Such functions are often referred to as simple functions.) Then $\mathbf{f}$ is measurable.

Proof: Letting U be open, $\mathbf{f}^{-1}(U)=\cup\{A_k : \mathbf{x}_k \in U\}$, a finite union of measurable sets.

In the Lebesgue integral, the simple functions play a role similar to step functions in the theory of the Riemann integral. Also there is a fundamental theorem about measurable functions and simple functions which says essentially that the measurable functions are those which are pointwise limits of simple functions.

Theorem 7.7.12 *Let $f \geq 0$ be measurable with respect to the measure space $(\Omega,\mathcal{F},\mu)$. Then there exists a sequence of nonnegative simple functions $\{s_n\}$ satisfying*

$$0 \leq s_n(\omega) \tag{7.20}$$

$$\cdots\ s_n(\omega) \leq s_{n+1}(\omega)\cdots$$

$$f(\omega)=\lim_{n\to\infty}s_n(\omega) \text{ for all } \omega \in \Omega. \tag{7.21}$$

If f is bounded the convergence is actually uniform.

Proof: First note that

$$\begin{aligned}f^{-1}([a,b)) &= f^{-1}((-\infty,a))^C \cap f^{-1}((-\infty,b))\\ &= \left(f^{-1}((-\infty,a)) \cup f^{-1}((-\infty,b))^C\right)^C \in \mathcal{F}.\end{aligned}$$

Letting $I \equiv \{\omega : f(\omega)=\infty\}$, define

$$t_n(\omega)=\sum_{k=0}^{2^n}\frac{k}{n}\mathcal{X}_{f^{-1}([k/n,(k+1)/n))}(\omega)+n\mathcal{X}_I(\omega).$$

Then $t_n(\omega) \leq f(\omega)$ for all ω and $\lim_{n\to\infty} t_n(\omega) = f(\omega)$ for all ω. This is because $t_n(\omega) = n$ for $\omega \in I$ and if $f(\omega) \in [0, \frac{2^n+1}{n})$, then

$$0 \leq f(\omega) - t_n(\omega) \leq \frac{1}{n}. \tag{7.22}$$

Thus whenever $\omega \notin I$, the above inequality will hold for all n large enough. Let

$$s_1 = t_1,\ s_2 = \max(t_1, t_2),\ s_3 = \max(t_1, t_2, t_3), \cdots.$$

Then the sequence $\{s_n\}$ satisfies 7.20-7.21.

To verify the last claim, note that in this case the term $n\mathcal{X}_I(\omega)$ is not present. Therefore, for all n large enough that $2^n n \geq f(\omega)$ for all ω, 7.22 holds for all ω. Thus the convergence is uniform. This proves the theorem. ■

7.8 Exercises

1. Let $\mathcal{C}$ be a set whose elements are σ algebras of subsets of Ω. Show $\cap\mathcal{C}$ is a σ algebra also.

2. Let Ω be any set. Show $\mathcal{P}(\Omega)$, the set of all subsets of Ω is a σ algebra. Now let $\mathcal{L}$ denote some subset of $\mathcal{P}(\Omega)$. Consider all σ algebras which contain $\mathcal{L}$. Show the intersection of all these σ algebras which contain $\mathcal{L}$ is a σ algebra containing $\mathcal{L}$ and it is the smallest σ algebra containing $\mathcal{L}$, denoted by $\sigma(\mathcal{L})$. When Ω is a normed vector space, and $\mathcal{L}$ consists of the open sets $\sigma(\mathcal{L})$ is called the σ algebra of Borel sets.

3. Consider $\Omega = [0,1]$ and let $\mathcal{S}$ denote all subsets of $[0,1], F$ such that either F^C or F is countable. Note the empty set must be countable. Show $\mathcal{S}$ is a σ algebra. (This is a sick σ algebra.) Now let $\mu : \mathcal{S} \to [0,\infty]$ be defined by $\mu(F) = 1$ if F^C is countable and $\mu(F) = 0$ if F is countable. Show μ is a measure on $\mathcal{S}$.

4. Let $\Omega = \mathbb{N}$, the positive integers and let a σ algebra be given by $\mathcal{F} = \mathcal{P}(\mathbb{N})$, the set of all subsets of $\mathbb{N}$. What are the measurable functions having values in $\mathbb{C}$? Let $\mu(E)$ be the number of elements of E where E is a subset of $\mathbb{N}$. Show μ is a measure.

5. Let $\mathcal{F}$ be a σ algebra of subsets of Ω and suppose $\mathcal{F}$ has infinitely many elements. Show that $\mathcal{F}$ is uncountable. **Hint:** You might try to show there exists a countable sequence of disjoint sets of $\mathcal{F}$, $\{A_i\}$. It might be easiest to verify this by contradiction if it doesn't exist rather than a direct construction however, I have seen this done several ways. Once this has been done, you can define a map, θ, from $\mathcal{P}(\mathbb{N})$ into $\mathcal{F}$ which is one to one by $\theta(S) = \cup_{i\in S} A_i$. Then argue $\mathcal{P}(\mathbb{N})$ is uncountable and so $\mathcal{F}$ is also uncountable.

6. A probability space is a measure space, $(\Omega, \mathcal{F}, P)$ where the measure, P has the property that $P(\Omega) = 1$. Such a measure is called a probability measure. Random vectors are measurable functions, $\mathbf{X}$, mapping a probability space, $(\Omega, \mathcal{F}, P)$ to $\mathbb{R}^n$. Thus $\mathbf{X}(\omega) \in \mathbb{R}^n$ for each $\omega \in \Omega$ and P is a probability measure defined on the sets of $\mathcal{F}$, a σ algebra of subsets of Ω. For E a Borel set in $\mathbb{R}^n$, define

$$\mu(E) \equiv P\left(\mathbf{X}^{-1}(E)\right) \equiv \text{ probability that } \mathbf{X} \in E.$$

Show this is a well defined probability measure on the Borel sets of $\mathbb{R}^n$. Thus $\mu(E) = P(\mathbf{X}(\omega) \in E)$. It is called the distribution. Explain why μ must be regular.

7. Suppose $(\Omega, \mathcal{S}, \mu)$ is a measure space which may not be complete. Show that another way to complete the measure space is to define $\overline{\mathcal{S}}$ to consist of all sets of the form E where there exists $F \in \mathcal{S}$ such that $(F \setminus E) \cup (E \setminus F) \subseteq N$ for some $N \in \mathcal{S}$ which has measure zero and then let $\mu(E) = \mu_1(F)$? Explain.

Chapter 8

The Abstract Lebesgue Integral

The general Lebesgue integral requires a measure space, $(\Omega, \mathcal{F}, \mu)$ and, to begin with, a nonnegative measurable function. I will use Lemma 2.3.3 about interchanging two supremums frequently. Also, I will use the observation that if $\{a_n\}$ is an increasing sequence of points of $[0, \infty]$, then $\sup_n a_n = \lim_{n\to\infty} a_n$ which is obvious from the definition of sup.

8.1 Definition For Nonnegative Measurable Functions

8.1.1 Riemann Integrals For Decreasing Functions

First of all, the notation

$$[g < f]$$

is short for

$$\{\omega \in \Omega : g(\omega) < f(\omega)\}$$

with other variants of this notation being similar. Also, the convention, $0 \cdot \infty = 0$ will be used to simplify the presentation whenever it is convenient to do so.

Definition 8.1.1 *For f a nonnegative decreasing function defined on a finite interval $[a, b]$, define*

$$\int_a^b f(\lambda)\, d\lambda \equiv \lim_{M\to\infty} \int_a^b M \wedge f(\lambda)\, d\lambda = \sup_M \int_a^b M \wedge f(\lambda)\, d\lambda$$

where $a \wedge b$ means the minimum of a and b. Note that for f bounded,

$$\sup_M \int_a^b M \wedge f(\lambda)\, d\lambda = \int_a^b f(\lambda)\, d\lambda$$

where the integral on the right is the usual Riemann integral because eventually $M > f$. For f a nonnegative decreasing function defined on $[0, \infty)$,

$$\int_0^\infty f d\lambda \equiv \lim_{R\to\infty} \int_0^R f d\lambda = \sup_{R>1} \int_0^R f d\lambda = \sup_R \sup_{M>0} \int_0^R f \wedge M d\lambda$$

Since decreasing bounded functions are Riemann integrable, the above definition is well defined. Now here are some obvious properties.

Lemma 8.1.2 *Let f be a decreasing nonnegative function defined on an interval $[a,b]$. Then if $[a,b] = \cup_{k=1}^{m} I_k$ where $I_k \equiv [a_k, b_k]$ and the intervals I_k are non overlapping, it follows*

$$\int_a^b f d\lambda = \sum_{k=1}^{m} \int_{a_k}^{b_k} f d\lambda.$$

Proof: This follows from the computation,

$$\begin{aligned} \int_a^b f d\lambda &\equiv \lim_{M\to\infty} \int_a^b f \wedge M d\lambda \\ &= \lim_{M\to\infty} \sum_{k=1}^{m} \int_{a_k}^{b_k} f \wedge M d\lambda = \sum_{k=1}^{m} \int_{a_k}^{b_k} f d\lambda \end{aligned}$$

Note both sides could equal $+\infty$. ■

8.1.2 The Lebesgue Integral For Nonnegative Functions

Here is the definition of the Lebesgue integral of a function which is measurable and has values in $[0,\infty]$.

Definition 8.1.3 *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and suppose $f : \Omega \to [0,\infty]$ is measurable. Then define*

$$\int f d\mu \equiv \int_0^\infty \mu([f > \lambda])\, d\lambda$$

which makes sense because $\lambda \to \mu([f > \lambda])$ is nonnegative and decreasing.

Lemma 8.1.4 *In the situation of the above definition,*

$$\int f d\mu = \sup_{h>0} \sum_{i=1}^{\infty} \mu([f > hi])\, h$$

Proof:

$$\int f d\mu \equiv \int_0^\infty \mu([f > \lambda])\, d\lambda = \sup_M \sup_{R>1} \int_0^R \mu([f > \lambda]) \wedge M d\lambda$$

$$= \sup_M \sup_{R>1} \sup_{h>0} \sum_{k=1}^{M(R)} h\,(\mu([f > kh]) \wedge M)$$

where $M(R)$ is such that $R - h \le M(R)\, h \le R$. The sum is just a lower sum for the integral. Hence this equals

$$\begin{aligned} &= \sup_{R>1} \sup_{h>0} \sup_M \sum_{k=1}^{M(R)} h\,(\mu([f > kh]) \wedge M) \\ &= \sup_{R>1} \sup_{h>0} \lim_{M\to\infty} \sum_{k=1}^{M(R)} h\,(\mu([f > kh]) \wedge M) \\ &= \sup_{h>0} \sup_{R>1} \sum_{k=1}^{M(R)} h\mu([f > kh]) = \sup_{h>0} \sum_{k=1}^{\infty} h\mu([f > kh]) \end{aligned}$$

■

8.2 The Lebesgue Integral For Nonnegative Simple Functions

To begin with, here is a useful lemma.

Lemma 8.2.1 *If $f(\lambda) = 0$ for all $\lambda > a$, where f is a decreasing nonnegative function, then*

$$\int_0^\infty f(\lambda)\, d\lambda = \int_0^a f(\lambda)\, d\lambda.$$

Proof: From the definition,

$$\begin{aligned}
\int_0^\infty f(\lambda)\, d\lambda &= \lim_{R\to\infty} \int_0^R f(\lambda)\, d\lambda = \sup_{R>1} \int_0^R f(\lambda)\, d\lambda \\
&= \sup_{R>1} \sup_{M} \int_0^R f(\lambda) \wedge M d\lambda \\
&= \sup_{M} \sup_{R>1} \int_0^R f(\lambda) \wedge M d\lambda \\
&= \sup_{M} \sup_{R>1} \int_0^a f(\lambda) \wedge M d\lambda \\
&= \sup_{M} \int_0^a f(\lambda) \wedge M d\lambda \equiv \int_0^a f(\lambda)\, d\lambda.
\end{aligned}$$

■

Now the Lebesgue integral for a nonnegative function has been defined, what does it do to a nonnegative simple function? Recall a nonnegative simple function is one which has finitely many nonnegative values which it assumes on measurable sets. Thus a simple function can be written in the form

$$s(\omega) = \sum_{i=1}^n c_i \mathcal{X}_{E_i}(\omega)$$

where the c_i are each nonnegative real numbers, the distinct values of s.

Lemma 8.2.2 *Let $s(\omega) = \sum_{i=1}^p a_i \mathcal{X}_{E_i}(\omega)$ be a nonnegative simple function where the E_i are distinct but the a_i might not be. Then*

$$\int s d\mu = \sum_{i=1}^p a_i \mu(E_i). \tag{8.1}$$

Proof: Without loss of generality, assume $0 \equiv a_0 < a_1 \leq a_2 \leq \cdots \leq a_p$ and that $\mu(E_i) < \infty, i > 0$. Here is why. If $\mu(E_i) = \infty$, then letting $a \in (a_{i-1}, a_i)$, by Lemma 8.2.1, the left side would be

$$\begin{aligned}
\int_0^{a_p} \mu([s > \lambda])\, d\lambda &\geq \int_{a_0}^{a_i} \mu([s > \lambda])\, d\lambda \\
&\equiv \sup_M \int_0^{a_i} \mu([s > \lambda]) \wedge M d\lambda \\
&= \sup_M M a_i = \infty
\end{aligned}$$

and so both sides are equal to ∞. Thus it can be assumed for each $i, \mu(E_i) < \infty$. Then it follows from Lemma 8.2.1 and Lemma 8.1.2,

$$\int_0^\infty \mu([s > \lambda])\, d\lambda = \int_0^{a_p} \mu([s > \lambda])\, d\lambda = \sum_{k=1}^{p} \int_{a_{k-1}}^{a_k} \mu([s > \lambda])\, d\lambda$$

$$= \sum_{k=1}^{p} \sum_{i=k}^{p} (a_k - a_{k-1})\, \mu(E_i) = \sum_{i=1}^{p} \mu(E_i) \sum_{k=1}^{i} (a_k - a_{k-1}) = \sum_{i=1}^{p} a_i \mu(E_i)$$

■

Lemma 8.2.3 *If $a, b \geq 0$ and if s and t are nonnegative simple functions, then*

$$\int as + btd\mu = a \int sd\mu + b \int td\mu.$$

Proof: Let

$$s(\omega) = \sum_{i=1}^{n} \alpha_i \mathcal{X}_{A_i}(\omega),\ t(\omega) = \sum_{i=1}^{m} \beta_j \mathcal{X}_{B_j}(\omega)$$

where α_i are the distinct values of s and the β_j are the distinct values of t. Clearly $as+bt$ is a nonnegative simple function because it has finitely many values on measurable sets In fact,

$$(as + bt)(\omega) = \sum_{j=1}^{m} \sum_{i=1}^{n} (a\alpha_i + b\beta_j) \mathcal{X}_{A_i \cap B_j}(\omega)$$

where the sets $A_i \cap B_j$ are disjoint and measurable. By Lemma 8.2.2,

$$\begin{aligned} \int as + btd\mu &= \sum_{j=1}^{m} \sum_{i=1}^{n} (a\alpha_i + b\beta_j) \mu(A_i \cap B_j) \\ &= \sum_{i=1}^{n} a \sum_{j=1}^{m} \alpha_i \mu(A_i \cap B_j) + b \sum_{j=1}^{m} \sum_{i=1}^{n} \beta_j \mu(A_i \cap B_j) \\ &= a \sum_{i=1}^{n} \alpha_i \mu(A_i) + b \sum_{j=1}^{m} \beta_j \mu(B_j) \\ &= a \int sd\mu + b \int td\mu. \end{aligned}$$

■

8.3 The Monotone Convergence Theorem

The following is called the monotone convergence theorem. This theorem and related convergence theorems are the reason for using the Lebesgue integral.

Theorem 8.3.1 *(Monotone Convergence theorem) Let f have values in $[0, \infty]$ and suppose $\{f_n\}$ is a sequence of nonnegative measurable functions having values in $[0, \infty]$ and satisfying*

$$\lim_{n \to \infty} f_n(\omega) = f(\omega) \text{ for each } \omega.$$

$$\cdots f_n(\omega) \leq f_{n+1}(\omega) \cdots$$

Then f is measurable and

$$\int fd\mu = \lim_{n \to \infty} \int f_n d\mu.$$

Proof: By Lemma 8.1.4

$$\lim_{n\to\infty}\int f_n d\mu = \sup_n \int f_n d\mu$$

$$\begin{aligned} &= \sup_n \sup_{h>0} \sum_{k=1}^{\infty} \mu\left([f_n > kh]\right) h = \sup_{h>0} \sup_N \sup_n \sum_{k=1}^{N} \mu\left([f_n > kh]\right) h \\ &= \sup_{h>0} \sup_N \sum_{k=1}^{N} \mu\left([f > kh]\right) h = \sup_{h>0} \sum_{k=1}^{\infty} \mu\left([f > kh]\right) h = \int f d\mu \end{aligned}$$

■

To illustrate what goes wrong without the Lebesgue integral, consider the following example.

Example 8.3.2 *Let $\{r_n\}$ denote the rational numbers in $[0,1]$ and let*

$$f_n(t) \equiv \begin{cases} 1 \text{ if } t \notin \{r_1, \cdots, r_n\} \\ 0 \text{ otherwise} \end{cases}$$

Then $f_n(t) \uparrow f(t)$ where f is the function which is one on the rationals and zero on the irrationals. Each f_n is Riemann integrable (why?) but f is not Riemann integrable. Therefore, you can't write $\int f dx = \lim_{n\to\infty} \int f_n dx$.

A meta-mathematical observation related to this type of example is this. If you can choose your functions, you don't need the Lebesgue integral. The Riemann Darboux integral is just fine. It is when you can't choose your functions and they come to you as pointwise limits that you really need the superior Lebesgue integral or at least something more general than the Riemann integral. The Riemann integral is entirely adequate for evaluating the seemingly endless lists of boring problems found in calculus books.

8.4 Other Definitions

To review and summarize the above, if $f \geq 0$ is measurable,

$$\int f d\mu \equiv \int_0^{\infty} \mu\left([f > \lambda]\right) d\lambda \tag{8.2}$$

another way to get the same thing for $\int f d\mu$ is to take an increasing sequence of nonnegative simple functions, $\{s_n\}$ with $s_n(\omega) \to f(\omega)$ and then by monotone convergence theorem,

$$\int f d\mu = \lim_{n\to\infty} \int s_n$$

where if $s_n(\omega) = \sum_{j=1}^{m} c_i \mathcal{X}_{E_i}(\omega)$,

$$\int s_n d\mu = \sum_{i=1}^{m} c_i \mu\left(E_i\right).$$

Similarly this also shows that for such nonnegative measurable function,

$$\int f d\mu = \sup\left\{\int s : 0 \leq s \leq f,\ s \text{ simple}\right\}$$

Here is an equivalent definition of the integral of a nonnegative measurable function. The fact it is well defined has been discussed above.

Definition 8.4.1 *For s a nonnegative simple function,*

$$s(\omega) = \sum_{k=1}^{n} c_k \mathcal{X}_{E_k}(\omega), \int s = \sum_{k=1}^{n} c_k \mu(E_k).$$

For f a nonnegative measurable function,

$$\int f d\mu = \sup\left\{\int s : 0 \leq s \leq f,\ s \text{ simple}\right\}.$$

8.5 Fatou's Lemma

The next theorem, known as Fatou's lemma is another important theorem which justifies the use of the Lebesgue integral.

Theorem 8.5.1 *(Fatou's lemma) Let f_n be a nonnegative measurable function with values in $[0,\infty]$. Let $g(\omega) = \liminf_{n\to\infty} f_n(\omega)$. Then g is measurable and*

$$\int g d\mu \leq \liminf_{n\to\infty} \int f_n d\mu.$$

In other words,

$$\int \left(\liminf_{n\to\infty} f_n\right) d\mu \leq \liminf_{n\to\infty} \int f_n d\mu$$

Proof: Let $g_n(\omega) = \inf\{f_k(\omega) : k \geq n\}$. Then

$$\begin{aligned} g_n^{-1}([a,\infty]) &= \cap_{k=n}^{\infty} f_k^{-1}([a,\infty]) \\ &= \left(\cup_{k=n}^{\infty} f_k^{-1}([a,\infty])^C\right)^C \in \mathcal{F}. \end{aligned}$$

Thus g_n is measurable by Lemma 7.7.1. Also $g(\omega) = \lim_{n\to\infty} g_n(\omega)$ so g is measurable because it is the pointwise limit of measurable functions. Now the functions g_n form an increasing sequence of nonnegative measurable functions so the monotone convergence theorem applies. This yields

$$\int g d\mu = \lim_{n\to\infty} \int g_n d\mu \leq \liminf_{n\to\infty} \int f_n d\mu.$$

The last inequality holding because

$$\int g_n d\mu \leq \int f_n d\mu.$$

(Note that it is not known whether $\lim_{n\to\infty} \int f_n d\mu$ exists.) This proves the theorem. ■

8.6 The Righteous Algebraic Desires Of The Lebesgue Integral

The monotone convergence theorem shows the integral wants to be linear. This is the essential content of the next theorem.

Theorem 8.6.1 *Let f, g be nonnegative measurable functions and let a, b be nonnegative numbers. Then $af + bg$ is measurable and*

$$\int (af + bg) d\mu = a \int f d\mu + b \int g d\mu. \tag{8.3}$$

Proof: By Theorem 7.7.12 on Page 184 there exist increasing sequences of nonnegative simple functions, $s_n \to f$ and $t_n \to g$. Then $af + bg$, being the pointwise limit of the simple functions $as_n + bt_n$, is measurable. Now by the monotone convergence theorem and Lemma 8.2.3,

$$\begin{aligned}\int (af + bg)\, d\mu &= \lim_{n\to\infty} \int as_n + bt_n d\mu \\ &= \lim_{n\to\infty} \left(a \int s_n d\mu + b \int t_n d\mu \right) \\ &= a \int f d\mu + b \int g d\mu.\end{aligned}$$

This proves the theorem. ■

As long as you are allowing functions to take the value $+\infty$, you cannot consider something like $f + (-g)$ and so you can't very well expect a satisfactory statement about the integral being linear until you restrict yourself to functions which have values in a vector space. This is discussed next.

8.7 The Lebesgue Integral, L^1

The functions considered here have values in $\mathbb{C}$, a vector space.

Definition 8.7.1 *Let $(\Omega, \mathcal{S}, \mu)$ be a measure space and suppose $f : \Omega \to \mathbb{C}$. Then f is said to be measurable if both $\operatorname{Re} f$ and $\operatorname{Im} f$ are measurable real valued functions.*

Definition 8.7.2 *A complex simple function will be a function which is of the form*

$$s(\omega) = \sum_{k=1}^{n} c_k \mathcal{X}_{E_k}(\omega)$$

where $c_k \in \mathbb{C}$ and $\mu(E_k) < \infty$. For s a complex simple function as above, define

$$I(s) \equiv \sum_{k=1}^{n} c_k \mu(E_k).$$

Lemma 8.7.3 *The definition, 8.7.2 is well defined. Furthermore, I is linear on the vector space of complex simple functions. Also the triangle inequality holds,*

$$|I(s)| \le I(|s|).$$

Proof: Suppose $\sum_{k=1}^{n} c_k \mathcal{X}_{E_k}(\omega) = 0$. Does it follow that $\sum_k c_k \mu(E_k) = 0$? The supposition implies

$$\sum_{k=1}^{n} \operatorname{Re} c_k \mathcal{X}_{E_k}(\omega) = 0, \ \sum_{k=1}^{n} \operatorname{Im} c_k \mathcal{X}_{E_k}(\omega) = 0. \tag{8.4}$$

Choose λ large and positive so that $\lambda + \operatorname{Re} c_k \ge 0$. Then adding $\sum_k \lambda \mathcal{X}_{E_k}$ to both sides of the first equation above,

$$\sum_{k=1}^{n} (\lambda + \operatorname{Re} c_k) \mathcal{X}_{E_k}(\omega) = \sum_{k=1}^{n} \lambda \mathcal{X}_{E_k}$$

and by Lemma 8.2.3 on Page 190, it follows upon taking $\int$ of both sides that

$$\sum_{k=1}^{n} (\lambda + \operatorname{Re} c_k)\, \mu(E_k) = \sum_{k=1}^{n} \lambda \mu(E_k)$$

which implies $\sum_{k=1}^{n} \operatorname{Re} c_k \mu(E_k) = 0$. Similarly,

$$\sum_{k=1}^{n} \operatorname{Im} c_k \mu(E_k) = 0$$

and so $\sum_{k=1}^{n} c_k \mu(E_k) = 0$. Thus if

$$\sum_j c_j \mathcal{X}_{E_j} = \sum_k d_k \mathcal{X}_{F_k}$$

then $\sum_j c_j \mathcal{X}_{E_j} + \sum_k (-d_k) \mathcal{X}_{F_k} = 0$ and so the result just established verifies

$$\sum_j c_j \mu(E_j) - \sum_k d_k \mu(F_k) = 0$$

which proves I is well defined.

That I is linear is now obvious. It only remains to verify the triangle inequality.

Let s be a simple function,

$$s = \sum_j c_j \mathcal{X}_{E_j}$$

Then pick $\theta \in \mathbb{C}$ such that $\theta I(s) = |I(s)|$ and $|\theta| = 1$. Then from the triangle inequality for sums of complex numbers,

$$\begin{aligned} |I(s)| &= \theta I(s) = I(\theta s) = \sum_j \theta c_j \mu(E_j) \\ &= \left| \sum_j \theta c_j \mu(E_j) \right| \leq \sum_j |\theta c_j| \, \mu(E_j) = I(|s|). \end{aligned}$$

■

Note that for any simple function $s = \sum_{k=1}^{n} c_k \mathcal{X}_{E_k}$ where $c_k > 0$, $\mu(E_k) < \infty$, it follows from Lemma 8.2.2 that $\int s d\mu = I(s)$ since they both equal $\sum_{k=1}^{n} c_k \mathcal{X}_{E_k}$.

With this lemma, the following is the definition of $L^1(\Omega)$.

Definition 8.7.4 *$f \in L^1(\Omega)$ means there exists a sequence of complex simple functions, $\{s_n\}$ such that*

$$\begin{gathered} s_n(\omega) \to f(\omega) \text{ for all } \omega \in \Omega \\ \lim_{m,n\to\infty} I(|s_n - s_m|) = \lim_{n,m\to\infty} \int |s_n - s_m| \, d\mu = 0 \end{gathered} \tag{8.5}$$

Then

$$I(f) \equiv \lim_{n\to\infty} I(s_n). \tag{8.6}$$

Lemma 8.7.5 *Definition 8.7.4 is well defined. Also $L^1(\Omega)$ is a vector space.*

Proof: There are several things which need to be verified. First suppose 8.5. Then by Lemma 8.7.3

$$|I(s_n) - I(s_m)| = |I(s_n - s_m)| \leq I(|s_n - s_m|)$$

and for m, n large enough, this last is given to be small so $\{I(s_n)\}$ is a Cauchy sequence in $\mathbb{C}$ and so it converges. This verifies the limit in 8.6 at least exists. It remains to consider another sequence $\{t_n\}$ having the same properties as $\{s_n\}$ and verifying $I(f)$ determined by this other sequence is the same. By Lemma 8.7.3 and Fatou's lemma, Theorem 8.5.1 on Page 192,

$$\begin{aligned} |I(s_n) - I(t_n)| &\leq I(|s_n - t_n|) = \int |s_n - t_n| \, d\mu \\ &\leq \int |s_n - f| + |f - t_n| \, d\mu \\ &\leq \liminf_{k\to\infty} \int |s_n - s_k| \, d\mu + \liminf_{k\to\infty} \int |t_n - t_k| \, d\mu < \varepsilon \end{aligned}$$

whenever n is large enough. Since ε is arbitrary, this shows the limit from using the t_n is the same as the limit from using s_n.

Why is $L^1(\Omega)$ a vector space? Let f, g be in $L^1(\Omega)$ and let $a, b \in \mathbb{C}$. Then let $\{s_n\}$ and $\{t_n\}$ be sequences of complex simple functions associated with f and g respectively as described in Definition 8.7.4. Consider $\{as_n + bt_n\}$, another sequence of complex simple functions. Then $as_n(\omega) + bt_n(\omega) \to af(\omega) + bg(\omega)$ for each ω. Also, from Theorem 8.6.1,

$$\int |as_n + bt_n - (as_m + bt_m)| \, d\mu \leq |a| \int |s_n - s_m| \, d\mu + |b| \int |t_n - t_m| \, d\mu$$

and the sum of the two terms on the right converge to zero as $m, n \to \infty$. Thus $af + bg \in L^1(\Omega)$. ■

Now here is another characterization for a function to be in $L^1(\Omega)$.

Corollary 8.7.6 *Let $(\Omega, \mathcal{S}, \mu)$ be a measure space and let $f : \Omega \to \mathbb{C}$. Then $f \in L^1(\Omega)$ if and only if f is measurable and $\int |f| \, d\mu < \infty$.*

Proof: First suppose $f \in L^1$. Then there exists a sequence $\{s_n\}$ of the sort described above attached to f. It follows that f is measurable because it is the limit of these measurable functions. Also for the same reasoning $|f| = \lim_{n\to\infty} |s_n|$ so $|f|$ is measurable as a real valued function. Now from I being linear,

$$\left| \int |s_n| \, d\mu - \int |s_m| \, d\mu \right| =$$

$$\begin{aligned} |I(|s_n|) - I(|s_m|)| &= |I(|s_n| - |s_m|)| \leq I(||s_n| - |s_m||) \\ &= \int ||s_n| - |s_m|| \, d\mu \leq \int |s_n - s_m| \, d\mu \end{aligned}$$

which is small whenever n, m are large. As to $\int |f| \, d\mu$ being finite, this follows from Fatou's lemma.

$$\int |f| \, d\mu \leq \liminf_{n\to\infty} \int |s_n| \, d\mu < \infty$$

Next suppose f is measurable and absolutely integrable. First suppose $f \geq 0$. Then by the approximation theorem involving simple functions, Theorem 7.7.12, there exists a sequence of nonnegative simple functions s_n which increases pointwise to f. Each of these must be nonzero only on a set of finite measure because $\int f d\mu < \infty$. Note that

$$\int 2f - (f - s_n) \, d\mu + \int f - s_n d\mu = \int 2f$$

and so

$$\int 2f - (f - s_n)\, d\mu = \int 2f d\mu - \int (f - s_n)\, d\mu$$

Then by the monotone convergence theorem,

$$\int 2f - (f - s_n)\, d\mu = \int 2f d\mu - \int (f - s_n)\, d\mu \to \int 2f$$

which shows that $\int |f - s_n|\, d\mu \to 0$. It follows that

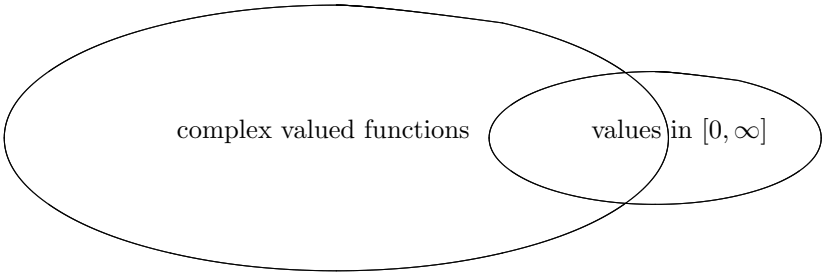
$$I(|s_n - s_m|) = \int |s_n - s_m|\, d\mu \leq \int |s_n - f|\, d\mu + \int |f - s_m|\, d\mu$$

both of which converge to 0. Thus there exists the right sort of sequence attached to f and this shows $f \in L^1(\Omega)$ as claimed. Now in the case where f has complex values, just write

$$f = \operatorname{Re} f^+ - \operatorname{Re} f^- + i\left(\operatorname{Im} f^+ - \operatorname{Im} f^-\right)$$

for $h^+ \equiv \frac{1}{2}(|h| + h)$ and $h^- \equiv \frac{1}{2}(|h| - h)$. Each of the above is nonnegative, measurable with finite integral and so from the above argument, each is in $L^1(\Omega)$ from what was just shown. Therefore, by Lemma 8.7.5 so is f. ■

Consider the following picture. I have just given a definition of an integral for functions having values in $\mathbb{C}$. However, $[0, \infty) \subseteq \mathbb{C}$.



What if f has values in $[0, \infty)$? Earlier $\int f d\mu$ was defined for such functions and now $I(f)$ has been defined. Are they the same? If so, I can be regarded as an extension of $\int d\mu$ to a larger class of functions.

Lemma 8.7.7 *Suppose f has values in $[0, \infty)$ and $f \in L^1(\Omega)$. Then f is measurable and*

$$I(f) = \int f d\mu.$$

Proof: Since f is the pointwise limit of a sequence of complex simple functions, $\{s_n\}$ having the properties described in Definition 8.7.4, it follows

$$f(\omega) = \lim_{n \to \infty} \operatorname{Re} s_n(\omega)$$

Also it is always the case that if a, b are real numbers,

$$\left|a^+ - b^+\right| \leq |a - b|$$

and so

$$\int \left|(\operatorname{Re} s_n)^+ - (\operatorname{Re} s_m)^+\right| d\mu \leq \int |\operatorname{Re} s_n - \operatorname{Re} s_m| \, d\mu \leq \int |s_n - s_m| \, d\mu$$

where $x^+ \equiv \frac{1}{2}(|x| + x)$, the positive part of the real number x. [1]Thus there is no loss of generality in assuming $\{s_n\}$ is a sequence of complex simple functions having values in $[0, \infty)$. By Corollary 8.7.6, $\int f d\mu < \infty$.

Therefore, there exists a nonnegative simple function $t \leq f$ such that

$$\int f d\mu \leq \int t d\mu + \varepsilon.$$

Then since, for such nonnegative complex simple functions, $I(s) = \int s d\mu$,

$$\begin{aligned} \left| I(f) - \int f d\mu \right| &\leq \left| I(f) - \int t d\mu \right| + \varepsilon \leq |I(f) - I(s_n)| \\ + \left| \int s_n d\mu - \int t d\mu \right| + \varepsilon &= |I(f) - I(s_n)| + |I(s_n) - I(t)| + \varepsilon \\ \leq \varepsilon + \int |s_n - t| \, d\mu + \varepsilon &\leq \varepsilon + \int |s_n - f| \, d\mu + \int |f - t| \, d\mu + \varepsilon \\ \leq 3\varepsilon + \lim_{k \to \infty} \inf \int |s_n - s_k| \, d\mu &< 4\varepsilon \end{aligned}$$

whenever n is large enough. Since ε is arbitrary, this shows $I(f) = \int f d\mu$ as claimed. ■

As explained above, I can be regarded as an extension of $\int d\mu$, so from now on, the usual symbol, $\int d\mu$ will be used. It is now easy to verify $\int d\mu$ is linear on the vector space $L^1(\Omega)$.

8.8 Approximation With Simple Functions

The next theorem says the integral as defined above is linear and also gives a way to compute the integral in terms of real and imaginary parts. In addition, functions in L^1 can be approximated with simple functions.

Theorem 8.8.1 *$\int d\mu$ is linear on $L^1(\Omega)$ and $L^1(\Omega)$ is a complex vector space. If $f \in L^1(\Omega)$, then $\operatorname{Re} f, \operatorname{Im} f$, and $|f|$ are all in $L^1(\Omega)$. Furthermore, for $f \in L^1(\Omega)$,*

$$\int f d\mu = \int (\operatorname{Re} f)^+ d\mu - \int (\operatorname{Re} f)^- d\mu + i\left(\int (\operatorname{Im} f)^+ d\mu - \int (\operatorname{Im} f)^- d\mu\right),$$

and the triangle inequality holds,

$$\left| \int f d\mu \right| \leq \int |f| \, d\mu$$

Also for every $f \in L^1(\Omega)$, for every $\varepsilon > 0$ there exists a simple function s such that

$$\int |f - s| \, d\mu < \varepsilon.$$

[1]The negative part of the real number x is defined to be $x^- \equiv \frac{1}{2}(|x| - x)$. Thus $|x| = x^+ + x^-$ and $x = x^+ - x^-$. .

Proof: Why is the integral linear? Let $\{s_n\}$ and $\{t_n\}$ be sequences of simple functions attached to f and g respectively according to the definition.

$$\begin{aligned}\int (af+bg)\,d\mu &\equiv \lim_{n\to\infty}\int (as_n+bt_n)\,d\mu\\ &= \lim_{n\to\infty}\left(a\int s_n d\mu + b\int t_n d\mu\right)\\ &= a\lim_{n\to\infty}\int s_n d\mu + b\lim_{n\to\infty}\int t_n d\mu\\ &= a\int f d\mu + b\int g d\mu.\end{aligned}$$

The fact that $\int$ is linear makes the triangle inequality easy to verify. Let $f\in L^1(\Omega)$ and let $\theta\in\mathbb{C}$ such that $|\theta|=1$ and $\theta\int f d\mu = \left|\int f d\mu\right|$. Then

$$\begin{aligned}\left|\int f d\mu\right| &= \int \theta f d\mu = \int \mathrm{Re}\,(\theta f)\,d\mu = \int \mathrm{Re}\,(\theta f)^+ - \mathrm{Re}\,(\theta f)^-\,d\mu\\ &\le \int \mathrm{Re}\,(\theta f)^+\,d\mu \le \int |\mathrm{Re}\,(\theta f)|\,d\mu \le \int |f|\,d\mu\end{aligned}$$

Now the last assertion follows from the definition. There exists a sequence of simple functions $\{s_n\}$ converging pointwise to f such that for all m,n large enough,

$$\frac{\varepsilon}{2} > \int |s_n - s_m|\,d\mu$$

Fix such an m and let $n\to\infty$. By Fatou's lemma

$$\varepsilon > \frac{\varepsilon}{2} \ge \lim\inf_{n\to\infty}\int |s_n-s_m|\,d\mu \ge \int |f-s_m|\,d\mu.$$

Let $s=s_m$. ■

One of the major theorems in this theory is the dominated convergence theorem. Before presenting it, here is a technical lemma about lim sup and lim inf .

Lemma 8.8.2 *Let $\{a_n\}$ be a sequence in $[-\infty,\infty]$. Then $\lim_{n\to\infty} a_n$ exists if and only if*

$$\lim\inf_{n\to\infty} a_n = \lim\sup_{n\to\infty} a_n$$

and in this case, the limit equals the common value of these two numbers.

Proof: Suppose first $\lim_{n\to\infty} a_n = a\in\mathbb{R}$. Then, letting $\varepsilon>0$ be given, $a_n\in(a-\varepsilon,a+\varepsilon)$ for all n large enough, say $n\ge N$. Therefore, both $\inf\{a_k : k\ge n\}$ and $\sup\{a_k : k\ge n\}$ are contained in $[a-\varepsilon,a+\varepsilon]$ whenever $n\ge N$. It follows $\lim\sup_{n\to\infty} a_n$ and $\lim\inf_{n\to\infty} a_n$ are both in $[a-\varepsilon,a+\varepsilon]$, showing

$$\left|\lim\inf_{n\to\infty} a_n - \lim\sup_{n\to\infty} a_n\right| < 2\varepsilon.$$

Since ε is arbitrary, the two must be equal and they both must equal a. Next suppose $\lim_{n\to\infty} a_n = \infty$. Then if $l\in\mathbb{R}$, there exists N such that for $n\ge N$,

$$l \le a_n$$

and therefore, for such n,

$$l \le \inf\{a_k : k\ge n\} \le \sup\{a_k : k\ge n\}$$

and this shows, since l is arbitrary that

$$\lim_{n\to\infty} \inf a_n = \lim_{n\to\infty} \sup a_n = \infty.$$

The case for $-\infty$ is similar.

Conversely, suppose $\liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n = a$. Suppose first that $a \in \mathbb{R}$. Then, letting $\varepsilon > 0$ be given, there exists N such that if $n \geq N$,

$$\sup\{a_k : k \geq n\} - \inf\{a_k : k \geq n\} < \varepsilon$$

therefore, if $k, m > N$, and $a_k > a_m$,

$$|a_k - a_m| = a_k - a_m \leq \sup\{a_k : k \geq n\} - \inf\{a_k : k \geq n\} < \varepsilon$$

showing that $\{a_n\}$ is a Cauchy sequence. Therefore, it converges to $a \in \mathbb{R}$, and as in the first part, the lim inf and lim sup both equal a. If $\liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n = \infty$, then given $l \in \mathbb{R}$, there exists N such that for $n \geq N$,

$$\inf_{n>N} a_n > l.$$

Therefore, $\lim_{n\to\infty} a_n = \infty$. The case for $-\infty$ is similar. This proves the lemma. ■

8.9 The Dominated Convergence Theorem

The dominated convergence theorem is one of the most important theorems in the theory of the integral. It is one of those big theorems which justifies the study of the Lebesgue integral.

Theorem 8.9.1 *(Dominated Convergence theorem) Let $f_n \in L^1(\Omega)$ and suppose*

$$f(\omega) = \lim_{n\to\infty} f_n(\omega),$$

and there exists a measurable function g, with values in $[0,\infty]$,[2] such that

$$|f_n(\omega)| \leq g(\omega) \text{ and } \int g(\omega)d\mu < \infty.$$

Then $f \in L^1(\Omega)$ and

$$0 = \lim_{n\to\infty} \int |f_n - f|\, d\mu = \lim_{n\to\infty} \left| \int f d\mu - \int f_n d\mu \right|$$

Proof: f is measurable by Theorem 7.7.3. Since $|f| \leq g$, it follows that

$$f \in L^1(\Omega) \text{ and } |f - f_n| \leq 2g.$$

By Fatou's lemma (Theorem 8.5.1),

$$\begin{aligned} \int 2g d\mu &\leq \lim_{n\to\infty} \inf \int 2g - |f - f_n| d\mu \\ &= \int 2g d\mu - \lim_{n\to\infty} \sup \int |f - f_n| d\mu. \end{aligned}$$

[2]Note that, since g is allowed to have the value ∞, it is not known that $g \in L^1(\Omega)$.

Subtracting $\int 2gd\mu$,

$$0 \leq -\limsup_{n\to\infty} \int |f-f_n| d\mu.$$

Hence

$$\begin{aligned} 0 &\geq \limsup_{n\to\infty} \left(\int |f-f_n| d\mu \right) \\ &\geq \liminf_{n\to\infty} \left(\int |f-f_n| d\mu \right) \geq \left| \int f d\mu - \int f_n d\mu \right| \geq 0. \end{aligned}$$

This proves the theorem by Lemma 8.8.2 because the lim sup and lim inf are equal. ■

Corollary 8.9.2 *Suppose $f_n \in L^1(\Omega)$ and $f(\omega) = \lim_{n\to\infty} f_n(\omega)$. Suppose also there exist measurable functions, g_n, g with values in $[0,\infty]$ such that $\lim_{n\to\infty} \int g_n d\mu = \int g d\mu$, $g_n(\omega) \to g(\omega)$ μ a.e. and both $\int g_n d\mu$ and $\int g d\mu$ are finite. Also suppose $|f_n(\omega)| \leq g_n(\omega)$. Then*

$$\lim_{n\to\infty} \int |f-f_n| \, d\mu = 0.$$

Proof: It is just like the above. This time $g + g_n - |f-f_n| \geq 0$ and so by Fatou's lemma,

$$\int 2g d\mu - \limsup_{n\to\infty} \int |f-f_n| \, d\mu =$$

$$\begin{aligned} &\liminf_{n\to\infty} \int (g_n + g) \, d\mu - \limsup_{n\to\infty} \int |f-f_n| \, d\mu \\ = \; &\liminf_{n\to\infty} \int ((g_n+g) - |f-f_n|) \, d\mu \geq \int 2g d\mu \end{aligned}$$

and so $-\limsup_{n\to\infty} \int |f-f_n| \, d\mu \geq 0$. Thus

$$\begin{aligned} 0 &\geq \limsup_{n\to\infty} \left(\int |f-f_n| d\mu \right) \\ &\geq \liminf_{n\to\infty} \left(\int |f-f_n| d\mu \right) \geq \left| \int f d\mu - \int f_n d\mu \right| \geq 0. \end{aligned}$$

This proves the corollary. ■

Definition 8.9.3

Let E be a measurable subset of Ω.

$$\int_E f d\mu \equiv \int f \mathcal{X}_E d\mu.$$

If $L^1(E)$ is written, the σ algebra is defined as

$$\{E \cap A : A \in \mathcal{F}\}$$

and the measure is μ restricted to this smaller σ algebra. Clearly, if $f \in L^1(\Omega)$, then

$$f\mathcal{X}_E \in L^1(E)$$

and if $f \in L^1(E)$, then letting $\tilde{f}$ be the 0 extension of f off of E, it follows $\tilde{f} \in L^1(\Omega)$.

8.10 Approximation With $C_c(Y)$

Let $(Y,\mathcal{F},\mu)$ be a measure space where Y is a closed subset of X a finite dimensional normed vector space and $\mathcal{F}\supseteq\mathcal{B}(Y)$, the Borel sets in Y. Also suppose that for every $E\in\mathcal{F}$, there are Borel sets H,G such that $\mu(G\setminus H)=0$ and $H\subseteq E\subseteq G$. This assumption of regularity will be tacitly assumed in what follows. Suppose also that $\mu(K)<\infty$ whenever K is a compact set in Y. By Theorem 7.4.6 it follows μ is regular. This regularity of μ implies an important approximation result valid for any $f\in L^1(Y)$. It turns out that in this situation, for all $\varepsilon>0$, there exists g a continuous function defined on Y with g equal to 0 outside some compact set and

$$\int|f-g|\,d\mu<\varepsilon.$$

Definition 8.10.1 *Let $\mathbf{f}:X\to Y$ where X is a normed vector space. Then the support of $\mathbf{f}$, denoted by $\operatorname{spt}(\mathbf{f})$ is the closure of the set where $\mathbf{f}$ is not equal to zero. Thus*

$$\operatorname{spt}(\mathbf{f})\equiv\overline{\{\mathbf{x}:\mathbf{f}(\mathbf{x})\neq\mathbf{0}\}}$$

Also, if U is an open set, $\mathbf{f}\in C_c(U)$ means $\mathbf{f}$ is continuous on U and $\operatorname{spt}(\mathbf{f})\subseteq U$. Similarly $\mathbf{f}\in C_c^m(U)$ if $\mathbf{f}$ has m continuous derivatives and $\operatorname{spt}(\mathbf{f})\subseteq U$ and $\mathbf{f}\in C_c^\infty(U)$ if $\operatorname{spt}(\mathbf{f})\subseteq U$ and $\mathbf{f}$ has continuous derivatives of every order on U.

Lemma 8.10.2 *Let Y be a closed subset of X a finite dimensional normed vector space. Let $K\subseteq V$ where K is compact in Y and V is open in Y. Then there exists a continuous function $f:Y\to[0,1]$ such that $\operatorname{spt}(f)\subseteq V$, $f(\mathbf{x})=1$ for all $\mathbf{x}\in K$. If $(Y,\mathcal{F},\mu)$ is a measure space with $\mu(K)<\infty$, for every compact K, then if $\mu(E)<\infty$ where $E\in\mathcal{F}$, there exists a sequence of functions in $C_c(Y)$ $\{f_k\}$ such that*

$$\lim_{k\to\infty}\int_Y|f_k(\mathbf{x})-\mathcal{X}_E(\mathbf{x})|\,d\mu=0.$$

Proof: For each $\mathbf{x}\in K$, there exists $r_{\mathbf{x}}$ such that

$$B(\mathbf{x},r_{\mathbf{x}})\equiv\{\mathbf{y}\in Y:||\mathbf{x}-\mathbf{y}||<r_{\mathbf{x}}\}\subseteq V.$$

Since K is compact, there are finitely many balls, $\{B(\mathbf{x}_k,r_{\mathbf{x}_k})\}_{k=1}^m$ which cover K. Let $W=\cup_{k=1}^m B(\mathbf{x}_k,r_{\mathbf{x}_k})$. Since there are only finitely many of these,

$$\overline{W}=\cup_{k=1}^m\overline{B(\mathbf{x},r_{\mathbf{x}_k})}$$

and $\overline{W}$ is a compact subset of V because it is closed and bounded, being the finite union of closed and bounded sets. Now define

$$f(\mathbf{x})\equiv\frac{\operatorname{dist}(\mathbf{x},W^C)}{\operatorname{dist}(\mathbf{x},W^C)+\operatorname{dist}(\mathbf{x},K)}$$

The denominator is never equal to 0 because if $\operatorname{dist}(\mathbf{x},K)=0$ then since K is closed, $\mathbf{x}\in K$ and so since $K\subseteq W$, an open set, $\operatorname{dist}(\mathbf{x},W^C)>0$. Therefore, f is continuous. When $\mathbf{x}\in K$, $f(\mathbf{x})=1$. If $\mathbf{x}\notin W$, then $f(\mathbf{x})=0$ and so $\operatorname{spt}(f)\subseteq\overline{W}\subseteq V$. In the above situation the following notation is often used.

$$K\prec f\prec V. \tag{8.7}$$

It remains to prove the last assertion. By Theorem 7.4.6, μ is regular and so there exist compact sets $\{K_k\}$ and open sets $\{V_k\}$ such that $V_k\supseteq V_{k+1}$, $K_k\subseteq K_{k+1}$ for all k, and

$$K_k\subseteq E\subseteq V_k,\ \mu(V_k\setminus K_k)<2^{-k}.$$

From the first part of the lemma, there exists a sequence $\{f_k\}$ such that

$$K_k \prec f_k \prec V_k.$$

Then $f_k(\mathbf{x})$ converges to $\mathcal{X}_E(\mathbf{x})$ a.e. because if convergence fails to take place, then $\mathbf{x}$ must be in infinitely many of the sets $V_k \setminus K_k$. Thus $\mathbf{x}$ is in

$$\cap_{m=1}^{\infty} \cup_{k=m}^{\infty} V_k \setminus K_k$$

and for each p

$$\begin{aligned} \mu\left(\cap_{m=1}^{\infty} \cup_{k=m}^{\infty} V_k \setminus K_k\right) &\leq \mu\left(\cup_{k=p}^{\infty} V_k \setminus K_k\right) \\ &\leq \sum_{k=p}^{\infty} \mu\left(V_k \setminus K_k\right) \\ &< \sum_{k=p}^{\infty} \frac{1}{2^k} \leq 2^{-(p-1)} \end{aligned}$$

Now the functions are all bounded above by 1 and below by 0 and are equal to zero off V_1, a set of finite measure so by the dominated convergence theorem,

$$\lim_{k\to\infty} \int \left|\mathcal{X}_E(\mathbf{x}) - f_k(\mathbf{x})\right| d\mu = 0,$$

the dominating function being $\mathcal{X}_E(\mathbf{x}) + \mathcal{X}_{V_1}(\mathbf{x})$. This proves the lemma. ■

With this lemma, here is an important major theorem.

Theorem 8.10.3 *Let Y be a closed subset of X a finite dimensional normed vector space. Let $(Y, \mathcal{F}, \mu)$ be a measure space with $\mathcal{F} \supseteq \mathcal{B}(Y)$ and $\mu(K) < \infty$, for every compact K in Y. Let $f \in L^1(Y)$ and let $\varepsilon > 0$ be given. Then there exists $g \in C_c(Y)$ such that*

$$\int_Y \left|f(\mathbf{x}) - g(\mathbf{x})\right| d\mu < \varepsilon.$$

Proof: By considering separately the positive and negative parts of the real and imaginary parts of f it suffices to consider only the case where $f \geq 0$. Then by Theorem 7.7.12 and the monotone convergence theorem, there exists a simple function,

$$s(\mathbf{x}) \equiv \sum_{m=1}^{p} c_m \mathcal{X}_{E_m}(\mathbf{x}),\ s(\mathbf{x}) \leq f(\mathbf{x})$$

such that

$$\int \left|f(\mathbf{x}) - s(\mathbf{x})\right| d\mu < \varepsilon/2.$$

By Lemma 8.10.2, there exists $\{h_{mk}\}_{k=1}^{\infty}$ be functions in $C_c(Y)$ such that

$$\lim_{k\to\infty} \int_Y \left|\mathcal{X}_{E_m} - f_{mk}\right| d\mu = 0.$$

Let

$$g_k(\mathbf{x}) \equiv \sum_{m=1}^{p} c_m h_{mk}.$$

Thus for k large enough,

$$\begin{aligned}\int |s(\mathbf{x}) - g_k(\mathbf{x})|\, d\mu &= \int \left| \sum_{m=1}^{p} c_m \left(\mathcal{X}_{E_m} - h_{mk}\right) \right| d\mu \\ &\leq \sum_{m=1}^{p} c_m \int |\mathcal{X}_{E_m} - h_{mk}|\, d\mu < \varepsilon/2\end{aligned}$$

Thus for k this large,

$$\begin{aligned}\int |f(\mathbf{x}) - g_k(\mathbf{x})|\, d\mu &\leq \int |f(\mathbf{x}) - s(\mathbf{x})|\, d\mu + \int |s(\mathbf{x}) - g_k(\mathbf{x})|\, d\mu \\ &< \varepsilon/2 + \varepsilon/2 = \varepsilon.\end{aligned}$$

This proves the theorem. ■

People think of this theorem as saying that f is approximated by g_k in $L^1(Y)$. It is customary to consider functions in $L^1(Y)$ as vectors and the norm of such a vector is given by

$$||f||_1 \equiv \int |f(\mathbf{x})|\, d\mu.$$

You should verify this mostly satisfies the axioms of a norm. The problem comes in asserting $f = 0$ if $||f|| = 0$ which strictly speaking is false. However, the other axioms of a norm do hold.

8.11 The One Dimensional Lebesgue Integral

Let F be an increasing function defined on $\mathbb{R}$. Let μ be the Lebesgue Stieltjes measure defined in Theorems 7.6.1 and 7.2.1. The conclusions of these theorems are reviewed here.

Theorem 8.11.1 *Let F be an increasing function defined on $\mathbb{R}$, an integrator function. There exists a function $\mu : \mathcal{P}(\mathbb{R}) \to [0, \infty]$ which satisfies the following properties.*

1. *If $A \subseteq B$, then $0 \leq \mu(A) \leq \mu(B), \mu(\emptyset) = 0$.*
2. *$\mu(\cup_{k=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$*
3. *$\mu([a,b]) = F(b+) - F(a-)$,*
4. *$\mu((a,b)) = F(b-) - F(a+)$*
5. *$\mu((a,b]) = F(b+) - F(a+)$*
6. *$\mu([a,b)) = F(b-) - F(a-)$ where*

$$F(b+) \equiv \lim_{t \to b+} F(t), F(b-) \equiv \lim_{t \to a-} F(t).$$

There also exists a σ algebra $\mathcal{S}$ of measurable sets on which μ is a measure which contains the open sets and also satisfies the regularity conditions,

$$\mu(E) = \sup\{\mu(K) : K \text{ is a closed and bounded set, } K \subseteq E\} \tag{8.8}$$

$$\mu(E) = \inf\{\mu(V) : V \text{ is an open set, } V \supseteq E\} \tag{8.9}$$

whenever E is a set in $\mathcal{S}$.

The Lebesgue integral taken with respect to this measure, is called the Lebesgue Stieltjes integral. Note that any real valued continuous function is measurable with respect to $\mathcal{S}$. This is because if f is continuous, inverse images of open sets are open and open sets are in $\mathcal{S}$. Thus f is measurable because $f^{-1}((a,b)) \in \mathcal{S}$. Similarly if f has complex values this argument applied to its real and imaginary parts yields the conclusion that f is measurable.

For f a continuous function, how does the Lebesgue Stieltjes integral compare with the Darboux Stieltjes integral? To answer this question, here is a technical lemma.

Lemma 8.11.2 *Let D be a countable subset of $\mathbb{R}$ and suppose $a, b \notin D$. Also suppose f is a continuous function defined on $[a,b]$. Then there exists a sequence of functions $\{s_n\}$ of the form*

$$s_n(x) \equiv \sum_{k=1}^{m_n} f\left(z_{k-1}^n\right) \mathcal{X}_{[z_{k-1}^n, z_k^n)}(x)$$

such that each $z_k^n \notin D$ and

$$\sup\{|s_n(x) - f(x)| : x \in [a,b]\} < 1/n.$$

Proof: First note that D contains no intervals. To see this let $D = \{d_k\}_{k=1}^{\infty}$. If D has an interval of length 2ε, let I_k be an interval centered at d_k which has length $\varepsilon/2^k$. Therefore, the sum of the lengths of these intervals is no more than

$$\sum_{k=1}^{\infty} \frac{\varepsilon}{2^k} = \varepsilon.$$

Thus D cannot contain an interval of length 2ε. Since ε is arbitrary, D cannot contain any interval.

Since f is continuous, it follows from Theorem 5.4.2 on Page 96 that f is uniformly continuous. Therefore, there exists $\delta > 0$ such that if $|x - y| \leq 3\delta$, then

$$|f(x) - f(y)| < 1/n$$

Now let $\{x_0, \cdots, x_{m_n}\}$ be a partition of $[a,b]$ such that $|x_i - x_{i-1}| < \delta$ for each i. For $k = 1, 2, \cdots, m_n - 1$, let $z_k^n \notin D$ and $|z_k^n - x_k| < \delta$. Then

$$\left|z_k^n - z_{k-1}^n\right| \leq |z_k^n - x_k| + |x_k - x_{k-1}| + \left|x_{k-1} - z_{k-1}^n\right| < 3\delta.$$

It follows that for each $x \in [a,b]$

$$\left|\sum_{k=1}^{m_n} f\left(z_{k-1}^n\right) \mathcal{X}_{[z_{k-1}^n, z_k^n)}(x) - f(x)\right| < 1/n.$$

This proves the lemma. ■

Proposition 8.11.3 *Let f be a continuous function defined on $\mathbb{R}$. Also let F be an increasing function defined on $\mathbb{R}$. Then whenever c, d are not points of discontinuity of F and $[a,b] \supseteq [c,d]$,*

$$\int_a^b f \mathcal{X}_{[c,d]} dF = \int f d\mu$$

Here μ is the Lebesgue Stieltjes measure defined above.

Proof: Since F is an increasing function it can have only countably many discontinuities. The reason for this is that the only kind of discontinuity it can have is where $F(x+) > F(x-)$. Now since F is increasing, the intervals $(F(x-), F(x+))$ for x a point of discontinuity are disjoint and so since each must contain a rational number and the rational numbers are countable, and therefore so are these intervals.

Let D denote this countable set of discontinuities of F. Then if $l, r \notin D, [l,r] \subseteq [a,b]$, it follows quickly from the definition of the Darboux Stieltjes integral that

$$\begin{aligned}\int_a^b \mathcal{X}_{[l,r)} dF &= F(r) - F(l) = F(r-) - F(l-) \\ &= \mu([l,r)) = \int \mathcal{X}_{[l,r)} d\mu.\end{aligned}$$

Now let $\{s_n\}$ be the sequence of step functions of Lemma 8.11.2 such that these step functions converge uniformly to f on $[c,d]$ Then

$$\left| \int \left(\mathcal{X}_{[c,d]} f - \mathcal{X}_{[c,d]} s_n \right) d\mu \right| \leq \int \left| \mathcal{X}_{[c,d]} (f - s_n) \right| d\mu \leq \frac{1}{n} \mu([c,d])$$

and

$$\left| \int_a^b \left(\mathcal{X}_{[c,d]} f - \mathcal{X}_{[c,d]} s_n \right) dF \right| \leq \int_a^b \mathcal{X}_{[c,d]} |f - s_n| \, dF < \frac{1}{n} (F(b) - F(a)).$$

Also if s_n is given by the formula of Lemma 8.11.2,

$$\begin{aligned}\int \mathcal{X}_{[c,d]} s_n d\mu &= \int \sum_{k=1}^{m_n} f\left(z_{k-1}^n\right) \mathcal{X}_{[z_{k-1}^n, z_k^n)} d\mu \\ &= \sum_{k=1}^{m_n} \int f\left(z_{k-1}^n\right) \mathcal{X}_{[z_{k-1}^n, z_k^n)} d\mu \\ &= \sum_{k=1}^{m_n} f\left(z_{k-1}^n\right) \mu\left([z_{k-1}^n, z_k^n)\right) \\ &= \sum_{k=1}^{m_n} f\left(z_{k-1}^n\right) \left(F(z_k^n-) - F\left(z_{k-1}^n-\right)\right) \\ &= \sum_{k=1}^{m_n} f\left(z_{k-1}^n\right) \left(F(z_k^n) - F\left(z_{k-1}^n\right)\right) \\ &= \sum_{k=1}^{m_n} \int_a^b f\left(z_{k-1}^n\right) \mathcal{X}_{[z_{k-1}^n, z_k^n)} dF = \int_a^b s_n dF.\end{aligned}$$

Therefore,

$$\begin{aligned}&\left| \int \mathcal{X}_{[c,d]} f d\mu - \int_a^b \mathcal{X}_{[c,d]} f dF \right| \\ &\leq \left| \int \mathcal{X}_{[c,d]} f d\mu - \int \mathcal{X}_{[c,d]} s_n d\mu \right| \\ &+ \left| \int \mathcal{X}_{[c,d]} s_n d\mu - \int_a^b s_n dF \right| + \left| \int_a^b s_n dF - \int_a^b \mathcal{X}_{[c,d]} f dF \right|\end{aligned}$$

$$\leq \frac{1}{n}\mu([c,d]) + \frac{1}{n}(F(b) - F(a))$$

and since n is arbitrary, this shows

$$\int f d\mu - \int_a^b f dF = 0.$$

This proves the theorem. ■

In particular, in the special case where F is continuous and f is continuous,

$$\int_a^b f dF = \int \mathcal{X}_{[a,b]} f d\mu.$$

Thus, if $F(x) = x$ so the Darboux Stieltjes integral is the usual integral from calculus,

$$\int_a^b f(t)\, dt = \int \mathcal{X}_{[a,b]} f d\mu$$

where μ is the measure which comes from $F(x) = x$ as described above. This measure is often denoted by m. Thus when f is continuous

$$\int_a^b f(t)\, dt = \int \mathcal{X}_{[a,b]} f dm$$

and so there is no problem in writing

$$\int_a^b f(t)\, dt$$

for either the Lebesgue or the Riemann integral. Furthermore, when f is continuous, you can compute the Lebesgue integral by using the fundamental theorem of calculus because in this case, the two integrals are equal.

8.12 Exercises

1. Let $\Omega = \mathbb{N} = \{1, 2, \cdots\}$. Let $\mathcal{F} = \mathcal{P}(\mathbb{N})$, the set of all subsets of $\mathbb{N}$, and let $\mu(S) =$ number of elements in S. Thus $\mu(\{1\}) = 1 = \mu(\{2\})$, $\mu(\{1,2\}) = 2$, etc. Show $(\Omega, \mathcal{F}, \mu)$ is a measure space. It is called counting measure. What functions are measurable in this case? For a nonnegative function, f defined on $\mathbb{N}$, show

$$\int_{\mathbb{N}} f d\mu = \sum_{k=1}^{\infty} f(k)$$

What do the monotone convergence and dominated convergence theorems say about this example?

2. For the measure space of Problem 1, give an example of a sequence of nonnegative measurable functions $\{f_n\}$ converging pointwise to a function f, such that inequality is obtained in Fatou's lemma.

3. If $(\Omega, \mathcal{F}, \mu)$ is a measure space and $f \geq 0$ is measurable, show that if $g(\omega) = f(\omega)$ a.e. ω and $g \geq 0$, then $\int g d\mu = \int f d\mu$. Show that if $f, g \in L^1(\Omega)$ and $g(\omega) = f(\omega)$ a.e. then $\int g d\mu = \int f d\mu$.

4. An algebra $\mathcal{A}$ of subsets of Ω is a subset of the power set such that Ω is in the algebra and for $A, B \in \mathcal{A}$, $A \setminus B$ and $A \cup B$ are both in $\mathcal{A}$. Let $\mathcal{C} \equiv \{E_i\}_{i=1}^{\infty}$ be a countable collection of sets and let $\Omega_1 \equiv \cup_{i=1}^{\infty} E_i$. Show there exists an algebra of sets $\mathcal{A}$, such that $\mathcal{A} \supseteq \mathcal{C}$ and $\mathcal{A}$ is countable. Note the difference between this problem and Problem 5. **Hint:** Let $\mathcal{C}_1$ denote all finite unions of sets of $\mathcal{C}$ and Ω_1. Thus $\mathcal{C}_1$ is countable. Now let $\mathcal{B}_1$ denote all complements with respect to Ω_1 of sets of $\mathcal{C}_1$. Let $\mathcal{C}_2$ denote all finite unions of sets of $\mathcal{B}_1 \cup \mathcal{C}_1$. Continue in this way, obtaining an increasing sequence $\mathcal{C}_n$, each of which is countable. Let

$$\mathcal{A} \equiv \cup_{i=1}^{\infty} \mathcal{C}_i.$$

5. Let $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ where $\mathcal{P}(\Omega)$ denotes the set of all subsets of Ω. Let $\sigma(\mathcal{A})$ denote the intersection of all σ algebras which contain $\mathcal{A}$, one of these being $\mathcal{P}(\Omega)$. Show $\sigma(\mathcal{A})$ is also a σ algebra.

6. We say a function g mapping a normed vector space, Ω to a normed vector space is Borel measurable if whenever U is open, $g^{-1}(U)$ is a Borel set. (The Borel sets are those sets in the smallest σ algebra which contains the open sets.) Let $f : \Omega \to X$ and let $g : X \to Y$ where X is a normed vector space and Y equals $\mathbb{C}, \mathbb{R}$, or $(-\infty, \infty]$ and $\mathcal{F}$ is a σ algebra of sets of Ω. Suppose f is measurable and g is Borel measurable. Show $g \circ f$ is measurable.

7. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Define $\overline{\mu} : \mathcal{P}(\Omega) \to [0, \infty]$ by

$$\overline{\mu}(A) = \inf\{\mu(B) : B \supseteq A,\ B \in \mathcal{F}\}.$$

Show $\overline{\mu}$ satisfies

$$\begin{aligned} \overline{\mu}(\emptyset) &= 0, \text{ if } A \subseteq B,\ \overline{\mu}(A) \leq \overline{\mu}(B), \\ \overline{\mu}(\cup_{i=1}^{\infty} A_i) &\leq \sum_{i=1}^{\infty} \overline{\mu}(A_i),\ \mu(A) = \overline{\mu}(A) \text{ if } A \in \mathcal{F}. \end{aligned}$$

If $\overline{\mu}$ satisfies these conditions, it is called an outer measure. This shows every measure determines an outer measure on the power set.

8. Let $\{E_i\}$ be a sequence of measurable sets with the property that

$$\sum_{i=1}^{\infty} \mu(E_i) < \infty.$$

Let $S = \{\omega \in \Omega \text{ such that } \omega \in E_i \text{ for infinitely many values of } i\}$. Show $\mu(S) = 0$ and S is measurable. This is part of the Borel Cantelli lemma. **Hint:** Write S in terms of intersections and unions. Something is in S means that for every n there exists $k > n$ such that it is in E_k. Remember the tail of a convergent series is small.

9. $\uparrow$ Let $\{f_n\}, f$ be measurable functions with values in $\mathbb{C}$. $\{f_n\}$ converges in measure if

$$\lim_{n \to \infty} \mu(x \in \Omega : |f(x) - f_n(x)| \geq \varepsilon) = 0$$

for each fixed $\varepsilon > 0$. Prove the theorem of F. Riesz. If f_n converges to f in measure, then there exists a subsequence $\{f_{n_k}\}$ which converges to f a.e. **Hint:** Choose n_1 such that

$$\mu(x : |f(x) - f_{n_1}(x)| \geq 1) < 1/2.$$

Choose $n_2 > n_1$ such that

$$\mu(x : |f(x) - f_{n_2}(x)| \geq 1/2) < 1/2^2,$$

$n_3 > n_2$ such that

$$\mu(x : |f(x) - f_{n_3}(x)| \geq 1/3) < 1/2^3,$$

etc. Now consider what it means for $f_{n_k}(x)$ to fail to converge to $f(x)$. Then use Problem 8.

10. Suppose (Ω, μ) is a finite measure space ($\mu(\Omega) < \infty$) and $\mathfrak{S} \subseteq L^1(\Omega)$. Then $\mathfrak{S}$ is said to be uniformly integrable if for every $\varepsilon > 0$ there exists $\delta > 0$ such that if E is a measurable set satisfying $\mu(E) < \delta$, then

$$\int_E |f| \, d\mu < \varepsilon$$

for all $f \in \mathfrak{S}$. Show $\mathfrak{S}$ is uniformly integrable and bounded in $L^1(\Omega)$ if there exists an increasing function h which satisfies

$$\lim_{t\to\infty} \frac{h(t)}{t} = \infty, \ \sup\left\{\int_\Omega h(|f|) \, d\mu : f \in \mathfrak{S}\right\} < \infty.$$

$\mathfrak{S}$ is bounded if there is some number, M such that

$$\int |f| \, d\mu \leq M$$

for all $f \in \mathfrak{S}$.

11. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and suppose $f, g : \Omega \to (-\infty, \infty]$ are measurable. Prove the sets

$$\{\omega : f(\omega) < g(\omega)\} \text{ and } \{\omega : f(\omega) = g(\omega)\}$$

are measurable. **Hint:** The easy way to do this is to write

$$\{\omega : f(\omega) < g(\omega)\} = \cup_{r\in\mathbb{Q}} [f < r] \cap [g > r].$$

Note that $l(x, y) = x - y$ is not continuous on $(-\infty, \infty]$ so the obvious idea doesn't work.

12. Let $\{f_n\}$ be a sequence of real or complex valued measurable functions. Let

$$S = \{\omega : \{f_n(\omega)\} \text{ converges}\}.$$

Show S is measurable. **Hint:** You might try to exhibit the set where f_n converges in terms of countable unions and intersections using the definition of a Cauchy sequence.

13. Suppose $u_n(t)$ is a differentiable function for $t \in (a, b)$ and suppose that for $t \in (a, b)$,

$$|u_n(t)|, \ |u_n'(t)| < K_n$$

where $\sum_{n=1}^\infty K_n < \infty$. Show

$$(\sum_{n=1}^\infty u_n(t))' = \sum_{n=1}^\infty u_n'(t).$$

Hint: This is an exercise in the use of the dominated convergence theorem and the mean value theorem.

14. Let E be a countable subset of $\mathbb{R}$. Show $m(E) = 0$. **Hint:** Let the set be $\{e_i\}_{i=1}^{\infty}$ and let e_i be the center of an open interval of length $\varepsilon/2^i$.

15. $\uparrow$ If S is an uncountable set of irrational numbers, is it necessary that S has a rational number as a limit point? **Hint:** Consider the proof of Problem 14 when applied to the rational numbers. (This problem was shown to me by Lee Erlebach.)

16. Suppose $\{f_n\}$ is a sequence of nonnegative measurable functions defined on a measure space, $(\Omega, \mathcal{S}, \mu)$. Show that

$$\int \sum_{k=1}^{\infty} f_k d\mu = \sum_{k=1}^{\infty} \int f_k d\mu.$$

Hint: Use the monotone convergence theorem along with the fact the integral is linear.

17. The integral $\int_{-\infty}^{\infty} f(t)\, dt$ will denote the Lebesgue integral taken with respect to one dimensional Lebesgue measure as discussed earlier. Show that for $\alpha > 0, t \to e^{-at^2}$ is in $L^1(\mathbb{R})$. The gamma function is defined for $x > 0$ as

$$\Gamma(x) \equiv \int_0^{\infty} e^{-t} t^{x-1} dt$$

Show $t \to e^{-t} t^{x-1}$ is in $L^1(\mathbb{R})$ for all $x > 0$. Also show that

$$\Gamma(x+1) = x\Gamma(x),\ \Gamma(1) = 1.$$

How does $\Gamma(n)$ for n an integer compare with $(n-1)!$?

18. This problem outlines a treatment of Stirling's formula which is a very useful approximation to $n!$ based on a section in [34]. It is an excellent application of the monotone convergence theorem. Follow and justify the following steps using the convergence theorems for the Lebesgue integral as needed. Here $x > 0$.

$$\Gamma(x+1) = \int_0^{\infty} e^{-t} t^x dt$$

First change the variables letting $t = x(1+u)$ to get $\Gamma(x+1) =$

$$e^{-x} x^{x+1} \int_{-1}^{\infty} \left(e^{-u}(1+u)\right)^x du$$

Next make the change of variables $u = s\sqrt{\frac{2}{x}}$ to obtain $\Gamma(x+1) =$

$$\sqrt{2} e^{-x} x^{x+(1/2)} \int_{-\sqrt{\frac{x}{2}}}^{\infty} \left(e^{-s\sqrt{\frac{2}{x}}} \left(1 + s\sqrt{\frac{2}{x}}\right)\right)^x ds$$

The integrand is increasing in x. This is most easily seen by taking ln of the integrand and then taking the derivative with respect to x. This derivative is positive. Next show the limit of the integrand as $x \to \infty$ is e^{-s^2}. This isn't too bad if you take ln and then use L'Hospital's rule. Consider the integral. Explain why it must be increasing in x. Next justify the following assertion. Remember the monotone convergence theorem applies to a sequence of functions.

$$\lim_{x\to\infty} \int_{-\sqrt{\frac{x}{2}}}^{\infty} \left(e^{-s\sqrt{\frac{2}{x}}} \left(1 + s\sqrt{\frac{2}{x}}\right)\right)^x ds = \int_{-\infty}^{\infty} e^{-s^2} ds$$

Now Stirling's formula is

$$\lim_{x\to\infty}\frac{\Gamma(x+1)}{\sqrt{2}e^{-x}x^{x+(1/2)}}=\int_{-\infty}^{\infty}e^{-s^2}ds$$

where this last improper integral equals a well defined constant (why?). It is very easy, when you know something about multiple integrals of functions of more than one variable to verify this constant is $\sqrt{\pi}$ but the necessary mathematical machinery has not yet been presented. It can also be done through much more difficult arguments in the context of functions of only one variable. See [34] for these clever arguments.

19. To show you the power of Stirling's formula, find whether the series

$$\sum_{n=1}^{\infty}\frac{n!e^n}{n^n}$$

converges. The ratio test falls flat but you can try it if you like. Now explain why, if n is large enough

$$n!\geq\frac{1}{2}\left(\int_{-\infty}^{\infty}e^{-s^2}ds\right)\sqrt{2}e^{-n}n^{n+(1/2)}\equiv c\sqrt{2}e^{-n}n^{n+(1/2)}.$$

Use this.

20. The Riemann integral is only defined for functions which are bounded which are also defined on a bounded interval. If either of these two criteria are not satisfied, then the integral is not the Riemann integral. Suppose f is Riemann integrable on a bounded interval, $[a,b]$. Show that it must also be Lebesgue integrable with respect to one dimensional Lebesgue measure and the two integrals coincide.

21. Give a theorem in which the improper Riemann integral coincides with a suitable Lebesgue integral. (There are many such situations just find one.)

22. Note that $\int_0^{\infty}\frac{\sin x}{x}dx$ is a valid improper Riemann integral defined by

$$\lim_{R\to\infty}\int_0^R\frac{\sin x}{x}dx$$

but this function, $\sin x/x$ is not in $L^1([0,\infty))$. Why?

23. Let f be a nonnegative strictly decreasing function defined on $[0,\infty)$. For $0\leq y\leq f(0)$, let $f^{-1}(y)=x$ where $y\in[f(x+),f(x-)]$. (Draw a picture. f could have jump discontinuities.) Show that f^{-1} is nonincreasing and that

$$\int_0^{\infty}f(t)\,dt=\int_0^{f(0)}f^{-1}(y)\,dy.$$

Hint: Use the distribution function description.

24. Consider the following nested sequence of compact sets $\{P_n\}$. We let $P_1=[0,1]$, $P_2=\left[0,\frac{1}{3}\right]\cup\left[\frac{2}{3},1\right]$, etc. To go from P_n to P_{n+1}, delete the open interval which is the middle third of each closed interval in P_n. Let $P=\cap_{n=1}^{\infty}P_n$. Since P is the intersection of nested nonempty compact sets, it follows from advanced calculus that $P\neq\emptyset$. Show $m(P)=0$. Show there is a one to one onto mapping of $[0,1]$ to P. The set P is called the Cantor set. Thus, although P has measure zero, it has the same number of points in it as $[0,1]$ in the sense that there is a one to one and onto mapping from one to the other. **Hint:** There are various ways of doing this last part but the most enlightenment is obtained by exploiting the construction of the Cantor set.

25. ↑ Consider the sequence of functions defined in the following way. Let $f_1(x) = x$ on $[0,1]$. To get from f_n to f_{n+1}, let $f_{n+1} = f_n$ on all intervals where f_n is constant. If f_n is nonconstant on $[a,b]$, let $f_{n+1}(a) = f_n(a)$, $f_{n+1}(b) = f_n(b)$, f_{n+1} is piecewise linear and equal to $\frac{1}{2}(f_n(a) + f_n(b))$ on the middle third of $[a,b]$. Sketch a few of these and you will see the pattern. The process of modifying a nonconstant section of the graph of this function is illustrated in the following picture.

Show $\{f_n\}$ converges uniformly on $[0,1]$. If $f(x) = \lim_{n\to\infty} f_n(x)$, show that $f(0) = 0$, $f(1) = 1$, f is continuous, and $f'(x) = 0$ for all $x \notin P$ where P is the Cantor set. This function is called the Cantor function.It is a very important example to remember. Note it has derivative equal to zero a.e. and yet it succeeds in climbing from 0 to 1. Thus

$$\int_0^1 f'(t)\,dt = 0 \neq f(1) - f(0).$$

Is this somehow contradictory to the fundamental theorem of calculus? **Hint:** This isn't too hard if you focus on getting a careful estimate on the difference between two successive functions in the list considering only a typical small interval in which the change takes place. The above picture should be helpful.

26. Let $m(W) > 0$, W is measurable, $W \subseteq [a,b]$. Show there exists a nonmeasurable subset of W. **Hint**: Let $x \sim y$ if $x - y \in \mathbb{Q}$. Observe that $\sim$ is an equivalence relation on $\mathbb{R}$. See Definition 2.1.9 on Page 17 for a review of this terminology. Let $\mathcal{C}$ be the set of equivalence classes and let $\mathcal{D} \equiv \{C \cap W : C \in \mathcal{C} \text{ and } C \cap W \neq \emptyset\}$. By the axiom of choice, there exists a set, A, consisting of exactly one point from each of the nonempty sets which are the elements of $\mathcal{D}$. Show

$$W \subseteq \cup_{r\in\mathbb{Q}} A + r \tag{a.}$$

$$A + r_1 \cap A + r_2 = \emptyset \text{ if } r_1 \neq r_2,\ r_i \in \mathbb{Q}. \tag{b.}$$

Observe that since $A \subseteq [a,b]$, then $A + r \subseteq [a-1, b+1]$ whenever $|r| < 1$. Use this to show that if $m(A) = 0$, or if $m(A) > 0$ a contradiction results.Show there exists some set, S such that $\overline{m}(S) < \overline{m}(S \cap A) + \overline{m}(S \setminus A)$ where $\overline{m}$ is the outer measure determined by m.

27. ↑ This problem gives a very interesting example found in the book by McShane [31]. Let $g(x) = x + f(x)$ where f is the strange function of Problem 25. Let P be the Cantor set of Problem 24. Let $[0,1] \setminus P = \cup_{j=1}^{\infty} I_j$ where I_j is open and $I_j \cap I_k = \emptyset$ if $j \neq k$. These intervals are the connected components of the complement of the Cantor set. Show $m(g(I_j)) = m(I_j)$ so

$$m(g(\cup_{j=1}^{\infty} I_j)) = \sum_{j=1}^{\infty} m(g(I_j)) = \sum_{j=1}^{\infty} m(I_j) = 1.$$

Thus $m(g(P)) = 1$ because $g([0,1]) = [0,2]$. By Problem 26 there exists a set, $A \subseteq g(P)$ which is non measurable. Define $\phi(x) = \mathcal{X}_A(g(x))$. Thus $\phi(x) = 0$ unless $x \in P$. Tell why ϕ is measurable. (Recall $m(P) = 0$ and Lebesgue measure is complete.) Now show that $\mathcal{X}_A(y) = \phi(g^{-1}(y))$ for $y \in [0,2]$. Tell why g^{-1} is

continuous but $\phi \circ g^{-1}$ is not measurable. (This is an example of measurable $\circ$ continuous $\neq$ measurable.) Show there exist Lebesgue measurable sets which are not Borel measurable. **Hint**: The function, ϕ is Lebesgue measurable. Now show that Borel $\circ$ measurable = measurable.

28. If A is $m \lfloor S$ measurable, it does not follow that A is m measurable. Give an example to show this is the case.

29. If f is a nonnegative Lebesgue measurable function, show there exists g a Borel measurable function such that $g(x) = f(x)$ a.e.

Chapter 9

The Lebesgue Integral For Functions Of p Variables

9.1 π Systems

The approach to p dimensional Lebesgue measure will be based on a very elegant idea due to Dynkin.

Definition 9.1.1 *Let Ω be a set and let $\mathcal{K}$ be a collection of subsets of Ω. Then $\mathcal{K}$ is called a π system if $\emptyset, \Omega \in \mathcal{K}$ and whenever $A, B \in \mathcal{K}$, it follows $A \cap B \in \mathcal{K}$.*

For example, if $\mathbb{R}^p = \Omega$, an example of a π system would be the set of all open sets. Another example would be sets of the form $\prod_{k=1}^{p} A_k$ where A_k is a Lebesgue measurable set.

The following is the fundamental lemma which shows these π systems are useful.

Lemma 9.1.2 *Let $\mathcal{K}$ be a π system of subsets of Ω, a set. Also let $\mathcal{G}$ be a collection of subsets of Ω which satisfies the following three properties.*

1. *$\mathcal{K} \subseteq \mathcal{G}$*
2. *If $A \in \mathcal{G}$, then $A^C \in \mathcal{G}$*
3. *If $\{A_i\}_{i=1}^{\infty}$ is a sequence of disjoint sets from $\mathcal{G}$ then $\cup_{i=1}^{\infty} A_i \in \mathcal{G}$.*

Then $\mathcal{G} \supseteq \sigma(\mathcal{K})$, where $\sigma(\mathcal{K})$ is the smallest σ algebra which contains $\mathcal{K}$.

Proof: First note that if

$$\mathcal{H} \equiv \{\mathcal{G} : 1 \text{ - } 3 \text{ all hold}\}$$

then $\cap\mathcal{H}$ yields a collection of sets which also satisfies 1 - 3. Therefore, I will assume in the argument that $\mathcal{G}$ is the smallest collection of sets satisfying 1 - 3, the intersection of all such collections. Let $A \in \mathcal{K}$ and define

$$\mathcal{G}_A \equiv \{B \in \mathcal{G} : A \cap B \in \mathcal{G}\}.$$

I want to show $\mathcal{G}_A$ satisfies 1 - 3 because then it must equal $\mathcal{G}$ since $\mathcal{G}$ is the smallest collection of subsets of Ω which satisfies 1 - 3. This will give the conclusion that for $A \in \mathcal{K}$ and $B \in \mathcal{G}$, $A \cap B \in \mathcal{G}$. This information will then be used to show that if

$A, B \in \mathcal{G}$ then $A \cap B \in \mathcal{G}$. From this it will follow very easily that $\mathcal{G}$ is a σ algebra which will imply it contains $\sigma(\mathcal{K})$. Now here are the details of the argument.

Since $\mathcal{K}$ is given to be a π system, $\mathcal{K} \subseteq \mathcal{G}_A$. Property 3 is obvious because if $\{B_i\}$ is a sequence of disjoint sets in $\mathcal{G}_A$, then

$$A \cap \cup_{i=1}^{\infty} B_i = \cup_{i=1}^{\infty} A \cap B_i \in \mathcal{G}$$

because $A \cap B_i \in \mathcal{G}$ and the property 3 of $\mathcal{G}$.

It remains to verify Property 2 so let $B \in \mathcal{G}_A$. I need to verify that $B^C \in \mathcal{G}_A$. In other words, I need to show that $A \cap B^C \in \mathcal{G}$. However,

$$A \cap B^C = \left(A^C \cup (A \cap B)\right)^C \in \mathcal{G}$$

Here is why. Since $B \in \mathcal{G}_A$, $A \cap B \in \mathcal{G}$ and since $A \in \mathcal{K} \subseteq \mathcal{G}$ it follows $A^C \in \mathcal{G}$. It follows the union of the disjoint sets A^C and $(A \cap B)$ is in $\mathcal{G}$ and then from 2 the complement of their union is in $\mathcal{G}$. Thus $\mathcal{G}_A$ satisfies 1 - 3 and this implies since $\mathcal{G}$ is the smallest such, that $\mathcal{G}_A \supseteq \mathcal{G}$. However, $\mathcal{G}_A$ is constructed as a subset of $\mathcal{G}$ and so $\mathcal{G} = \mathcal{G}_A$. This proves that for every $B \in \mathcal{G}$ and $A \in \mathcal{K}$, $A \cap B \in \mathcal{G}$. Now pick $B \in \mathcal{G}$ and consider

$$\mathcal{G}_B \equiv \{A \in \mathcal{G} : A \cap B \in \mathcal{G}\}.$$

I just proved $\mathcal{K} \subseteq \mathcal{G}_B$. The other arguments are identical to show $\mathcal{G}_B$ satisfies 1 - 3 and is therefore equal to $\mathcal{G}$. This shows that whenever $A, B \in \mathcal{G}$ it follows $A \cap B \in \mathcal{G}$.

This implies $\mathcal{G}$ is a σ algebra. To show this, all that is left is to verify $\mathcal{G}$ is closed under countable unions because then it follows $\mathcal{G}$ is a σ algebra. Let $\{A_i\} \subseteq \mathcal{G}$. Then let $A_1' = A_1$ and

$$\begin{aligned} A_{n+1}' &\equiv A_{n+1} \setminus (\cup_{i=1}^{n} A_i) \\ &= A_{n+1} \cap \left(\cap_{i=1}^{n} A_i^C\right) \\ &= \cap_{i=1}^{n} \left(A_{n+1} \cap A_i^C\right) \in \mathcal{G} \end{aligned}$$

because finite intersections of sets of $\mathcal{G}$ are in $\mathcal{G}$. Since the A_i' are disjoint, it follows

$$\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} A_i' \in \mathcal{G}$$

Therefore, $\mathcal{G} \supseteq \sigma(\mathcal{K})$ because it is a σ algebra which contains $\mathcal{K}$ and This proves the lemma. ■

9.2 p Dimensional Lebesgue Measure And Integrals

9.2.1 Iterated Integrals

Let m denote one dimensional Lebesgue measure. That is, it is the Lebesgue Stieltjes measure which comes from the integrator function, $F(x) = x$. Also let the σ algebra of measurable sets be denoted by $\mathcal{F}$. Recall this σ algebra contained the open sets. Also from the construction given above,

$$m([a,b]) = m((a,b)) = b - a$$

Definition 9.2.1 *Let f be a function of p variables and consider the symbol*

$$\int \cdots \int f(x_1, \cdots, x_p)\, dx_{i_1} \cdots dx_{i_p}. \tag{9.1}$$

where $(i_1, \cdots, i_p)$ is a permutation of the integers $\{1, 2, \cdots, p\}$. The symbol means to first do the Lebesgue integral

$$\int f(x_1, \cdots, x_p)\, dx_{i_1}$$

yielding a function of the other $p-1$ variables given above. Then you do

$$\int \left(\int f(x_1, \cdots, x_p)\, dx_{i_1} \right) dx_{i_2}$$

and continue this way. The iterated integral is said to make sense if the process just described makes sense at each step. Thus, to make sense, it is required

$$x_{i_1} \to f(x_1, \cdots, x_p)$$

can be integrated. Either the function has values in $[0, \infty]$ and is measurable or it is a function in L^1. Then it is required

$$x_{i_2} \to \int f(x_1, \cdots, x_p)\, dx_{i_1}$$

can be integrated and so forth. The symbol in 9.1 is called an iterated integral.

With the above explanation of iterated integrals, it is now time to define p dimensional Lebesgue measure.

9.2.2 p Dimensional Lebesgue Measure And Integrals

With the Lemma about π systems given above and the monotone convergence theorem, it is possible to give a very elegant and fairly easy definition of the Lebesgue integral of a function of p real variables. This is done in the following proposition.

Proposition 9.2.2 *There exists a σ algebra of sets of $\mathbb{R}^p$ which contains the open sets, $\mathcal{F}^p$ and a measure m_p defined on this σ algebra such that if $f : \mathbb{R}^p \to [0, \infty)$ is measurable with respect to $\mathcal{F}^p$ then for any permutation $(i_1, \cdots, i_p)$ of $\{1, \cdots, p\}$ it follows*

$$\int_{\mathbb{R}^p} f dm_p = \int \cdots \int f(x_1, \cdots, x_p)\, dx_{i_1} \cdots dx_{i_p} \tag{9.2}$$

In particular, this implies that if A_i is Lebesgue measurable for each $i = 1, \cdots, p$ then

$$m_p \left(\prod_{i=1}^{p} A_i \right) = \prod_{i=1}^{p} m(A_i).$$

Proof: Define a π system as

$$\mathcal{K} \equiv \left\{ \prod_{i=1}^{p} A_i : A_i \text{ is Lebesgue measurable} \right\}$$

Also let $R_n \equiv [-n, n]^p$, the p dimensional rectangle having sides $[-n, n]$. A set $F \subseteq \mathbb{R}^p$ will be said to satisfy property $\mathcal{P}$ if for every $n \in \mathbb{N}$ and any two permutations of $\{1, 2, \cdots, p\}$, $(i_1, \cdots, i_p)$ and $(j_1, \cdots, j_p)$ the two iterated integrals

$$\int \cdots \int \mathcal{X}_{R_n \cap F} dx_{i_1} \cdots dx_{i_p}, \int \cdots \int \mathcal{X}_{R_n \cap F} dx_{j_1} \cdots dx_{j_p}$$

make sense and are equal. Now define $\mathcal{G}$ to be those subsets of $\mathbb{R}^p$ which have property $\mathcal{P}$.

Thus $\mathcal{K} \subseteq \mathcal{G}$ because if $(i_1, \cdots, i_p)$ is any permutation of $\{1, 2, \cdots, p\}$ and

$$A = \prod_{i=1}^{p} A_i \in \mathcal{K}$$

then

$$\int \cdots \int \mathcal{X}_{R_n \cap A} dx_{i_1} \cdots dx_{i_p} = \prod_{i=1}^{p} m\left([-n, n] \cap A_i\right).$$

Now suppose $F \in \mathcal{G}$ and let $(i_1, \cdots, i_p)$ and $(j_1, \cdots, j_p)$ be two permutations. Then

$$R_n = \left(R_n \cap F^C\right) \cup \left(R_n \cap F\right)$$

and so

$$\int \cdots \int \mathcal{X}_{R_n \cap F^C} dx_{i_1} \cdots dx_{i_p} = \int \cdots \int \left(\mathcal{X}_{R_n} - \mathcal{X}_{R_n \cap F}\right) dx_{i_1} \cdots dx_{i_p}$$

Since $R_n \in \mathcal{G}$ the iterated integrals on the right and hence on the left make sense. Then continuing with the expression on the right and using that $F \in \mathcal{G}$, it equals

$$\begin{aligned}
& (2n)^p - \int \cdots \int \mathcal{X}_{R_n \cap F} dx_{i_1} \cdots dx_{i_p} \\
= \; & (2n)^p - \int \cdots \int \mathcal{X}_{R_n \cap F} dx_{j_1} \cdots dx_{j_p} \\
= \; & \int \cdots \int \left(\mathcal{X}_{R_n} - \mathcal{X}_{R_n \cap F}\right) dx_{j_1} \cdots dx_{j_p} \\
= \; & \int \cdots \int \mathcal{X}_{R_n \cap F^C} dx_{j_1} \cdots dx_{j_p}
\end{aligned}$$

which shows that if $F \in \mathcal{G}$ then so is F^C.

Next suppose $\{F_i\}_{i=1}^{\infty}$ is a sequence of disjoint sets in $\mathcal{G}$. Let $F = \cup_{i=1}^{\infty} F_i$. I need to show $F \in \mathcal{G}$. Since the sets are disjoint,

$$\int \cdots \int \mathcal{X}_{R_n \cap F} dx_{i_1} \cdots dx_{i_p} = \int \cdots \int \sum_{k=1}^{\infty} \mathcal{X}_{R_n \cap F_k} dx_{i_1} \cdots dx_{i_p}$$

$$= \int \cdots \int \lim_{N \to \infty} \sum_{k=1}^{N} \mathcal{X}_{R_n \cap F_k} dx_{i_1} \cdots dx_{i_p}$$

Do the iterated integrals make sense? Note that the iterated integral makes sense for $\sum_{k=1}^{N} \mathcal{X}_{R_n \cap F_k}$ as the integrand because it is just a finite sum of functions for which the iterated integral makes sense. Therefore,

$$x_{i_1} \to \sum_{k=1}^{\infty} \mathcal{X}_{R_n \cap F_k}(\mathbf{x})$$

is measurable and by the monotone convergence theorem,

$$\int \sum_{k=1}^{\infty} \mathcal{X}_{R_n \cap F_k}(\mathbf{x}) \, dx_{i_1} = \lim_{N \to \infty} \int \sum_{k=1}^{N} \mathcal{X}_{R_n \cap F_k} dx_{i_1}$$

Now each of the functions,

$$x_{i_2} \to \int \sum_{k=1}^{N} \mathcal{X}_{R_n \cap F_k} dx_{i_1}$$

is measurable and so the limit of these functions,

$$\int \sum_{k=1}^{\infty} \mathcal{X}_{R_n \cap F_k}(\mathbf{x})\, dx_{i_1}$$

is also measurable. Therefore, one can do another integral to this function. Continuing this way using the monotone convergence theorem, it follows the iterated integral makes sense. The same reasoning shows the iterated integral makes sense for any other permutation.

Now applying the monotone convergence theorem as needed,

$$\int \cdots \int \mathcal{X}_{R_n \cap F} dx_{i_1} \cdots dx_{i_p} = \int \cdots \int \sum_{k=1}^{\infty} \mathcal{X}_{R_n \cap F_k} dx_{i_1} \cdots dx_{i_p}$$

$$\begin{aligned}
&= \int \cdots \int \lim_{N \to \infty} \sum_{k=1}^{N} \mathcal{X}_{R_n \cap F_k} dx_{i_1} \cdots dx_{i_p} \\
&= \int \cdots \int \lim_{N \to \infty} \sum_{k=1}^{N} \int \mathcal{X}_{R_n \cap F_k} dx_{i_1} \cdots dx_{i_p} \\
&= \int \cdots \lim_{N \to \infty} \sum_{k=1}^{N} \int \int \mathcal{X}_{R_n \cap F_k} dx_{i_1} \cdots dx_{i_p} \cdots \\
&= \lim_{N \to \infty} \sum_{k=1}^{N} \int \cdots \int \mathcal{X}_{R_n \cap F_k} dx_{i_1} \cdots dx_{i_p} \\
&= \lim_{N \to \infty} \sum_{k=1}^{N} \int \cdots \int \mathcal{X}_{R_n \cap F_k} dx_{j_1} \cdots dx_{j_p}
\end{aligned}$$

the last step holding because each $F_k \in \mathcal{G}$. Then repeating the steps above in the opposite order, this equals

$$\int \cdots \int \sum_{k=1}^{\infty} \mathcal{X}_{R_n \cap F_k} dx_{j_1} \cdots dx_{j_p} = \int \cdots \int \mathcal{X}_{R_n \cap F} dx_{j_1} \cdots dx_{j_p}$$

Thus $F \in \mathcal{G}$. By Lemma 9.1.2 $\mathcal{G} \supseteq \sigma(\mathcal{K})$.

Let $\mathcal{F}^p = \sigma(\mathcal{K})$. Each set of the form $\prod_{k=1}^{p} U_k$ where U_k is an open set is in $\mathcal{K}$. Also every open set in $\mathbb{R}^p$ is a countable union of open sets of this form. This follows from Lemma 7.7.7 on Page 182. Therefore, every open set is in $\mathcal{F}^p$.

For $F \in \mathcal{F}^p$ define

$$m_p(F) \equiv \lim_{n \to \infty} \int \cdots \int \mathcal{X}_{R_n \cap F} dx_{j_1} \cdots dx_{j_p}$$

where $(j_1, \cdots, j_p)$ is a permutation of $\{1, \cdots, p\}$. It doesn't matter which one. It was shown above they all give the same result. I need to verify m_p is a measure. Let $\{F_k\}$ be a sequence of disjoint sets of $\mathcal{F}^p$.

$$m_p(\cup_{k=1}^{\infty} F_k) = \lim_{n \to \infty} \int \cdots \int \sum_{k=1}^{\infty} \mathcal{X}_{R_n \cap F_k} dx_{j_1} \cdots dx_{j_p}.$$

Using the monotone convergence theorem repeatedly as in the first part of the argument, this equals

$$\sum_{k=1}^{\infty} \lim_{n\to\infty} \int \cdots \int \mathcal{X}_{R_n \cap F_k} dx_{j_1} \cdots dx_{j_p} \equiv \sum_{k=1}^{\infty} m_p\left(F_k\right).$$

Thus m_p is a measure. Now letting A_k be a Lebesgue measurable set,

$$\begin{aligned} m_p\left(\prod_{k=1}^{p} A_k\right) &= \lim_{n\to\infty} \int \cdots \int \prod_{k=1}^{p} \mathcal{X}_{[-n,n]\cap A_k}\left(x_k\right) dx_{j_1} \cdots dx_{j_p} \\ &= \lim_{n\to\infty} \prod_{k=1}^{p} m\left([-n,n] \cap A_k\right) = \prod_{k=1}^{p} m\left(A_k\right). \end{aligned}$$

It only remains to prove 9.2.

It was shown above that for $F \in \mathcal{F}$ it follows

$$\int_{\mathbb{R}^p} \mathcal{X}_F dm_p = \lim_{n\to\infty} \int \cdots \int \mathcal{X}_{R_n \cap F} dx_{j_1} \cdots dx_{j_p}$$

Applying the monotone convergence theorem repeatedly on the right, this yields that the iterated integral makes sense and

$$\int_{\mathbb{R}^p} \mathcal{X}_F dm_p = \int \cdots \int \mathcal{X}_F dx_{j_1} \cdots dx_{j_p}$$

It follows 9.2 holds for every nonnegative simple function in place of f because these are just linear combinations of functions, $\mathcal{X}_F$. Now taking an increasing sequence of nonnegative simple functions, $\{s_k\}$ which converges to a measurable nonnegative function f

$$\begin{aligned} \int_{\mathbb{R}^p} f dm_p &= \lim_{k\to\infty} \int_{\mathbb{R}^p} s_k dm_p \\ &= \lim_{k\to\infty} \int \cdots \int s_k dx_{j_1} \cdots dx_{j_p} \\ &= \int \cdots \int f dx_{j_1} \cdots dx_{j_p} \end{aligned}$$

This proves the proposition. ■

9.2.3 Fubini's Theorem

Formula 9.2 is often called Fubini's theorem. So is the following theorem. In general, people tend to refer to theorems about the equality of iterated integrals as Fubini's theorem and in fact Fubini did produce such theorems but so did Tonelli and some of these theorems presented here and above should be called Tonelli's theorem.

Theorem 9.2.3 *Let m_p be defined in Proposition 9.2.2 on the σ algebra of sets $\mathcal{F}^p$ given there. Suppose $f \in L^1\left(\mathbb{R}^p\right)$. Then if $(i_1, \cdots, i_p)$ is any permutation of $\{1, \cdots, p\}$,*

$$\int_{\mathbb{R}^p} f dm_p = \int \cdots \int f\left(\mathbf{x}\right) dx_{i_1} \cdots dx_{i_p}.$$

In particular, iterated integrals for any permutation of $\{1, \cdots, p\}$ are all equal.

Proof: It suffices to prove this for f having real values because if this is shown the general case is obtained by taking real and imaginary parts. Since $f \in L^1(\mathbb{R}^p)$,

$$\int_{\mathbb{R}^p} |f|\, dm_p < \infty$$

and so both $\frac{1}{2}(|f| + f)$ and $\frac{1}{2}(|f| - f)$ are in $L^1(\mathbb{R}^p)$ and are each nonnegative. Hence from Proposition 9.2.2,

$$\begin{aligned}
\int_{\mathbb{R}^p} f dm_p &= \int_{\mathbb{R}^p} \left[\frac{1}{2}(|f| + f) - \frac{1}{2}(|f| - f)\right] dm_p \\
&= \int_{\mathbb{R}^p} \frac{1}{2}(|f| + f)\, dm_p - \int_{\mathbb{R}^p} \frac{1}{2}(|f| - f)\, dm_p \\
&= \int \cdots \int \frac{1}{2}(|f(\mathbf{x})| + f(\mathbf{x}))\, dx_{i_1} \cdots dx_{i_p} \\
&\quad - \int \cdots \int \frac{1}{2}(|f(\mathbf{x})| - f(\mathbf{x}))\, dx_{i_1} \cdots dx_{i_p} \\
&= \int \cdots \int \frac{1}{2}(|f(\mathbf{x})| + f(\mathbf{x})) - \frac{1}{2}(|f(\mathbf{x})| - f(\mathbf{x}))\, dx_{i_1} \cdots dx_{i_p} \\
&= \int \cdots \int f(\mathbf{x})\, dx_{i_1} \cdots dx_{i_p}
\end{aligned}$$

This proves the theorem. ■

The following corollary is a convenient way to verify the hypotheses of the above theorem.

Corollary 9.2.4 *Suppose f is measurable with respect to $\mathcal{F}^p$ and suppose for some permutation, $(i_1, \cdots, i_p)$*

$$\int \cdots \int |f(\mathbf{x})|\, dx_{i_1} \cdots dx_{i_p} < \infty$$

Then $f \in L^1(\mathbb{R}^p)$.

Proof: By Proposition 9.2.2,

$$\int_{\mathbb{R}^p} |f|\, dm_p = \int \cdots \int |f(\mathbf{x})|\, dx_{i_1} \cdots dx_{i_p} < \infty$$

and so f is in $L^1(\mathbb{R}^p)$ by Corollary 8.7.6. This proves the corollary. ■

The following theorem is a summary of the above specialized to Borel sets along with an assertion about regularity.

Theorem 9.2.5 *Let $\mathcal{B}(\mathbb{R}^p)$ be the Borel sets on $\mathbb{R}^p$. There exists a measure m_p defined on $\mathcal{B}(\mathbb{R}^p)$ such that if f is a nonnegative Borel measurable function,*

$$\int_{\mathbb{R}^p} f dm_p = \int \cdots \int f(\mathbf{x})\, dx_{i_1} \cdots dx_{i_p} \tag{9.3}$$

whenever $(i_1, \cdots, i_p)$ is a permutation of $\{1. \cdots, p\}$. If $f \in L^1(\mathbb{R}^p)$ and f is Borel measurable, then the above equation holds for f and all integrals make sense. If f is Borel measurable and for some $(i_1, \cdots, i_p)$ a permutation of $\{1. \cdots, p\}$

$$\int \cdots \int |f(\mathbf{x})|\, dx_{i_1} \cdots dx_{i_p} < \infty,$$

then $f \in L^1(\mathbb{R}^p)$. The measure m_p is both inner and outer regular on the Borel sets. That is, if $E \in \mathcal{B}(\mathbb{R}^p)$,

$$m_p(E) = \sup\{m_p(K) : K \subseteq E \text{ and } K \text{ is compact}\}$$

$$m_p(E) = \inf\{m_p(V) : V \supseteq E \text{ and } V \text{ is open}\}.$$

Also if A_k is a Borel set in $\mathbb{R}$ then

$$\prod_{k=1}^{p} A_k$$

is a Borel set in $\mathbb{R}^p$ and

$$m_p\left(\prod_{k=1}^{p} A_k\right) = \prod_{k=1}^{p} m(A_k).$$

Proof: Most of it was shown earlier since $\mathcal{B}(\mathbb{R}^p) \subseteq \mathcal{F}^p$. The two assertions about regularity follow from observing that m_p is finite on compact sets and then using Theorem 7.4.6. It remains to show the assertion about the product of Borel sets. If each A_k is open, there is nothing to show because the result is an open set. Suppose then that whenever $A_1, \cdots, A_m, m \leq p$ are open, the product, $\prod_{k=1}^{p} A_k$ is a Borel set. Let $\mathcal{K}$ be the open sets in $\mathbb{R}$ and let $\mathcal{G}$ be those Borel sets such that if $A_m \in \mathcal{G}$ it follows $\prod_{k=1}^{p} A_k$ is Borel. Then $\mathcal{K}$ is a π system and is contained in $\mathcal{G}$. Now suppose $F \in \mathcal{G}$. Then

$$\begin{aligned} &\left(\prod_{k=1}^{m-1} A_k \times F \times \prod_{k=m+1}^{p} A_k\right) \cup \left(\prod_{k=1}^{m-1} A_k \times F^C \times \prod_{k=m+1}^{p} A_k\right) \\ = \quad &\left(\prod_{k=1}^{m-1} A_k \times \mathbb{R} \times \prod_{k=m+1}^{p} A_k\right) \end{aligned}$$

and by assumption this is of the form

$$B \cup A = D.$$

where B, A are disjoint and B and D are Borel. Therefore, $A = D \setminus B$ which is a Borel set. Thus $\mathcal{G}$ is closed with respect to complements. If $\{F_i\}$ is a sequence of disjoint elements of $\mathcal{G}$

$$\left(\prod_{k=1}^{m-1} A_k \times \cup_i F_i \times \prod_{k=m+1}^{p} A_k\right) = \cup_{i=1}^{\infty}\left(\prod_{k=1}^{m-1} A_k \times F_i \times \prod_{k=m+1}^{p} A_k\right)$$

which is a countable union of Borel sets and is therefore, Borel. Hence $\mathcal{G}$ is also closed with respect to countable unions of disjoint sets. Thus by the Lemma on π systems $\mathcal{G} \supseteq \sigma(\mathcal{K}) = \mathcal{B}(\mathbb{R})$ and this shows that A_m can be any Borel set. Thus the assertion about the product is true if only $A_1, \cdots, A_{m-1}$ are open while the rest are Borel. Continuing this way shows the assertion remains true for each A_i being Borel. Now the final formula about the measure of a product follows from 9.3.

$$\begin{aligned} \int_{\mathbb{R}^p} \mathcal{X}_{\prod_{k=1}^{p} A_k} dm_p &= \int \cdots \int \mathcal{X}_{\prod_{k=1}^{p} A_k}(\mathbf{x})\, dx_1 \cdots dx_p \\ &= \int \cdots \int \prod_{k=1}^{p} \mathcal{X}_{A_k}(x_k)\, dx_1 \cdots dx_p = \prod_{k=1}^{p} m(A_k). \end{aligned}$$

This proves the theorem. ■

Of course iterated integrals can often be used to compute the Lebesgue integral. Sometimes the iterated integral taken in one order will allow you to compute the Lebesgue integral and it does not work well in the other order. Here is a simple example.

Example 9.2.6 *Find the iterated integral*

$$\int_0^1 \int_x^1 \frac{\sin(y)}{y} dy dx$$

Notice the limits. The iterated integral equals

$$\int_{\mathbb{R}^2} \mathcal{X}_A(x,y) \frac{\sin(y)}{y} dm_2$$

where

$$A = \{(x,y) : x \leq y \text{ where } x \in [0,1]\}$$

Fubini's theorem can be applied because the function $(x,y) \to \sin(y)/y$ is continuous except at $y = 0$ and can be redefined to be continuous there. The function is also bounded so

$$(x,y) \to \mathcal{X}_A(x,y) \frac{\sin(y)}{y}$$

clearly is in $L^1(\mathbb{R}^2)$. Therefore,

$$\begin{aligned} \int_{\mathbb{R}^2} \mathcal{X}_A(x,y) \frac{\sin(y)}{y} dm_2 &= \int\int \mathcal{X}_A(x,y) \frac{\sin(y)}{y} dx dy \\ &= \int_0^1 \int_0^y \frac{\sin(y)}{y} dx dy \\ &= \int_0^1 \sin(y)\, dy = 1 - \cos(1) \end{aligned}$$

9.3 Exercises

1. Find $\int_0^2 \int_0^{6-2z} \int_{\frac{1}{2}x}^{3-z} (3-z)\cos\left(y^2\right) dy\, dx\, dz$.
2. Find $\int_0^1 \int_0^{18-3z} \int_{\frac{1}{3}x}^{6-z} (6-z)\exp\left(y^2\right) dy\, dx\, dz$.
3. Find $\int_0^2 \int_0^{24-4z} \int_{\frac{1}{4}y}^{6-z} (6-z)\exp\left(x^2\right) dx\, dy\, dz$.
4. Find $\int_0^1 \int_0^{12-4z} \int_{\frac{1}{4}y}^{3-z} \frac{\sin x}{x} dx\, dy\, dz$.
5. Find $\int_0^{20} \int_0^1 \int_{\frac{1}{5}y}^{5-z} \frac{\sin x}{x} dx\, dz\, dy + \int_{20}^{25} \int_0^{5-\frac{1}{5}y} \int_{\frac{1}{5}y}^{5-z} \frac{\sin x}{x} dx\, dz\, dy$. **Hint:** You might try doing it in the order, $dy\, dx\, dz$
6. Explain why for each $t > 0, x \to e^{-tx}$ is a function in $L^1(\mathbb{R})$ and

$$\int_0^\infty e^{-tx} dx = \frac{1}{t}.$$

Thus

$$\int_0^R \frac{\sin(t)}{t} dt = \int_0^R \int_0^\infty \sin(t)\, e^{-tx} dx dt$$

Now explain why you can change the order of integration in the above iterated integral. Then compute what you get. Next pass to a limit as $R \to \infty$ and show

$$\int_0^\infty \frac{\sin(t)}{t} dt = \frac{1}{2}\pi$$

7. Explain why $\int_a^\infty f(t)\,dt \equiv \lim_{r\to\infty}\int_a^r f(t)\,dt$ whenever $f \in L^1(a,\infty)$; that is $f\mathcal{X}_{[a,\infty)} \in L^1(\mathbb{R})$.

8. $B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx, \Gamma(p) = \int_0^\infty e^{-t}t^{p-1}dt$ for $p,q>0$. The first of these is called the beta function, while the second is the gamma function. Show a.) $\Gamma(p+1) = p\Gamma(p)$; b.) $\Gamma(p)\Gamma(q) = B(p,q)\Gamma(p+q)$.Explain why the gamma function makes sense for any $p>0$.

9. Let $f(y) = g(y) = |y|^{-1/2}$if $y \in (-1,0)\cup(0,1)$ and $f(y) = g(y) = 0$ if $y \notin (-1,0)\cup(0,1)$. For which values of x does it make sense to write the integral $\int_{\mathbb{R}} f(x-y)\,g(y)\,dy$?

10. Let $\{a_n\}$ be an increasing sequence of numbers in $(0,1)$ which converges to 1. Let g_n be a nonnegative function which equals zero outside (a_n, a_{n+1}) such that $\int g_n dx = 1$. Now for $(x,y) \in [0,1)\times[0,1)$ define

$$f(x,y) \equiv \sum_{k=1}^{\infty} g_n(y)\left(g_n(x) - g_{n+1}(x)\right).$$

Explain why this is actually a finite sum for each such (x,y) so there are no convergence questions in the infinite sum. Explain why f is a continuous function on $[0,1)\times[0,1)$. You can extend f to equal zero off $[0,1)\times[0,1)$ if you like. Show the iterated integrals exist but are not equal. In fact, show

$$\int_0^1\int_0^1 f(x,y)\,dydx = 1 \neq 0 = \int_0^1\int_0^1 f(x,y)\,dxdy.$$

Does this example contradict the Fubini theorem? Explain why or why not.

9.4 Lebesgue Measure On $\mathbb{R}^p$

The σ algebra of Lebesgue measurable sets is larger than the above σ algebra of Borel sets or of the earlier σ algebra which came from an application of the π system lemma. It is convenient to use this larger σ algebra, especially when considering change of variables formulas, although it is certainly true that one can do most interesting theorems with the Borel sets only. However, it is in some ways easier to consider the more general situation and this will be done next.

Definition 9.4.1 *The completion of $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), m_p)$ is the Lebesgue measure space. Denote this by $(\mathbb{R}^p, \mathcal{F}_p, m_p)$.*

Thus for each $E \in \mathcal{F}_p$,

$$m_p(E) = \inf\{m_p(F) : F \supseteq E \text{ and } F \in \mathcal{B}(\mathbb{R}^p)\}$$

It follows that for each $E \in \mathcal{F}_p$ there exists $F \in \mathcal{B}(\mathbb{R}^p)$ such that $F \supseteq E$ and

$$m_p(E) = m_p(F).$$

Theorem 9.4.2 *m_p is regular on $\mathcal{F}_p$. In fact, if $E \in \mathcal{F}_p$, there exist sets in $\mathcal{B}(\mathbb{R}^p), F, G$ such that*

$$F \subseteq E \subseteq G,$$

F is a countable union of compact sets, G is a countable intersection of open sets and

$$m_p(G \setminus F) = 0.$$

If A_k is Lebesgue measurable then $\prod_{k=1}^p A_k \in \mathcal{F}_p$ and

$$m_p\left(\prod_{k=1}^p A_k\right) = \prod_{k=1}^p m\left(A_k\right).$$

In addition to this, m_p is translation invariant. This means that if $E \in \mathcal{F}_p$ and $\mathbf{x} \in \mathbb{R}^p$, then

$$m_p\left(\mathbf{x}+E\right) = m_p\left(E\right).$$

The expression $\mathbf{x}+E$ means $\{\mathbf{x}+\mathbf{e} : \mathbf{e} \in E\}$.

Proof: m_p is regular on $\mathcal{B}\left(\mathbb{R}^p\right)$ by Theorem 7.4.6 because it is finite on compact sets. Then from Theorem 7.5.7, it follows m_p is regular on $\mathcal{F}_p$. This is because the regularity of m_p on the Borel sets and the definition of the completion of a measure given there implies the uniqueness part of Theorem 7.5.6 can be used to conclude

$$\left(\mathbb{R}^p, \mathcal{F}_p, m_p\right) = \left(\mathbb{R}^p, \overline{\mathcal{B}\left(\mathbb{R}^p\right)}, \overline{m_p}\right)$$

Now for $E \in \mathcal{F}_p$, having finite measure, there exists $F \in \mathcal{B}\left(\mathbb{R}^p\right)$ such that $m_p\left(F\right) = m_p\left(E\right)$ and $F \supseteq E$. Thus $m_p\left(F \setminus E\right) = 0$. By regularity of m_p on the Borel sets, there exists G a countable intersection of open sets such that $G \supseteq F$ and $m_p\left(G\right) = m_p\left(F\right) = m_p\left(E\right)$. Thus $m_p\left(G \setminus E\right) = 0$. If E does not have finite measure, then letting

$$E_m \equiv \left(B\left(\mathbf{0}, m\right) \setminus B\left(\mathbf{0}, m-1\right)\right) \cap E,$$

it follows there exists G_m, an intersection of open sets such that $m_p\left(G_m \setminus E_m\right) = 0$. Hence if $G = \cup_m G_m$, it follows $m_p\left(G \setminus E\right) \leq \sum_m m_p\left(G_m \setminus E_m\right) = 0$. Thus G is a countable intersection of open sets.

To obtain F a countable union of compact sets contained in E such that $m_p\left(E \setminus F\right) = 0$, consider the closed sets $A_m = \overline{B\left(\mathbf{0},m\right)} \setminus B\left(\mathbf{0}, m-1\right)$ and let $E_m = A_m \cap E$. Then from what was just shown, there exists $G_m \supseteq \left(A_m \setminus E_m\right)$ such that

$$m_p\left(G_m \setminus \left(A_m \setminus E_m\right)\right) = 0.$$

and G_m is the countable intersection of open sets. The set on the inside equals $\left(G_m \cap A_m^C\right) \cup \left(G_m \cap E_m\right)$. Also

$$G_m^C \subseteq A_m^C \cup E_m \text{ so } G_m^C \cap A_m \subseteq E_m$$

and $G_m^C \cap A_m$ is the countable union of closed sets. Also

$$\begin{aligned} m_p\left(E_m \setminus \left(G_m^C \cap A_m\right)\right) &= m_p\left(\left(E_m \cap G_m\right) \cup \left(\overbrace{E_m \cap A_m^C}^{=\emptyset}\right)\right) \\ &\leq m_p\left(\left(G_m \cap A_m^C\right) \cup \left(G_m \cap E_m\right)\right) = 0. \end{aligned}$$

Denote this set $G_m^C \cap A_m$ by F_m. It is a countable union of closed sets and $m_p\left(E_m \setminus F_m\right) = 0$. Let $F = \cup_{m=1}^\infty F_m$. Then F is a countable union of compact sets and

$$m_p\left(E \setminus F\right) \leq \sum_{m=1}^\infty m_p\left(E_m \setminus F_m\right) = 0.$$

Consider the next assertion about the measure of a Cartesian product. By regularity of m there exists $B_k, C_k \in \mathcal{B}\left(\mathbb{R}^p\right)$ such that $B_k \supseteq A_k \supseteq C_k$ and $m\left(B_k\right) = m\left(A_k\right) =$

$m(C_k)$. In fact, you can have B_k equal a countable intersection of open sets and C_k a countable union of compact sets. Then

$$\begin{aligned}\prod_{k=1}^{p} m(A_k) &= \prod_{k=1}^{p} m(C_k) \leq m_p\left(\prod_{k=1}^{p} C_k\right)\\ &\leq m_p\left(\prod_{k=1}^{p} A_k\right) \leq m_p\left(\prod_{k=1}^{p} B_k\right)\\ &= \prod_{k=1}^{p} m(B_k) = \prod_{k=1}^{p} m(A_k).\end{aligned}$$

It remains to prove the claim about the measure being translation invariant.

Let $\mathcal{K}$ denote all sets of the form

$$\prod_{k=1}^{p} U_k$$

where each U_k is an open set in $\mathbb{R}$. Thus $\mathcal{K}$ is a π system.

$$\mathbf{x} + \prod_{k=1}^{p} U_k = \prod_{k=1}^{p} (x_k + U_k)$$

which is also a finite Cartesian product of finitely many open sets. Also,

$$\begin{aligned}m_p\left(\mathbf{x} + \prod_{k=1}^{p} U_k\right) &= m_p\left(\prod_{k=1}^{p} (x_k + U_k)\right)\\ &= \prod_{k=1}^{p} m(x_k + U_k)\\ &= \prod_{k=1}^{p} m(U_k) = m_p\left(\prod_{k=1}^{p} U_k\right)\end{aligned}$$

The step to the last line is obvious because an arbitrary open set in $\mathbb{R}$ is the disjoint union of open intervals and the lengths of these intervals are unchanged when they are slid to another location.

Now let $\mathcal{G}$ denote those Borel sets E with the property that for each $n \in \mathbb{N}$

$$m_p(\mathbf{x} + E \cap (-n,n)^p) = m_p(E \cap (-n,n)^p)$$

and the set, $\mathbf{x} + E \cap (-n,n)^p$ is a Borel set. Thus $\mathcal{K} \subseteq \mathcal{G}$. If $E \in \mathcal{G}$ then

$$\left(\mathbf{x} + E^C \cap (-n,n)^p\right) \cup (\mathbf{x} + E \cap (-n,n)^p) = \mathbf{x} + (-n,n)^p$$

which implies $\mathbf{x} + E^C \cap (-n,n)^p$ is a Borel set since it equals a difference of two Borel sets. Now consider the following.

$$\begin{aligned}& m_p\left(\mathbf{x} + E^C \cap (-n,n)^p\right) + m_p(E \cap (-n,n)^p)\\ = \;& m_p\left(\mathbf{x} + E^C \cap (-n,n)^p\right) + m_p(\mathbf{x} + E \cap (-n,n)^p)\\ = \;& m_p(\mathbf{x} + (-n,n)^p) = m_p((-n,n)^p)\\ & = m_p\left(E^C \cap (-n,n)^p\right) + m_p(E \cap (-n,n)^p)\end{aligned}$$

which shows

$$m_p\left(\mathbf{x}+E^C\cap(-n,n)^p\right)=m_p\left(E^C\cap(-n,n)^p\right)$$

showing that $E^C\in\mathcal{G}$.

If $\{E_k\}$ is a sequence of disjoint sets of $\mathcal{G}$,

$$m_p\left(\mathbf{x}+\cup_{k=1}^{\infty}E_k\cap(-n,n)^p\right)=m_p\left(\cup_{k=1}^{\infty}\mathbf{x}+E_k\cap(-n,n)^p\right)$$

Now the sets $\{\mathbf{x}+E_k\cap(-p,p)^n\}$ are also disjoint and so the above equals

$$\begin{aligned}\sum_k m_p\left(\mathbf{x}+E_k\cap(-n,n)^p\right) &= \sum_k m_p\left(E_k\cap(-n,n)^p\right)\\ &= m_p\left(\cup_{k=1}^{\infty}E_k\cap(-n,n)^p\right)\end{aligned}$$

Thus $\mathcal{G}$ is also closed with respect to countable disjoint unions. It follows from the lemma on π systems that $\mathcal{G}\supseteq\sigma(\mathcal{K})$. But from Lemma 7.7.7 on Page 182, every open set is a countable union of sets of $\mathcal{K}$ and so $\sigma(\mathcal{K})$ contains the open sets. Therefore, $\mathcal{B}(\mathbb{R}^p)\supseteq\mathcal{G}\supseteq\sigma(\mathcal{K})\supseteq\mathcal{B}(\mathcal{K})$ which shows $\mathcal{G}=\mathcal{B}(\mathbb{R}^p)$.

I have just shown that for every $E\in\mathcal{B}(\mathbb{R}^p)$, and any $n\in\mathbb{N}$,

$$m_p\left(\mathbf{x}+E\cap(-n,n)^p\right)=m_p\left(E\cap(-n,n)^p\right)$$

Taking the limit as $n\to\infty$ yields

$$m_p\left(\mathbf{x}+E\right)=m_p\left(E\right).$$

This proves translation invariance on Borel sets.

Now suppose $m_p(S)=0$ so that S is a set of measure zero. From outer regularity, there exists a Borel set, F such that $F\supseteq S$ and $m_p(F)=0$. Therefore from what was just shown,

$$m_p\left(\mathbf{x}+S\right)\le m_p\left(\mathbf{x}+F\right)=m_p\left(F\right)=m_p\left(S\right)=0$$

which shows that if $m_p(S)=0$ then so does $m_p(\mathbf{x}+S)$. Let F be any set of $\mathcal{F}_p$. By regularity, there exists $E\supseteq F$ where $E\in\mathcal{B}(\mathbb{R}^p)$ and $m_p(E\setminus F)=0$. Then

$$m_p\left(F\right)=m_p\left(E\right)=m_p\left(\mathbf{x}+E\right)=m_p\left(\mathbf{x}+(E\setminus F)\cup F\right)$$

$$=m_p\left(\mathbf{x}+E\setminus F\right)+m_p\left(\mathbf{x}+F\right)=m_p\left(\mathbf{x}+F\right).$$

■

9.5 Mollifiers

From Theorem 8.10.3, every function in $L^1(\mathbb{R}^p)$ can be approximated by one in $C_c(\mathbb{R}^p)$ but even more incredible things can be said. In fact, you can approximate an arbitrary function in $L^1(\mathbb{R}^p)$ with one which is infinitely differentiable having compact support. This is very important in partial differential equations. I am just giving a short introduction to this concept here. Consider the following example.

Example 9.5.1 *Let $U=B(\mathbf{z},2r)$*

$$\psi(\mathbf{x})=\begin{cases}\exp\left[\left(\left|\mathbf{x}-\mathbf{z}\right|^2-r^2\right)^{-1}\right] & \text{if } |\mathbf{x}-\mathbf{z}|<r,\\ 0 \text{ if } |\mathbf{x}-\mathbf{z}|\ge r.\end{cases}$$

Then a little work shows $\psi\in C_c^{\infty}(U)$. Also note that if $\mathbf{z}=\mathbf{0}$, then $\psi(\mathbf{x})=\psi(-\mathbf{x})$.

You show this by verifying the partial derivatives all exist and are continuous. The only place this is hard is when $|\mathbf{x}-\mathbf{z}| = r$. It is left as an exercise. You might consider a simpler example,

$$f(x) = \begin{cases} e^{-1/x^2} \text{ if } x \neq 0 \\ 0 \text{ if } x = 0 \end{cases}$$

and reduce the above to a consideration of something like this simpler case.

Lemma 9.5.2 *Let U be any open set. Then $C_c^\infty(U) \neq \emptyset$.*

Proof: Pick $\mathbf{z} \in U$ and let r be small enough that $B(\mathbf{z}, 2r) \subseteq U$. Then let $\psi \in C_c^\infty(B(\mathbf{z}, 2r)) \subseteq C_c^\infty(U)$ be the function of the above example. ■

Definition 9.5.3 *Let $U = \{\mathbf{x} \in \mathbb{R}^p : |\mathbf{x}| < 1\}$. A sequence $\{\psi_m\} \subseteq C_c^\infty(U)$ is called a mollifier if*

$$\psi_m(\mathbf{x}) \geq 0,\ \psi_m(\mathbf{x}) = 0,\ \text{if } |\mathbf{x}| \geq \frac{1}{m},$$

and $\int \psi_m(\mathbf{x}) = 1$. Sometimes it may be written as $\{\psi_\varepsilon\}$ where ψ_ε satisfies the above conditions except $\psi_\varepsilon(\mathbf{x}) = 0$ if $|\mathbf{x}| \geq \varepsilon$. In other words, ε takes the place of $1/m$ and in everything that follows $\varepsilon \to 0$ instead of $m \to \infty$.

$\int f(\mathbf{x}, \mathbf{y}) dm_p(\mathbf{y})$ will mean $\mathbf{x}$ is fixed and the function $\mathbf{y} \to f(\mathbf{x}, \mathbf{y})$ is being integrated. To make the notation more familiar, dx is written instead of $dm_p(x)$.

Example 9.5.4 *Let*

$$\psi \in C_c^\infty(B(0,1)) \ \ (B(0,1) = \{\mathbf{x} : |\mathbf{x}| < 1\})$$

with $\psi(\mathbf{x}) \geq 0$ and $\int \psi dm = 1$. Let $\psi_m(\mathbf{x}) = c_m \psi(m\mathbf{x})$ where c_m is chosen in such a way that $\int \psi_m dm = 1$.

Definition 9.5.5 *A function, f, is said to be in $L^1_{loc}(\mathbb{R}^p)$ if f is Lebesgue measurable and if $|f|\mathcal{X}_K \in L^1(\mathbb{R}^p)$ for every compact set, K. If $f \in L^1_{loc}(\mathbb{R}^p)$, and $g \in C_c(\mathbb{R}^p)$,*

$$f * g(\mathbf{x}) \equiv \int f(\mathbf{y}) g(\mathbf{x}-\mathbf{y}) dy.$$

This is called the convolution of f and g.

The following is an important property of the convolution.

Proposition 9.5.6 *Let f and g be as in the above definition. Then*

$$\int f(\mathbf{y}) g(\mathbf{x}-\mathbf{y}) dy = \int f(\mathbf{x}-\mathbf{y}) g(\mathbf{y}) dy$$

Proof: This follows right away from the change of variables formula. In the left, let $\mathbf{x}-\mathbf{y} \equiv \mathbf{u}$. Then the left side equals

$$\int f(\mathbf{x}-\mathbf{u}) g(\mathbf{u}) du$$

because the absolute value of the determinant of the derivative is 1. Now replace $\mathbf{u}$ with $\mathbf{y}$ and This proves the proposition. ■

The following lemma will be useful in what follows. It says among other things that one of these very unregular functions in $L^1_{loc}(\mathbb{R}^p)$ is smoothed out by convolving with a mollifier.

Lemma 9.5.7 *Let $f \in L^1_{loc}(\mathbb{R}^p)$, and $g \in C_c^\infty(\mathbb{R}^p)$. Then $f * g$ is an infinitely differentiable function. Also, if $\{\psi_m\}$ is a mollifier and U is an open set and $f \in C^0(U) \cap L^1_{loc}(\mathbb{R}^p)$, then at every $\mathbf{x} \in U$,*

$$\lim_{m\to\infty} f * \psi_m(\mathbf{x}) = f(\mathbf{x}).$$

If $f \in C^1(U) \cap L^1_{loc}(\mathbb{R}^p)$ and $\mathbf{x} \in U$,

$$(f * \psi_m)_{x_i}(\mathbf{x}) = f_{x_i} * \psi_m(\mathbf{x}).$$

Proof: Consider the difference quotient for calculating a partial derivative of $f * g$.

$$\frac{f * g(\mathbf{x} + t\mathbf{e}_j) - f * g(\mathbf{x})}{t} = \int f(\mathbf{y}) \frac{g(\mathbf{x} + t\mathbf{e}_j - \mathbf{y}) - g(\mathbf{x} - \mathbf{y})}{t} dy.$$

Using the fact that $g \in C_c^\infty(\mathbb{R}^p)$, the quotient,

$$\frac{g(\mathbf{x} + t\mathbf{e}_j - \mathbf{y}) - g(\mathbf{x} - \mathbf{y})}{t},$$

is uniformly bounded. To see this easily, use Theorem 6.5.2 on Page 131 to get the existence of a constant, M depending on

$$\max\{||Dg(\mathbf{x})|| : \mathbf{x} \in \mathbb{R}^p\}$$

such that

$$|g(\mathbf{x} + t\mathbf{e}_j - \mathbf{y}) - g(\mathbf{x} - \mathbf{y})| \leq M|t|$$

for any choice of $\mathbf{x}$ and $\mathbf{y}$. Therefore, there exists a dominating function for the integrand of the above integral which is of the form $C|f(\mathbf{y})|\mathcal{X}_K$ where K is a compact set depending on the support of g. It follows from the dominated convergence theorem the limit of the difference quotient above passes inside the integral as $t \to 0$ and so

$$\frac{\partial}{\partial x_j}(f * g)(\mathbf{x}) = \int f(\mathbf{y}) \frac{\partial}{\partial x_j} g(\mathbf{x} - \mathbf{y})\, dy.$$

Now letting $\frac{\partial}{\partial x_j} g$ play the role of g in the above argument, a repeat of the above reasoning shows partial derivatives of all orders exist. A similar use of the dominated convergence theorem shows all these partial derivatives are also continuous.

It remains to verify the claim about the mollifier. Let $\mathbf{x} \in U$ and let m be large enough that $B\left(\mathbf{x}, \frac{1}{m}\right) \subseteq U$. Then

$$|f * g(\mathbf{x}) - f(\mathbf{x})| \leq \int_{B\left(\mathbf{0}, \frac{1}{m}\right)} |f(\mathbf{x} - \mathbf{y}) - f(\mathbf{x})| \psi_m(\mathbf{y})\, dy$$

By continuity of f at $\mathbf{x}$, for all m sufficiently large, the above is dominated by

$$\varepsilon \int_{B\left(\mathbf{0}, \frac{1}{m}\right)} \psi_m(\mathbf{y})\, dy = \varepsilon$$

and this proves the claim.

Now consider the formula in the case where $f \in C^1(U)$. Using Proposition 9.5.6,

$$\frac{f * \psi_m(\mathbf{x} + h\mathbf{e}_i) - f * \psi_m(\mathbf{x})}{h} =$$

$$\frac{1}{h}\left(\int_{B\left(\mathbf{0}, \frac{1}{m}\right)} f(\mathbf{x} + h\mathbf{e}_i - \mathbf{y}) \psi_m(\mathbf{y})\, dy - \int_{B\left(\mathbf{0}, \frac{1}{m}\right)} f(\mathbf{x} - \mathbf{y}) \psi_m(\mathbf{y})\, dy\right)$$

$$= \int_{B\left(\mathbf{0},\frac{1}{m}\right)} \frac{\left(f\left(\mathbf{x}+h\mathbf{e}_i-\mathbf{y}\right)-f\left(\mathbf{x}-\mathbf{y}\right)\right)}{h}\psi_m\left(\mathbf{y}\right)dy$$

Now letting m be small enough and using the continuity of the partial derivatives, it follows the difference quotients are uniformly bounded for all h sufficiently small and so one can apply the dominated convergence theorem and pass to the limit obtaining

$$\int_{\mathbb{R}^p} f_{x_i}\left(\mathbf{x}-\mathbf{y}\right)\psi_m\left(\mathbf{y}\right)dy \equiv f_{x_i} * \psi_m\left(\mathbf{x}\right)$$

This proves the lemma. ■

Theorem 9.5.8 *Let K be a compact subset of an open set, U. Then there exists a function, $h \in C_c^\infty(U)$, such that $h(\mathbf{x}) = 1$ for all $\mathbf{x} \in K$ and $h(\mathbf{x}) \in [0,1]$ for all $\mathbf{x}$. Also there exists an open set W such that*

$$K \subseteq W \subseteq \overline{W} \subseteq U$$

such that $\overline{W}$ is compact.

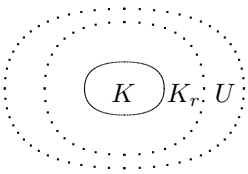
Proof: Let $r > 0$ be small enough that $K+B(\mathbf{0},3r) \subseteq U$. The symbol, $K+B(\mathbf{0},3r)$ means

$$\left\{\mathbf{k}+\mathbf{x} : \mathbf{k} \in K \text{ and } \mathbf{x} \in B\left(\mathbf{0},3r\right)\right\}.$$

Thus this is simply a way to write

$$\cup\left\{B\left(\mathbf{k},3r\right) : \mathbf{k} \in K\right\}.$$

Think of it as fattening up the set, K. Let $K_r = K + B(0,r)$. A picture of what is happening follows.



Consider $\mathcal{X}_{K_r} * \psi_m$where $\psi_m$is a mollifier. Let m be so large that $\frac{1}{m} < r$. Then from the definition of what is meant by a convolution, and using that ψ_m has support in $B\left(\mathbf{0},\frac{1}{m}\right)$, $\mathcal{X}_{K_r} * \psi_m = 1$ on K and its support is in $K + B\left(\mathbf{0},3r\right)$, a bounded set. Now using Lemma 9.5.7, $\mathcal{X}_{K_r} * \psi_m$ is also infinitely differentiable. Therefore, let $h = \mathcal{X}_{K_r} * \psi_m$.

As to the existence of the open set W, let it equal the closed and bounded set $h^{-1}\left(\left[\frac{1}{2},1\right]\right)$. This proves the theorem. ■

The following is the remarkable theorem mentioned above. First, here is some notation.

Definition 9.5.9 *Let $\mathbf{g}$ be a function defined on a vector space. Then $\mathbf{g}_{\mathbf{y}}\left(\mathbf{x}\right) \equiv \mathbf{g}\left(\mathbf{x}-\mathbf{y}\right)$.*

Theorem 9.5.10 *$C_c^\infty(\mathbb{R}^p)$ is dense in $L^1(\mathbb{R}^p)$. Here the measure is Lebesgue measure.*

Proof: Let $f \in L^1(\mathbb{R}^p)$ and let $\varepsilon > 0$ be given. Choose $g \in C_c(\mathbb{R}^p)$ such that

$$\int |g-f|\, dm_p < \varepsilon/2$$

This can be done by using Theorem 8.10.3. Now let

$$g_m(\mathbf{x}) = g * \psi_m(\mathbf{x}) \equiv \int g(\mathbf{x}-\mathbf{y})\,\psi_m(\mathbf{y})\,dm_p(y) = \int g(\mathbf{y})\,\psi_m(\mathbf{x}-\mathbf{y})\,dm_p(y)$$

where $\{\psi_m\}$ is a mollifier. It follows from Lemma 9.5.7 $g_m \in C_c^\infty(\mathbb{R}^p)$. It vanishes if $\mathbf{x} \notin \mathrm{spt}(g) + B(0, \frac{1}{m})$.

$$\begin{aligned}
\int |g - g_m|\,dm_p &= \int |g(\mathbf{x}) - \int g(\mathbf{x}-\mathbf{y})\psi_m(\mathbf{y})dm_p(\mathbf{y})|dm_p(\mathbf{x}) \\
&\leq \int (\int |g(\mathbf{x}) - g(\mathbf{x}-\mathbf{y})|\psi_m(\mathbf{y})dm_p(\mathbf{y}))dm_p(\mathbf{x}) \\
&\leq \int\int |g(\mathbf{x}) - g(\mathbf{x}-\mathbf{y})|dm_p(\mathbf{x})\psi_m(\mathbf{y})dm_p(\mathbf{y}) \\
&= \int_{B(0,\frac{1}{m})} \int |g - g_{\mathbf{y}}|dm_p(\mathbf{x})\psi_m(\mathbf{y})dm_p(\mathbf{y}) < \frac{\varepsilon}{2}
\end{aligned}$$

whenever m is large enough. This follows because since g has compact support, it is uniformly continuous on $\mathbb{R}^p$ and so if $\eta > 0$ is given, then whenever $|\mathbf{y}|$ is sufficiently small,

$$|g(\mathbf{x}) - g(\mathbf{x}-\mathbf{y})| < \eta$$

for all $\mathbf{x}$. Thus, since $\mathbf{g}$ has compact support, if $\mathbf{y}$ is small enough, it follows

$$\int |g - g_{\mathbf{y}}|dm_p(\mathbf{x}) < \varepsilon/2.$$

There is no measurability problem in the use of Fubini's theorem because the function

$$(\mathbf{x}, \mathbf{y}) \to |g(\mathbf{x}) - g(\mathbf{x}-\mathbf{y})|\psi_m(\mathbf{y})$$

is continuous. Thus when m is large enough,

$$\int |f - g_m|\,dm_p \leq \int |f-g|\,dm_p + \int |g - g_m|\,dm_p < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

This proves the theorem. ■

Another important application of Theorem 9.5.8 has to do with a partition of unity.

Definition 9.5.11 *A collection of sets $\mathcal{H}$ is called locally finite if for every $\mathbf{x}$, there exists $r > 0$ such that $B(\mathbf{x}, r)$ has nonempty intersection with only finitely many sets of $\mathcal{H}$. Of course every finite collection of sets is locally finite. This is the case of most interest in this book but the more general notion is interesting.*

The thing about locally finite collection of sets is that the closure of their union equals the union of their closures. This is clearly true of a finite collection.

Lemma 9.5.12 *Let $\mathcal{H}$ be a locally finite collection of sets of a normed vector space V. Then*

$$\overline{\cup\mathcal{H}} = \cup\left\{\overline{H} : H \in \mathcal{H}\right\}.$$

Proof: It is obvious $\supseteq$ holds in the above claim. It remains to go the other way. Suppose then that $\mathbf{p}$ is a limit point of $\cup\mathcal{H}$ and $\mathbf{p} \notin \cup\mathcal{H}$. There exists $r > 0$ such that $B(\mathbf{p}, r)$ has nonempty intersection with only finitely many sets of $\mathcal{H}$ say these are $H_1, \cdots, H_m$. Then I claim $\mathbf{p}$ must be a limit point of one of these. If this is not so, there would exist r' such that $0 < r' < r$ with $B(\mathbf{p}, r')$ having empty intersection with each

of these H_i. But then $\mathbf{p}$ would fail to be a limit point of $\cup\mathcal{H}$. Therefore, $\mathbf{p}$ is contained in the right side. It is clear $\cup\mathcal{H}$ is contained in the right side and so This proves the lemma. ■

A good example to consider is the rational numbers each being a set in $\mathbb{R}$. This is **not** a locally finite collection of sets and you note that $\overline{\mathbb{Q}} = \mathbb{R} \neq \cup\{\overline{x} : x \in \mathbb{Q}\}$. By contrast, $\mathbb{Z}$ is a locally finite collection of sets, the sets consisting of individual integers. The closure of $\mathbb{Z}$ is equal to $\mathbb{Z}$ because $\mathbb{Z}$ has no limit points so it contains them all.

Notation 9.5.13 *I will write $\phi \prec V$ to symbolize $\phi \in C_c(V)$, ϕ has values in $[0,1]$, and ϕ has compact support in V. I will write $K \prec \phi \prec V$ for K compact and V open to symbolize ϕ is 1 on K and ϕ has values in $[0,1]$ with compact support contained in V.*

A version of the following lemma is valid for locally finite coverings, but we are only using it when the covering is finite.

Lemma 9.5.14 *Let K be a closed set in $\mathbb{R}^p$ and let $\{V_i\}_{i=1}^n$ be a finite list of bounded open sets whose union contains K. Then there exist functions, $\psi_i \in C_c^\infty(V_i)$ such that for all $\mathbf{x} \in K$,*

$$1 = \sum_{i=1}^n \psi_i(\mathbf{x})$$

and the function $f(\mathbf{x})$ given by

$$f(\mathbf{x}) = \sum_{i=1}^n \psi_i(\mathbf{x})$$

is in $C^\infty(\mathbb{R}^p)$.

Proof: Let $K_1 = K \setminus \cup_{i=2}^n V_i$. Thus K_1 is compact because $K_1 \subseteq V_1$. Let W_1 be an open set having compact closure which satisfies

$$K_1 \subseteq W_1 \subseteq \overline{W}_1 \subseteq V_1$$

Thus $W_1, V_2, \cdots, V_n$ covers K and $\overline{W}_1 \subseteq V_1$. Suppose $W_1, \cdots, W_r$ have been defined such that $\overline{W_i} \subseteq V_i$ for each i, and $W_1, \cdots, W_r, V_{r+1}, \cdots, V_n$ covers K. Then let

$$K_{r+1} \equiv K \setminus ((\cup_{i=r+2}^n V_i) \cup (\cup_{j=1}^r W_j)).$$

It follows K_{r+1} is compact because $K_{r+1} \subseteq V_{r+1}$. Let W_{r+1} satisfy

$$K_{r+1} \subseteq W_{r+1} \subseteq \overline{W}_{r+1} \subseteq V_{r+1},\ \overline{W}_{r+1} \text{ is compact}$$

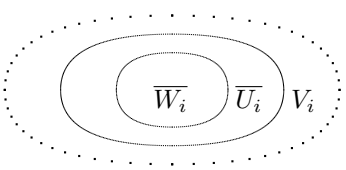
Continuing this way defines a sequence of open sets $\{W_i\}_{i=1}^n$ having compact closures with the property

$$\overline{W_i} \subseteq V_i,\ K \subseteq \cup_{i=1}^n W_i.$$

Note $\{W_i\}_{i=1}^n$ is locally finite because the original list, $\{V_i\}_{i=1}^n$ was locally finite. Now let U_i be open sets which satisfy

$$\overline{W}_i \subseteq U_i \subseteq \overline{U}_i \subseteq V_i,\ \overline{U}_i \text{ is compact.}$$

Similarly, $\{U_i\}_{i=1}^n$ is locally finite.



Now $\overline{\cup_{i=1}^n W_i} = \cup_{i=1}^n \overline{W_i}$ and so it is possible to define ϕ_i and γ, infinitely differentiable functions having compact support such that

$$\overline{U}_i \prec \phi_i \prec V_i,\ \cup_{i=1}^n \overline{W}_i \prec \gamma \prec \cup_{i=1}^n U_i.$$

Now define

$$\psi_i(\mathbf{x}) = \begin{cases} \gamma(\mathbf{x})\phi_i(\mathbf{x}) / \sum_{j=1}^n \phi_j(\mathbf{x}) \text{ if } \sum_{j=1}^n \phi_j(\mathbf{x}) \neq 0, \\ 0 \text{ if } \sum_{j=1}^n \phi_j(\mathbf{x}) = 0. \end{cases}$$

If $\mathbf{x}$ is such that $\sum_{j=1}^n \phi_j(\mathbf{x}) = 0$, then $\mathbf{x} \notin \cup_{i=1}^n \overline{U_i}$ because ϕ_i equals one on $\overline{U_i}$. Consequently $\gamma(\mathbf{y}) = 0$ for all $\mathbf{y}$ near $\mathbf{x}$ thanks to the fact that $\cup_{i=1}^n \overline{U_i}$ is closed and so $\psi_i(\mathbf{y}) = 0$ for all $\mathbf{y}$ near $\mathbf{x}$. Hence ψ_i is infinitely differentiable at such $\mathbf{x}$. If $\sum_{j=1}^n \phi_j(\mathbf{x}) \neq 0$, this situation persists near $\mathbf{x}$ because each ϕ_j is continuous and so ψ_i is infinitely differentiable at such points also. Therefore ψ_i is infinitely differentiable. If $\mathbf{x} \in K$, then $\gamma(\mathbf{x}) = 1$ and so $\sum_{j=1}^n \psi_j(\mathbf{x}) = 1$. Clearly $0 \leq \psi_i(\mathbf{x}) \leq 1$ and $\mathrm{spt}(\psi_j) \subseteq V_j$. This proves the theorem. ■

The functions, $\{\psi_i\}$ are called a C^∞ partition of unity.

Since K is compact, one often uses the above in the following form which follows from the same method of proof.

Corollary 9.5.15 *If H is a compact subset of V_i for some V_i there exists a partition of unity such that $\psi_i(x) = 1$ for all $x \in H$ in addition to the conclusion of Lemma 9.5.14.*

Proof: Keep V_i the same but replace V_j with $\widetilde{V_j} \equiv V_j \setminus H$. Now in the proof above, applied to this modified collection of open sets, if $j \neq i, \phi_j(x) = 0$ whenever $x \in H$. Therefore, $\psi_i(x) = 1$ on H. ■

9.6 The Vitali Covering Theorem

The Vitali covering theorem is a profound result about coverings of a set in $\mathbb{R}^p$ with open balls. The balls can be defined in terms of any norm for $\mathbb{R}^p$. For example, the norm could be

$$||\mathbf{x}|| \equiv \max\{|x_k| : k = 1, \cdots, p\}$$

or the usual norm

$$|\mathbf{x}| = \sqrt{\sum_k |x_k|^2}$$

or any other. The proof given here is from Basic Analysis [27]. It first considers the case of open balls and then generalizes to balls which may be neither open nor closed.

Lemma 9.6.1 *Let $\mathcal{F}$ be a countable collection of balls satisfying*

$$\infty > M \equiv \sup\{r : B(\mathbf{p}, r) \in \mathcal{F}\} > 0$$

and let $k \in (0, \infty)$. Then there exists $\mathcal{G} \subseteq \mathcal{F}$ such that

$$\text{If } B(\mathbf{p}, r) \in \mathcal{G} \text{ then } r > k, \tag{9.4}$$

$$\text{If } B_1, B_2 \in \mathcal{G} \text{ then } B_1 \cap B_2 = \emptyset, \tag{9.5}$$

$$\mathcal{G} \text{ is maximal with respect to 9.4 and 9.5.} \tag{9.6}$$

By this is meant that if $\mathcal{H}$ is a collection of balls satisfying 9.4 and 9.5, then $\mathcal{H}$ cannot properly contain $\mathcal{G}$.

Proof: If no ball of $\mathcal{F}$ has radius larger than k, let $\mathcal{G} = \emptyset$. Assume therefore, that some balls have radius larger than k. Let $\mathcal{F} \equiv \{B_i\}_{i=1}^{\infty}$. Now let B_{n_1} be the first ball in the list which has radius greater than k. If every ball having radius larger than k intersects this one, then stop. The maximal set is $\{B_{n_1}\}$. Otherwise, let B_{n_2} be the next ball having radius larger than k which is disjoint from B_{n_1}. Continue this way obtaining $\{B_{n_i}\}_{i=1}^{\infty}$, a finite or infinite sequence of disjoint balls having radius larger than k. Then let $\mathcal{G} \equiv \{B_{n_i}\}$. To see $\mathcal{G}$ is maximal with respect to 9.4 and 9.5, suppose $B \in \mathcal{F}$, B has radius larger than k, and $\mathcal{G} \cup \{B\}$ satisfies 9.4 and 9.5. Then at some point in the process, B would have been chosen because it would be the ball of radius larger than k which has the smallest index. Therefore, $B \in \mathcal{G}$ and this shows $\mathcal{G}$ is maximal with respect to 9.4 and 9.5. ■

For an open ball, $B = B(\mathbf{x}, r)$, denote by $\widetilde{B}$ the open ball, $B(\mathbf{x}, 4r)$.

Lemma 9.6.2 *Let $\mathcal{F}$ be a collection of open balls, and let*

$$A \equiv \cup \{B : B \in \mathcal{F}\}.$$

Suppose

$$\infty > M \equiv \sup \{r : B(\mathbf{p}, r) \in \mathcal{F}\} > 0.$$

Then there exists $\mathcal{G} \subseteq \mathcal{F}$ such that $\mathcal{G}$ consists of disjoint balls and

$$A \subseteq \cup\{\widetilde{B} : B \in \mathcal{G}\}.$$

Proof: Without loss of generality assume $\mathcal{F}$ is countable. This is because there is a countable subset of $\mathcal{F}$, $\mathcal{F}'$ such that $\cup\mathcal{F}' = A$. To see this, consider the set of balls having rational radii and centers having all components rational. This is a countable set of balls and you should verify that every open set is the union of balls of this form. Therefore, you can consider the subset of this set of balls consisting of those which are contained in some open set of $\mathcal{F}$, G so $\cup G = A$ and use the axiom of choice to define a subset of $\mathcal{F}$ consisting of a single set from $\mathcal{F}$ containing each set of G. Then this is $\mathcal{F}'$. The union of these sets equals A. Then consider $\mathcal{F}'$ instead of $\mathcal{F}$. Therefore, assume at the outset $\mathcal{F}$ is countable.

By Lemma 9.6.1, there exists $\mathcal{G}_1 \subseteq \mathcal{F}$ which satisfies 9.4, 9.5, and 9.6 with $k = \frac{2M}{3}$.

Suppose $\mathcal{G}_1, \cdots, \mathcal{G}_{m-1}$ have been chosen for $m \geq 2$. Let

$$\mathcal{F}_m = \{B \in \mathcal{F} : B \subseteq \mathbb{R}^p \setminus \overbrace{\cup\{\mathcal{G}_1 \cup \cdots \cup \mathcal{G}_{m-1}\}}^{\text{union of the balls in these } \mathcal{G}_j}\}$$

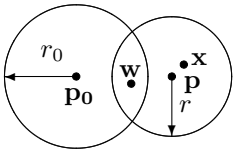
and using Lemma 9.6.1, let $\mathcal{G}_m$ be a maximal collection of disjoint balls from $\mathcal{F}_m$ with the property that each ball has radius larger than $\left(\frac{2}{3}\right)^m M$. Let $\mathcal{G} \equiv \cup_{k=1}^{\infty} \mathcal{G}_k$. Let $\mathbf{x} \in B(\mathbf{p}, r) \in \mathcal{F}$. Choose m such that

$$\left(\frac{2}{3}\right)^m M < r \leq \left(\frac{2}{3}\right)^{m-1} M$$

Then $B(\mathbf{p}, r)$ must have nonempty intersection with some ball from $\mathcal{G}_1 \cup \cdots \cup \mathcal{G}_m$ because if it didn't, then $\mathcal{G}_m$ would fail to be maximal. Denote by $B(\mathbf{p}_0, r_0)$ a ball in $\mathcal{G}_1 \cup \cdots \cup \mathcal{G}_m$ which has nonempty intersection with $B(\mathbf{p}, r)$. Thus

$$r_0 > \left(\frac{2}{3}\right)^m M.$$

Consider the picture, in which $\mathbf{w} \in B(\mathbf{p}_0, r_0) \cap B(\mathbf{p}, r)$.



Then

$$\begin{aligned} |\mathbf{x}-\mathbf{p}_0| &\leq |\mathbf{x}-\mathbf{p}|+|\mathbf{p}-\mathbf{w}|+\overbrace{|\mathbf{w}-\mathbf{p}_0|}^{<r_0} \\ &< r+r+r_0 \leq 2\overbrace{\left(\frac{2}{3}\right)^{m-1}}^{<\frac{3}{2}r_0} M+r_0 \\ &< 2\left(\frac{3}{2}\right)r_0+r_0 = 4r_0. \end{aligned}$$

This proves the lemma since it shows $B(\mathbf{p},r) \subseteq B(\mathbf{p}_0,4r_0)$. ■

With this Lemma consider a version of the Vitali covering theorem in which the balls do not have to be open. In this theorem, B will denote an open ball, $B(\mathbf{x},r)$ along with either part or all of the points where $||\mathbf{x}|| = r$ and $||\cdot||$ is any norm for $\mathbb{R}^p$.

Definition 9.6.3 *Let B be a ball centered at $\mathbf{x}$ having radius r. Denote by $\widehat{B}$ the open ball, $B(\mathbf{x},5r)$.*

Theorem 9.6.4 *(Vitali) Let $\mathcal{F}$ be a collection of balls, and let*

$$A \equiv \cup\{B : B \in \mathcal{F}\}.$$

Suppose

$$\infty > M \equiv \sup\{r : B(\mathbf{p},r) \in \mathcal{F}\} > 0.$$

Then there exists $\mathcal{G} \subseteq \mathcal{F}$ such that $\mathcal{G}$ consists of disjoint balls and

$$A \subseteq \cup\{\widehat{B} : B \in \mathcal{G}\}.$$

Proof: For B one of these balls, say $\overline{B(\mathbf{x},r)} \supseteq B \supseteq B(\mathbf{x},r)$, denote by B_1, the open ball $B\left(\mathbf{x},\frac{5r}{4}\right)$. Let $\mathcal{F}_1 \equiv \{B_1 : B \in \mathcal{F}\}$ and let A_1 denote the union of the balls in $\mathcal{F}_1$. Apply Lemma 9.6.2 to $\mathcal{F}_1$ to obtain

$$A_1 \subseteq \cup\{\widetilde{B_1} : B_1 \in \mathcal{G}_1\}$$

where $\mathcal{G}_1$ consists of disjoint balls from $\mathcal{F}_1$. Now let $\mathcal{G} \equiv \{B \in \mathcal{F} : B_1 \in \mathcal{G}_1\}$. Thus $\mathcal{G}$ consists of disjoint balls from $\mathcal{F}$ because they are contained in the disjoint open balls, $\mathcal{G}_1$. Then

$$A \subseteq A_1 \subseteq \cup\{\widetilde{B_1} : B_1 \in \mathcal{G}_1\} = \cup\{\widehat{B} : B \in \mathcal{G}\}$$

because for $B_1 = B\left(\mathbf{x},\frac{5r}{4}\right)$, it follows $\widetilde{B_1} = B(\mathbf{x},5r) = \widehat{B}$. This proves the theorem. ■

9.7 Vitali Coverings

There is another version of the Vitali covering theorem which is also of great importance. In this one, disjoint balls from the original set of balls almost cover the set, leaving out only a set of measure zero. It is like packing a truck with stuff. You keep trying to fill in the holes with smaller and smaller things so as to not waste space. It is remarkable that you can avoid wasting any space at all when you are dealing with balls of any sort provided you can use arbitrarily small balls.

Definition 9.7.1 *Let $\mathcal{F}$ be a collection of balls that cover a set, E, which have the property that if $\mathbf{x} \in E$ and $\varepsilon > 0$, then there exists $B \in \mathcal{F}$, diameter of $B < \varepsilon$ and $\mathbf{x} \in B$. Such a collection covers E in the sense of Vitali.*

In the following covering theorem, $\overline{m_p}$ denotes the outer measure determined by p dimensional Lebesgue measure. Thus, letting $\mathcal{F}$ denote the Lebesgue measurable sets,

$$\overline{m_p}(S) \equiv \inf\left\{\sum_{k=1}^{\infty} m_p(E_k) : S \subseteq \cup_k E_k, E_k \in \mathcal{F}\right\}$$

Recall that from this definition, if $S \subseteq \mathbb{R}^p$ there exists $E_1 \supseteq S$ such that $m_p(E_1) = \overline{m_p}(S)$. To see this, note that it suffices to assume in the above definition of $\overline{m_p}$ that the E_k are also disjoint. If not, replace with the sequence given by

$$F_1 = E_1, F_2 \equiv E_2 \setminus F_1, \cdots, F_m \equiv E_m \setminus F_{m-1},$$

etc. Then for each $l > \overline{m_p}(S)$, there exists $\{E_k\}$ such that

$$l > \sum_k m_p(E_k) \geq \sum_k m_p(F_k) = m_p(\cup_k E_k) \geq \overline{m_p}(S).$$

If $\overline{m_p}(S) = \infty$, let $E_1 = \mathbb{R}^p$. Otherwise, there exists $G_k \in \mathcal{F}$ such that

$$\overline{m_p}(S) \leq m_p(G_k) \leq \overline{m_p}(S) + 1/k.$$

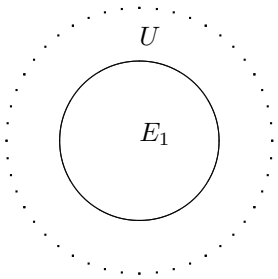
then let $E_1 = \cap_k G_k$.

Note this implies that if $\overline{m_p}(S) = 0$ then S must be in $\mathcal{F}$ because of completeness of Lebesgue measure.

Theorem 9.7.2 *Let $E \subseteq \mathbb{R}^p$ and suppose $0 < \overline{m_p}(E) < \infty$ where $\overline{m_p}$ is the outer measure determined by m_p, p dimensional Lebesgue measure, and let $\mathcal{F}$ be a collection of closed balls of bounded radii such that $\mathcal{F}$ covers E in the sense of Vitali. Then there exists a countable collection of disjoint balls from $\mathcal{F}$, $\{B_j\}_{j=1}^{\infty}$, such that $\overline{m_p}(E \setminus \cup_{j=1}^{\infty} B_j) = 0$.*

Proof: From the definition of outer measure there exists a Lebesgue measurable set, $E_1 \supseteq E$ such that $m_p(E_1) = \overline{m_p}(E)$. Now by outer regularity of Lebesgue measure, there exists U, an open set which satisfies

$$m_p(E_1) > (1 - 10^{-p}) m_p(U),\ U \supseteq E_1.$$



Each point of E is contained in balls of $\mathcal{F}$ of arbitrarily small radii and so there exists a covering of E with balls of $\mathcal{F}$ which are themselves contained in U. Therefore, by the Vitali covering theorem, there exist disjoint balls, $\{B_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ such that

$$E \subseteq \cup_{j=1}^{\infty} \widehat{B}_j,\ B_j \subseteq U.$$

Therefore,

$$\begin{aligned} m_p(E_1) &= \overline{m_p}(E) \leq m_p\left(\cup_{j=1}^{\infty}\widehat{B}_j\right) \leq \sum_j m_p\left(\widehat{B}_j\right) \\ &= 5^p \sum_j m_p(B_j) = 5^p m_p\left(\cup_{j=1}^{\infty}B_j\right) \end{aligned}$$

Then

$$\begin{aligned} m_p(E_1) &> (1-10^{-p})m_p(U) \\ &\geq (1-10^{-p})[m_p(E_1 \setminus \cup_{j=1}^{\infty}B_j) + m_p(\cup_{j=1}^{\infty}B_j)] \\ &\geq (1-10^{-p})[m_p(E_1 \setminus \cup_{j=1}^{\infty}B_j) + 5^{-p}\overbrace{\overline{m_p}(E)}^{=m_p(E_1)}]. \end{aligned}$$

and so

$$\left(1-\left(1-10^{-p}\right)5^{-p}\right)m_p(E_1) \geq (1-10^{-p})m_p(E_1 \setminus \cup_{j=1}^{\infty}B_j)$$

which implies

$$m_p(E_1 \setminus \cup_{j=1}^{\infty}B_j) \leq \frac{(1-(1-10^{-p})5^{-p})}{(1-10^{-p})}m_p(E_1)$$

Now a short computation shows

$$0 < \frac{(1-(1-10^{-p})5^{-p})}{(1-10^{-p})} < 1$$

Hence, denoting by θ_p a number such that

$$\frac{(1-(1-10^{-p})5^{-p})}{(1-10^{-p})} < \theta_p < 1,$$

$$\overline{m_p}\left(E \setminus \cup_{j=1}^{\infty}B_j\right) \leq m_p(E_1 \setminus \cup_{j=1}^{\infty}B_j) < \theta_p m_p(E_1) = \theta_p \overline{m_p}(E)$$

Now using Theorem 7.3.2 on Page 163 there exists N_1 large enough that

$$\theta_p\overline{m_p}(E) \geq m_p(E_1 \setminus \cup_{j=1}^{N_1}B_j) \geq \overline{m_p}(E \setminus \cup_{j=1}^{N_1}B_j) \tag{9.7}$$

Let $\mathcal{F}_1 = \{B \in \mathcal{F} : B_j \cap B = \emptyset,\ j = 1, \cdots, N_1\}$. If $E \setminus \cup_{j=1}^{N_1}B_j = \emptyset$, then $\mathcal{F}_1 = \emptyset$ and

$$\overline{m_p}\left(E \setminus \cup_{j=1}^{N_1}B_j\right) = 0$$

Therefore, in this case let $B_k = \emptyset$ for all $k > N_1$. Consider the case where

$$E \setminus \cup_{j=1}^{N_1}B_j \neq \emptyset.$$

In this case, since the balls are closed and $\mathcal{F}$ is a Vitali cover, $\mathcal{F}_1 \neq \emptyset$ and covers $E \setminus \cup_{j=1}^{N_1}B_j$ in the sense of Vitali. Repeat the same argument, letting $E \setminus \cup_{j=1}^{N_1}B_j$ play the role of E. (You pick a different E_1 whose measure equals the outer measure of $E \setminus \cup_{j=1}^{N_1}B_j$ and proceed as before.) Then choosing B_j for $j = N_1+1, \cdots, N_2$ as in the above argument,

$$\theta_p\overline{m_p}(E \setminus \cup_{j=1}^{N_1}B_j) \geq \overline{m_p}(E \setminus \cup_{j=1}^{N_2}B_j)$$

and so from 9.7,

$$\theta_p^2\overline{m_p}(E) \geq \overline{m_p}(E \setminus \cup_{j=1}^{N_2}B_j).$$

Continuing this way

$$\theta_p^k\overline{m_p}(E) \geq \overline{m_p}\left(E \setminus \cup_{j=1}^{N_k}B_j\right).$$

If it is ever the case that $E \setminus \cup_{j=1}^{N_k} B_j = \emptyset$, then as in the above argument,

$$\overline{m_p}\left(E \setminus \cup_{j=1}^{N_k} B_j\right) = 0.$$

Otherwise, the process continues and

$$\overline{m_p}\left(E \setminus \cup_{j=1}^{\infty} B_j\right) \leq \overline{m_p}\left(E \setminus \cup_{j=1}^{N_k} B_j\right) \leq \theta_p^k \overline{m_p}(E)$$

for every $k \in \mathbb{N}$. Therefore, the conclusion holds in this case also because $\theta_p < 1$. This proves the theorem. ■

There is an obvious corollary which removes the assumption that $0 < \overline{m_p}(E)$.

Corollary 9.7.3 *Let $E \subseteq \mathbb{R}^p$ and suppose $\overline{m_p}(E) < \infty$ where $\overline{m_p}$ is the outer measure determined by m_p, p dimensional Lebesgue measure, and let $\mathcal{F}$, be a collection of closed balls of bounded radii such that$\mathcal{F}$ covers E in the sense of Vitali. Then there exists a countable collection of disjoint balls from $\mathcal{F}$, $\{B_j\}_{j=1}^{\infty}$, such that $\overline{m_p}(E \setminus \cup_{j=1}^{\infty} B_j) = 0$.*

Proof: If $0 = \overline{m_p}(E)$ you simply pick any ball from $\mathcal{F}$ for your collection of disjoint balls. ■

It is also not hard to remove the assumption that $\overline{m_p}(E) < \infty$.

Corollary 9.7.4 *Let $E \subseteq \mathbb{R}^p$ and let $\mathcal{F}$, be a collection of closed balls of bounded radii such that $\mathcal{F}$ covers E in the sense of Vitali. Then there exists a countable collection of disjoint balls from $\mathcal{F}$, $\{B_j\}_{j=1}^{\infty}$, such that $\overline{m_p}(E \setminus \cup_{j=1}^{\infty} B_j) = 0$.*

Proof: Let $R_m \equiv (-m, m)^p$ be the open rectangle having sides of length $2m$ which is centered at $\mathbf{0}$ and let $R_0 = \emptyset$. Let $H_m \equiv \overline{R_m} \setminus R_m$. Since both $\overline{R_m}$ and R_m have the same measure, $(2m)^p$, it follows $m_p(H_m) = 0$. Now for all $k \in \mathbb{N}$, $R_k \subseteq \overline{R_k} \subseteq R_{k+1}$. Consider the disjoint open sets $U_k \equiv R_{k+1} \setminus \overline{R_k}$. Thus $\mathbb{R}^p = \cup_{k=0}^{\infty} U_k \cup N$ where N is a set of measure zero equal to the union of the H_k. Let $\mathcal{F}_k$ denote those balls of $\mathcal{F}$ which are contained in U_k and let $E_k \equiv U_k \cap E$. Then from Theorem 9.7.2, there exists a sequence of disjoint balls, $D_k \equiv \left\{B_i^k\right\}_{i=1}^{\infty}$ of $\mathcal{F}_k$ such that $\overline{m_p}(E_k \setminus \cup_{j=1}^{\infty} B_j^k) = 0$. Letting $\{B_i\}_{i=1}^{\infty}$ be an enumeration of all the balls of $\cup_k D_k$, it follows that

$$\overline{m_p}(E \setminus \cup_{j=1}^{\infty} B_j) \leq m_p(N) + \sum_{k=1}^{\infty} \overline{m_p}(E_k \setminus \cup_{j=1}^{\infty} B_j^k) = 0.$$

■

Also, you don't have to assume the balls are closed.

Corollary 9.7.5 *Let $E \subseteq \mathbb{R}^p$ and let $\mathcal{F}$, be a collection of open balls of bounded radii such that $\mathcal{F}$ covers E in the sense of Vitali. Then there exists a countable collection of disjoint balls from $\mathcal{F}$, $\{B_j\}_{j=1}^{\infty}$, such that $\overline{m_p}(E \setminus \cup_{j=1}^{\infty} B_j) = 0$.*

Proof: Let $\overline{\mathcal{F}}$ be the collection of closures of balls in $\mathcal{F}$. Then $\overline{\mathcal{F}}$ covers E in the sense of Vitali and so from Corollary 9.7.4 there exists a sequence of disjoint closed balls from $\overline{\mathcal{F}}$ satisfying $\overline{m_p}\left(E \setminus \cup_{i=1}^{\infty} \overline{B_i}\right) = 0$. Now boundaries of the balls, B_i have measure zero and so $\{B_i\}$ is a sequence of disjoint open balls satisfying $\overline{m_p}(E \setminus \cup_{i=1}^{\infty} B_i) = 0$. The reason for this is that

$$\left(E \setminus \cup_{i=1}^{\infty} B_i\right) \setminus \left(E \setminus \cup_{i=1}^{\infty} \overline{B_i}\right) \subseteq \cup_{i=1}^{\infty} \overline{B_i} \setminus \cup_{i=1}^{\infty} B_i \subseteq \cup_{i=1}^{\infty} \overline{B_i} \setminus B_i,$$

a set of measure zero. Therefore,

$$E \setminus \cup_{i=1}^{\infty} B_i \subseteq \left(E \setminus \cup_{i=1}^{\infty} \overline{B_i}\right) \cup \left(\cup_{i=1}^{\infty} \overline{B_i} \setminus B_i\right)$$

and so

$$\begin{aligned}\overline{m_p}\left(E \setminus \cup_{i=1}^{\infty} B_i\right) &\leq \overline{m_p}\left(E \setminus \cup_{i=1}^{\infty} \overline{B_i}\right) + m_p\left(\cup_{i=1}^{\infty} \overline{B_i} \setminus B_i\right) \\ &= \overline{m_p}\left(E \setminus \cup_{i=1}^{\infty} \overline{B_i}\right) = 0.\end{aligned}$$

■

This implies you can fill up an open set with balls which cover the open set in the sense of Vitali.

Corollary 9.7.6 *Let $U \subseteq \mathbb{R}^p$ be an open set and let $\mathcal{F}$ be a collection of closed or even open balls of bounded radii contained in U such that $\mathcal{F}$ covers U in the sense of Vitali. Then there exists a countable collection of disjoint balls from $\mathcal{F}$, $\{B_j\}_{j=1}^{\infty}$, such that $\overline{m_p}(U \setminus \cup_{j=1}^{\infty} B_j) = 0$.*

9.8 Change Of Variables For Linear Maps

To begin with certain kinds of functions map measurable sets to measurable sets. It will be assumed that U is an open set in $\mathbb{R}^p$ and that $\mathbf{h} : U \to \mathbb{R}^p$ satisfies

$$D\mathbf{h}(\mathbf{x}) \text{ exists for all } \mathbf{x} \in U, \tag{9.8}$$

Note that if

$$\mathbf{h}(\mathbf{x}) = L\mathbf{x}$$

where $L \in \mathcal{L}(\mathbb{R}^p, \mathbb{R}^p)$, then L is included in 9.8 because

$$L(\mathbf{x} + \mathbf{v}) = L(\mathbf{x}) + L(\mathbf{v}) + \mathbf{o}(\mathbf{v})$$

In fact, $\mathbf{o}(\mathbf{v}) = \mathbf{0}$.

It is convenient in the following lemma to use the norm on $\mathbb{R}^p$ given by

$$||\mathbf{x}|| = \max\{|x_k| : k = 1, 2, \cdots, p\}.$$

Thus $B(\mathbf{x}, r)$ is the open box,

$$\prod_{k=1}^{p} (x_k - r, x_k + r)$$

and so $m_p(B(\mathbf{x}, r)) = (2r)^p$.

Lemma 9.8.1 *Let $\mathbf{h}$ satisfy 9.8. If $T \subseteq U$ and $m_p(T) = 0$, then $m_p(\mathbf{h}(T)) = 0$.*

Proof: Let

$$T_k \equiv \{\mathbf{x} \in T : ||D\mathbf{h}(\mathbf{x})|| < k\}$$

and let $\varepsilon > 0$ be given. Now by outer regularity, there exists an open set V, containing T_k which is contained in U such that $m_p(V) < \varepsilon$. Let $\mathbf{x} \in T_k$. Then by differentiability,

$$\mathbf{h}(\mathbf{x} + \mathbf{v}) = \mathbf{h}(\mathbf{x}) + D\mathbf{h}(\mathbf{x})\mathbf{v} + o(\mathbf{v})$$

and so there exist arbitrarily small $r_{\mathbf{x}} < 1$ such that $B(\mathbf{x}, 5r_{\mathbf{x}}) \subseteq V$ and whenever $||\mathbf{v}|| \leq 5r_{\mathbf{x}}$, $||o(\mathbf{v})|| < k||\mathbf{v}||$. Thus

$$\mathbf{h}(B(\mathbf{x}, 5r_{\mathbf{x}})) \subseteq B(\mathbf{h}(\mathbf{x}), 6kr_{\mathbf{x}}).$$

From the Vitali covering theorem, there exists a countable disjoint sequence of these balls, $\{B(\mathbf{x}_i, r_i)\}_{i=1}^{\infty}$ such that $\{B(\mathbf{x}_i, 5r_i)\}_{i=1}^{\infty} = \left\{\widehat{B_i}\right\}_{i=1}^{\infty}$ covers T_k Then letting $\overline{m_p}$ denote the outer measure determined by m_p,

$$\overline{m_p}(\mathbf{h}(T_k)) \leq \overline{m_p}\left(\mathbf{h}\left(\cup_{i=1}^{\infty}\widehat{B}_i\right)\right)$$

$$\leq \sum_{i=1}^{\infty}\overline{m_p}\left(\mathbf{h}\left(\widehat{B}_i\right)\right) \leq \sum_{i=1}^{\infty} m_p(B(\mathbf{h}(\mathbf{x}_i), 6kr_{\mathbf{x}_i}))$$

$$\begin{aligned} &= \sum_{i=1}^{\infty} m_p(B(\mathbf{x}_i, 6kr_{\mathbf{x}_i})) = (6k)^p \sum_{i=1}^{\infty} m_p(B(\mathbf{x}_i, r_{\mathbf{x}_i})) \\ &\leq (6k)^p m_p(V) \leq (6k)^p \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, this shows $m_p(\mathbf{h}(T_k)) = 0$. Now

$$m_p(\mathbf{h}(T)) = \lim_{k\to\infty} m_p(\mathbf{h}(T_k)) = 0.$$

■

A somewhat easier result is the following about Lipschitz continuous functions.

Corollary 9.8.2 *In case* $\mathbf{h}$ *is Lipschitz,*

$$\|\mathbf{h}(\mathbf{x}) - \mathbf{h}(\mathbf{y})\| \leq K\|\mathbf{x} - \mathbf{y}\|$$

then the same conclusion holds as in Lemma 9.8.1.

Proof: In this case, $\|\mathbf{h}(\mathbf{x}+\mathbf{v}) - \mathbf{h}(\mathbf{x})\| \leq K\|\mathbf{v}\|$ and you can simply let $T \subseteq V$ where

$$m_p(V) < \varepsilon/(K^p 5^p).$$

Then there is a countable disjoint sequence of balls $\{B_i\}$ such that $\left\{\hat{B}_i\right\}$ covers T and each ball B_i is contained in V. Then $\mathbf{h}\left(\hat{B}_i\right) \subseteq B(\mathbf{h}(\mathbf{x}_i), 5K)$ and so

$$\bar{m}_p(\mathbf{h}(T)) \leq \sum_{i=1}^{\infty} m_p\left(\mathbf{h}\left(\hat{B}_i\right)\right) \leq 5^p K^p \sum_{i=1}^{\infty} m_p(B_i) \leq K^p 5^p m_p(V) < \varepsilon$$

Since ε is arbitrary, this shows that $\mathbf{h}(T)$ is measurable and $m_p(\mathbf{h}(T)) = 0$. ■

Lemma 9.8.3 *Let* $\mathbf{h}$ *satisfy 9.8. If* S *is a Lebesgue measurable subset of* U*, then* $\mathbf{h}(S)$ *is Lebesgue measurable.*

Proof: By Theorem 9.4.2 there exists F which is a countable union of compact sets $F = \cup_{k=1}^{\infty} K_k$ such that

$$F \subseteq S,\ m_p(S \setminus F) = 0.$$

Then since $\mathbf{h}$ is continuous

$$\mathbf{h}(F) = \cup_k \mathbf{h}(K_k) \in \mathcal{B}(\mathbb{R}^p)$$

because the continuous image of a compact set is compact. Also, $\mathbf{h}(S \setminus F)$ is a set of measure zero by Lemma 9.8.1 and so

$$\mathbf{h}(S) = \mathbf{h}(F) \cup \mathbf{h}(S \setminus F) \in \mathcal{F}_p$$

because it is the union of two sets which are in $\mathcal{F}_p$. This proves the lemma. ■

In particular, this proves the following corollary.

Corollary 9.8.4 *Suppose $A \in \mathcal{L}(\mathbb{R}^p, \mathbb{R}^p)$. Then if S is a Lebesgue measurable set, it follows AS is also a Lebesgue measurable set.*

In the next lemma, the norm used for defining balls will be the usual norm,

$$|\mathbf{x}| = \left(\sum_{k=1}^{p} |x_k|^2 \right)^{1/2}.$$

Thus a unitary transformation preserves distances measured with respect to this norm. In particular, if R is unitary, $(R^*R = RR^* = I)$ then

$$R(B(\mathbf{0}, r)) = B(\mathbf{0}, r).$$

Lemma 9.8.5 *Let R be unitary and let V be a an open set. Then $m_p(RV) = m_p(V)$.*

Proof: First assume V is a bounded open set. By Corollary 9.7.6 there is a disjoint sequence of closed balls, $\{B_i\}$ such that $U = \cup_{i=1}^{\infty} B_i \cup N$ where $m_p(N) = 0$. Denote by $\mathbf{x}_i$ the center of B_i and let r_i be the radius of B_i. Then by Lemma 9.8.1 $m_p(RV) = \sum_{i=1}^{\infty} m_p(RB_i)$. Now by invariance of translation of Lebesgue measure, this equals

$$\sum_{i=1}^{\infty} m_p(RB_i - R\mathbf{x}_i) = \sum_{i=1}^{\infty} m_p(RB(\mathbf{0}, r_i)).$$

Since R is unitary, it preserves all distances and so $RB(\mathbf{0}, r_i) = B(\mathbf{0}, r_i)$ and therefore,

$$m_p(RV) = \sum_{i=1}^{\infty} m_p(B(\mathbf{0}, r_i)) = \sum_{i=1}^{\infty} m_p(B_i) = m_p(V).$$

This proves the lemma in the case that V is bounded. Suppose now that V is just an open set. Let $V_k = V \cap B(\mathbf{0}, k)$. Then $m_p(RV_k) = m_p(V_k)$. Letting $k \to \infty$, this yields the desired conclusion. This proves the lemma in the case that V is open. ■

Lemma 9.8.6 *Let E be Lebesgue measurable set in $\mathbb{R}^p$ and let R be unitary. Then $m_p(RE) = m_p(E)$.*

Proof: Let $\mathcal{K}$ be the open sets. Thus $\mathcal{K}$ is a π system. Let $\mathcal{G}$ denote those Borel sets F such that for each $n \in \mathbb{N}$,

$$m_p(R(F \cap (-n, n)^p)) = m_n(F \cap (-n, n)^p).$$

Thus $\mathcal{G}$ contains $\mathcal{K}$ from Lemma 9.8.5. It is also routine to verify $\mathcal{G}$ is closed with respect to complements and countable disjoint unions. Therefore from the π systems lemma,

$$\mathcal{G} \supseteq \sigma(\mathcal{K}) = \mathcal{B}(\mathbb{R}^p) \supseteq \mathcal{G}$$

and this proves the lemma whenever $E \in \mathcal{B}(\mathbb{R}^p)$. If E is only in $\mathcal{F}_p$, it follows from Theorem 9.4.2

$$E = F \cup N$$

where $m_p(N) = 0$ and F is a countable union of compact sets. Thus by Lemma 9.8.1

$$m_p(RE) = m_p(RF) + m_p(RN) = m_p(RF) = m_p(F) = m_p(E).$$

This proves the theorem. ■

Lemma 9.8.7 *Let $D \in \mathcal{L}(\mathbb{R}^p, \mathbb{R}^p)$ be of the form*

$$D = \sum_j d_j \mathbf{e}_j \mathbf{e}_j$$

where $d_j \geq 0$ and $\{\mathbf{e}_j\}$ is the usual orthonormal basis of $\mathbb{R}^p$. Then for all $E \in \mathcal{F}_p$

$$m_p(DE) = |\det(D)| \, m_p(E).$$

Proof: Let $\mathcal{K}$ consist of open sets of the form

$$\prod_{k=1}^p (a_k, b_k) \equiv \left\{ \sum_{k=1}^p x_k \mathbf{e}_k \text{ such that } x_k \in (a_k, b_k) \right\}$$

Hence

$$\begin{aligned} D\left(\prod_{k=1}^p (a_k, b_k) \right) &= \left\{ \sum_{k=1}^p d_k x_k \mathbf{e}_k \text{ such that } x_k \in (a_k, b_k) \right\} \\ &= \prod_{k=1}^p (d_k a_k, d_k b_k). \end{aligned}$$

It follows

$$\begin{aligned} m_p\left(D\left(\prod_{k=1}^p (a_k, b_k) \right) \right) &= \left(\prod_{k=1}^p d_k \right) \left(\prod_{k=1}^p (b_k - a_k) \right) \\ &= |\det(D)| \, m_p\left(\prod_{k=1}^p (a_k, b_k) \right). \end{aligned}$$

Now let $\mathcal{G}$ consist of Borel sets F with the property that

$$m_p(D(F \cap (-n,n)^p)) = |\det(D)| \, m_p(F \cap (-n,n)^p).$$

Thus $\mathcal{K} \subseteq \mathcal{G}$.

Suppose now that $F \in \mathcal{G}$ and first assume D is one to one. Then

$$m_p\left(D\left(F^C \cap (-n,n)^p\right)\right) + m_p(D(F \cap (-n,n)^p)) = m_p(D(-n,n)^p)$$

and so

$$m_p\left(D\left(F^C \cap (-n,n)^p\right)\right) + |\det(D)| \, m_p(F \cap (-n,n)^p) = |\det(D)| \, m_p((-n,n)^p)$$

which shows

$$\begin{aligned} m_p\left(D\left(F^C \cap (-n,n)^p\right)\right) &= |\det(D)| \left[m_p((-n,n)^p) - m_p(F \cap (-n,n)^p) \right] \\ &= |\det(D)| \, m_p\left(F^C \cap (-n,n)^p\right) \end{aligned}$$

In case D is not one to one, it follows some $d_j = 0$ and so $|\det(D)| = 0$ and

$$\begin{aligned} 0 &\leq m_p\left(D\left(F^C \cap (-n,n)^p\right)\right) \leq m_p(D(-n,n)^p) = \prod_{i=1}^p (d_i p + d_i p) = 0 \\ &= |\det(D)| \, m_p\left(F^C \cap (-n,n)^p\right) \end{aligned}$$

so $F^C \in \mathcal{G}$.

If $\{F_k\}$ is a sequence of disjoint sets of $\mathcal{G}$ and D is one to one

$$\begin{aligned} m_p\left(D\left(\cup_{k=1}^{\infty}F_k\cap(-n,n)^p\right)\right) &= \sum_{k=1}^{\infty}m_p\left(D\left(F_k\cap(-n,n)^p\right)\right) \\ &= |\det(D)|\sum_{k=1}^{\infty}m_p\left(F_k\cap(-n,n)^p\right) \\ &= |\det(D)|\,m_p\left(\cup_k F_k\cap(-n,n)^p\right). \end{aligned}$$

If D is not one to one, then $\det(D)=0$ and so the right side of the above equals 0. The left side is also equal to zero because it is no larger than

$$m_p\left(D\left(-n,n\right)^p\right)=0.$$

Thus $\mathcal{G}$ is closed with respect to complements and countable disjoint unions. Hence it contains $\sigma(\mathcal{K})$, the Borel sets. But also $\mathcal{G}\subseteq\mathcal{B}(\mathbb{R}^p)$ and so $\mathcal{G}$ equals $\mathcal{B}(\mathbb{R}^p)$. Letting $p\to\infty$ yields the conclusion of the lemma in case $E\in\mathcal{B}(\mathbb{R}^p)$.

Now for $E\in\mathcal{F}_p$ arbitrary, it follows from Theorem 9.4.2

$$E=F\cup N$$

where N is a set of measure zero and F is a countable union of compact sets. Hence as before,

$$\begin{aligned} m_p\left(D\left(E\right)\right) &= m_p\left(DF\cup DN\right)\le m_p\left(DF\right)+m_p\left(DN\right) \\ &= |\det(D)|\,m_p\left(F\right)=|\det(D)|\,m_p\left(E\right) \end{aligned}$$

Also from Theorem 9.4.2 there exists G Borel such that

$$G=E\cup S$$

where S is a set of measure zero. Therefore,

$$\begin{aligned} |\det(D)|\,m_p\left(E\right) &= |\det(D)|\,m_p\left(G\right)=m_p\left(DG\right) \\ &= m_p\left(DE\cup DS\right)\le m_p\left(DE\right)+m_p\left(DS\right) \\ &= m_p\left(DE\right) \end{aligned}$$

This proves the theorem. ■

The main result follows.

Theorem 9.8.8 *Let $E\in\mathcal{F}_p$ and let $A\in\mathcal{L}(\mathbb{R}^p,\mathbb{R}^p)$. Then*

$$m_p\left(AV\right)=|\det(A)|\,m_p\left(V\right).$$

Proof: Let RU be the right polar decomposition (Theorem 3.9.3 on Page 68) of A. Thus R is unitary and

$$U=\sum_k d_k\mathbf{w}_k\mathbf{w}_k$$

where each $d_k\ge 0$. It follows $|\det(A)|=|\det(U)|$ because

$$|\det(A)|=|\det(R)\det(U)|=|\det(R)|\,|\det(U)|=|\det(U)|.$$

Recall from Lemma 3.9.5 on Page 70 the determinant of a unitary transformation has absolute value equal to 1. Then from Lemma 9.8.6,

$$m_p\left(AE\right)=m_p\left(RUE\right)=m_p\left(UE\right).$$

Let

$$Q = \sum_j \mathbf{w}_j \mathbf{e}_j$$

and so by Lemma 3.8.21 on Page 65,

$$Q^* = \sum_k \mathbf{e}_k \mathbf{w}_k.$$

Thus Q and Q^* are both unitary and a simple computation shows

$$U = Q \sum_i d_i \mathbf{e}_i \mathbf{e}_i Q^* \equiv QDQ^*.$$

Do both sides to $\mathbf{w}_k$ and observe both sides give $d_k \mathbf{w}_k$. Since the two linear operators agree on a basis, they must be the same. Thus

$$|\det(D)| = |\det(U)| = |\det(A)|.$$

Therefore, from Lemma 9.8.6 and Lemma 9.8.7

$$\begin{aligned} m_p(AE) &= m_p(QDQ^*E) = m_p(DQ^*E) \\ &= |\det(D)|\, m_p(Q^*E) = |\det(A)|\, m_p(E). \end{aligned}$$

This proves the theorem. ■

9.9 Change Of Variables For C^1 Functions

In this section theorems are proved which yield change of variables formulas for C^1 functions. More general versions can be seen in Kuttler [27], Kuttler [28], and Rudin [35]. You can obtain more by exploiting the Radon Nikodym theorem and the Lebesgue fundamental theorem of calculus, two topics which are best studied in a real analysis course. Instead, I will present some good theorems using the Vitali covering theorem directly.

A basic version of the theorems to be presented is the following. If you like, let the balls be defined in terms of the norm

$$||\mathbf{x}|| \equiv \max\{|x_k| : k = 1, \cdots, p\}$$

Lemma 9.9.1 *Let U and V be bounded open sets in $\mathbb{R}^p$ and let $\mathbf{h}, \mathbf{h}^{-1}$ be C^1 functions such that $\mathbf{h}(U) = V$. Also let $f \in C_c(V)$. Then*

$$\int_V f(\mathbf{y})\, dm_p = \int_U f(\mathbf{h}(\mathbf{x}))\, |\det(D\mathbf{h}(\mathbf{x}))|\, dm_p$$

Proof: First note $\mathbf{h}^{-1}(\mathrm{spt}(f))$ is a closed subset of the bounded set, U and so it is compact. Thus $\mathbf{x} \to f(\mathbf{h}(\mathbf{x}))\, |\det(D\mathbf{h}(\mathbf{x}))|$ is bounded and continuous.

Let $\mathbf{x} \in U$. By the assumption that $\mathbf{h}$ and $\mathbf{h}^{-1}$ are C^1,

$$\begin{aligned} \mathbf{h}(\mathbf{x}+\mathbf{v}) - \mathbf{h}(\mathbf{x}) &= D\mathbf{h}(\mathbf{x})\,\mathbf{v} + \mathbf{o}(\mathbf{v}) \\ &= D\mathbf{h}(\mathbf{x})\left(\mathbf{v} + D\mathbf{h}^{-1}(\mathbf{h}(\mathbf{x}))\,\mathbf{o}(\mathbf{v})\right) \\ &= D\mathbf{h}(\mathbf{x})\left(\mathbf{v} + \mathbf{o}(\mathbf{v})\right) \end{aligned}$$

and so if $r > 0$ is small enough then $B(\mathbf{x}, r)$ is contained in U and

$$\mathbf{h}(B(\mathbf{x}, r)) - \mathbf{h}(\mathbf{x}) =$$

$$\mathbf{h}(\mathbf{x}+B(\mathbf{0},r)) - \mathbf{h}(\mathbf{x}) \subseteq D\mathbf{h}(\mathbf{x})(B(\mathbf{0},(1+\varepsilon)r)). \tag{9.9}$$

Making r still smaller if necessary, one can also obtain

$$|f(\mathbf{y}) - f(\mathbf{h}(\mathbf{x}))| < \varepsilon \tag{9.10}$$

for any $\mathbf{y} \in \mathbf{h}(B(\mathbf{x},r))$ and also

$$|f(\mathbf{h}(\mathbf{x}_1))|\det(D\mathbf{h}(\mathbf{x}_1))| - f(\mathbf{h}(\mathbf{x}))|\det(D\mathbf{h}(\mathbf{x}))|| < \varepsilon \tag{9.11}$$

whenever $\mathbf{x}_1 \in B(\mathbf{x},r)$. The collection of such balls is a Vitali cover of U. By Corollary 9.7.6 there is a sequence of disjoint closed balls $\{B_i\}$ such that $U = \cup_{i=1}^{\infty} B_i \cup N$ where $m_p(N) = 0$. Denote by $\mathbf{x}_i$ the center of B_i and r_i the radius. Then by Lemma 9.8.1, the monotone convergence theorem, and 9.9 - 9.11,

$$\begin{aligned}
\textstyle\int_V f(\mathbf{y})\,dm_p &= \textstyle\sum_{i=1}^{\infty} \int_{\mathbf{h}(B_i)} f(\mathbf{y})\,dm_p \\
&\leq \varepsilon m_p(V) + \textstyle\sum_{i=1}^{\infty} \int_{\mathbf{h}(B_i)} f(\mathbf{h}(\mathbf{x}_i))\,dm_p \\
&\leq \varepsilon m_p(V) + \textstyle\sum_{i=1}^{\infty} f(\mathbf{h}(\mathbf{x}_i))\, m_p(\mathbf{h}(B_i)) \\
&\leq \varepsilon m_p(V) + \textstyle\sum_{i=1}^{\infty} f(\mathbf{h}(\mathbf{x}_i))\, m_p(D\mathbf{h}(\mathbf{x}_i)(B(\mathbf{0},(1+\varepsilon)r_i))) \\
&= \varepsilon m_p(V) + (1+\varepsilon)^p \textstyle\sum_{i=1}^{\infty} \int_{B_i} f(\mathbf{h}(\mathbf{x}_i))\,|\det(D\mathbf{h}(\mathbf{x}_i))|\,dm_p \\
&\leq \varepsilon m_p(V) + (1+\varepsilon)^p \textstyle\sum_{i=1}^{\infty} \left(\int_{B_i} f(\mathbf{h}(\mathbf{x}))\,|\det(D\mathbf{h}(\mathbf{x}))|\,dm_p + \varepsilon m_p(B_i) \right) \\
&\leq \varepsilon m_p(V) + (1+\varepsilon)^p \textstyle\sum_{i=1}^{\infty} \int_{B_i} f(\mathbf{h}(\mathbf{x}))\,|\det(D\mathbf{h}(\mathbf{x}))|\,dm_p + (1+\varepsilon)^p \varepsilon m_p(U) \\
&= \varepsilon m_p(V) + (1+\varepsilon)^p \textstyle\int_U f(\mathbf{h}(\mathbf{x}))\,|\det(D\mathbf{h}(\mathbf{x}))|\,dm_p + (1+\varepsilon)^p \varepsilon m_p(U)
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, this shows

$$\int_V f(\mathbf{y})\,dm_p \leq \int_U f(\mathbf{h}(\mathbf{x}))\,|\det(D\mathbf{h}(\mathbf{x}))|\,dm_p \tag{9.12}$$

whenever $f \in C_c(V)$. Now $\mathbf{x} \to f(\mathbf{h}(\mathbf{x}))\,|\det(D\mathbf{h}(\mathbf{x}))|$ is in $C_c(U)$ and so using the same argument with U and V switching roles and replacing $\mathbf{h}$ with $\mathbf{h}^{-1}$,

$$\begin{aligned}
& \int_U f(\mathbf{h}(\mathbf{x}))\,|\det(D\mathbf{h}(\mathbf{x}))|\,dm_p \\
\leq & \int_V f\left(\mathbf{h}\left(\mathbf{h}^{-1}(\mathbf{y})\right)\right) \left|\det\left(D\mathbf{h}\left(\mathbf{h}^{-1}(\mathbf{y})\right)\right)\right| \left|\det\left(D\mathbf{h}^{-1}(\mathbf{y})\right)\right| dm_p \\
= & \int_V f(\mathbf{y})\,dm_p
\end{aligned}$$

by the chain rule. This with 9.12 proves the lemma. ■

The next task is to relax the assumption that f is continuous.

Corollary 9.9.2 *Let U and V be bounded open sets in $\mathbb{R}^p$ and let $\mathbf{h}, \mathbf{h}^{-1}$ be C^1 functions such that $\mathbf{h}(U) = V$. Also let $E \subseteq V$ be measurable. Then*

$$\int_V \mathcal{X}_E(\mathbf{y})\,dm_p = \int_U \mathcal{X}_E(\mathbf{h}(\mathbf{x}))\,|\det(D\mathbf{h}(\mathbf{x}))|\,dm_p.$$

Proof: First suppose $E \subseteq H \subseteq V$ where H is compact. By regularity, there exist compact sets K_k and a decreasing sequence of open sets $\overline{G_k} \subseteq V$ such that

$$K_k \subseteq E \subseteq G_k$$

and $m_p(G_k \setminus K_k) < 2^{-k}$. By Lemma 8.10.2, there exist f_k such that $K_k \prec f_k \prec G_k$. Then $f_k(\mathbf{y}) \to \mathcal{X}_E(\mathbf{y})$ a.e. because if $\mathbf{y}$ is such that convergence fails, it must be the case that $\mathbf{y}$ is in $G_k \setminus K_k$ for infinitely many k and $\sum_k m_p(G_k \setminus K_k) < \infty$. This set equals

$$N = \cap_{m=1}^{\infty} \cup_{k=m}^{\infty} G_k \setminus K_k$$

and so for each $m \in \mathbb{N}$

$$\begin{aligned} m_p(N) &\leq m_p(\cup_{k=m}^{\infty} G_k \setminus K_k) \\ &\leq \sum_{k=m}^{\infty} m_p(G_k \setminus K_k) < \sum_{k=m}^{\infty} 2^{-k} = 2^{-(m-1)} \end{aligned}$$

showing $m_p(N) = 0$.

Then $f_k(\mathbf{h}(\mathbf{x}))$ must converge to $\mathcal{X}_E(\mathbf{h}(\mathbf{x}))$ for all $\mathbf{x} \notin \mathbf{h}^{-1}(N)$, a set of measure zero by Lemma 9.8.1. Thus

$$\int_V f_k(\mathbf{y})\, dm_p = \int_U f_k(\mathbf{h}(\mathbf{x}))\, |\det(D\mathbf{h}(\mathbf{x}))|\, dm_p.$$

Since V is bounded, $\overline{G_1}$ is compact. Therefore, $|\det(D\mathbf{h}(\mathbf{x}))|$ is bounded independent of k and so, by the dominated convergence theorem, using a dominating function, $\mathcal{X}_V$ in the integral on the left and $\mathcal{X}_{G_1} |\det(D\mathbf{h})|$ on the right, it follows

$$\int_V \mathcal{X}_E(\mathbf{y})\, dm_p = \int_U \mathcal{X}_E(\mathbf{h}(\mathbf{x}))\, |\det(D\mathbf{h}(\mathbf{x}))|\, dm_p.$$

For an arbitrary measurable E, let $E_k = H_k \cap E$ replace E in the above with E_k and use the monotone convergence theorem letting $k \to \infty$. ■

You don't need to assume the open sets are bounded.

Corollary 9.9.3 *Let U and V be open sets in $\mathbb{R}^p$ and let $\mathbf{h}, \mathbf{h}^{-1}$ be C^1 functions such that $\mathbf{h}(U) = V$. Also let $E \subseteq V$ be measurable. Then*

$$\int_V \mathcal{X}_E(\mathbf{y})\, dm_p = \int_U \mathcal{X}_E(\mathbf{h}(\mathbf{x}))\, |\det(D\mathbf{h}(\mathbf{x}))|\, dm_p.$$

Proof: For each $\mathbf{x} \in U$, there exists $r_{\mathbf{x}}$ such that $\overline{B(\mathbf{x}, r_{\mathbf{x}})} \subseteq U$ and $r_{\mathbf{x}} < 1$. Then by the mean value inequality Theorem 6.5.2, it follows $\mathbf{h}(B(\mathbf{x}, r_{\mathbf{x}}))$ is also bounded. These balls, $B(\mathbf{x}, r_{\mathbf{x}})$ give a Vitali cover of U and so by Corollary 9.7.6 there is a sequence of these balls, $\{B_i\}$ such that they are disjoint, $\mathbf{h}(B_i)$ is bounded and

$$m_p(U \setminus \cup_i B_i) = 0.$$

It follows from Lemma 9.8.1 that $\mathbf{h}(U \setminus \cup_i B_i)$ also has measure zero. Then from Corollary 9.9.2

$$\begin{aligned} \int_V \mathcal{X}_E(\mathbf{y})\, dm_p &= \sum_i \int_{\mathbf{h}(B_i)} \mathcal{X}_{E \cap \mathbf{h}(B_i)}(\mathbf{y})\, dm_p \\ &= \sum_i \int_{B_i} \mathcal{X}_E(\mathbf{h}(\mathbf{x}))\, |\det(D\mathbf{h}(\mathbf{x}))|\, dm_p \\ &= \int_U \mathcal{X}_E(\mathbf{h}(\mathbf{x}))\, |\det(D\mathbf{h}(\mathbf{x}))|\, dm_p. \end{aligned}$$

This proves the corollary. ■

With this corollary, the main theorem follows.

Theorem 9.9.4 *Let U and V be open sets in $\mathbb{R}^p$ and let $\mathbf{h}, \mathbf{h}^{-1}$ be C^1 functions such that $\mathbf{h}(U) = V$. Then if g is a nonnegative Lebesgue measurable function,*

$$\int_V g(\mathbf{y})\, dm_p = \int_U g(\mathbf{h}(\mathbf{x}))\, |\det(D\mathbf{h}(\mathbf{x}))|\, dm_p. \tag{9.13}$$

Proof: From Corollary 9.9.3, 9.13 holds for any nonnegative simple function in place of g. In general, let $\{s_k\}$ be an increasing sequence of simple functions which converges to g pointwise. Then from the monotone convergence theorem

$$\begin{aligned}\int_V g(\mathbf{y})\, dm_p &= \lim_{k\to\infty} \int_V s_k dm_p = \lim_{k\to\infty} \int_U s_k(\mathbf{h}(\mathbf{x}))\, |\det(D\mathbf{h}(\mathbf{x}))|\, dm_p \\ &= \int_U g(\mathbf{h}(\mathbf{x}))\, |\det(D\mathbf{h}(\mathbf{x}))|\, dm_p.\end{aligned}$$

This proves the theorem. ■

Of course this theorem implies the following corollary by splitting up the function into the positive and negative parts of the real and imaginary parts.

Corollary 9.9.5 *Let U and V be open sets in $\mathbb{R}^p$ and let $\mathbf{h}, \mathbf{h}^{-1}$ be C^1 functions such that $\mathbf{h}(U) = V$. Let $g \in L^1(V)$. Then*

$$\int_V g(\mathbf{y})\, dm_p = \int_U g(\mathbf{h}(\mathbf{x}))\, |\det(D\mathbf{h}(\mathbf{x}))|\, dm_p.$$

This is a pretty good theorem but it isn't too hard to generalize it. In particular, it is not necessary to assume $\mathbf{h}^{-1}$ is C^1.

Lemma 9.9.6 *Suppose V is an $p-1$ dimensional subspace of $\mathbb{R}^p$ and K is a compact subset of V. Then letting*

$$K_\varepsilon \equiv \cup_{\mathbf{x}\in K} B(\mathbf{x},\varepsilon) = K + B(\mathbf{0},\varepsilon),$$

it follows that

$$m_p(K_\varepsilon) \le 2^p \varepsilon (\operatorname{diam}(K) + \varepsilon)^{p-1}.$$

Proof: Using the Gram Schmidt procedure, there exists an orthonormal basis for V, $\{\mathbf{v}_1, \cdots, \mathbf{v}_{p-1}\}$ and let

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_{p-1}, \mathbf{v}_p\}$$

be an orthonormal basis for $\mathbb{R}^p$. Now define a linear transformation, Q by $Q\mathbf{v}_i = \mathbf{e}_i$. Thus $QQ^* = Q^*Q = I$ and Q preserves all distances and is a unitary transformation because

$$\left| Q \sum_i a_i \mathbf{e}_i \right|^2 = \left| \sum_i a_i \mathbf{v}_i \right|^2 = \sum_i |a_i|^2 = \left| \sum_i a_i \mathbf{e}_i \right|^2.$$

Thus $m_p(K_\varepsilon) = m_p(QK_\varepsilon)$. Letting $\mathbf{k}_0 \in K$, it follows $K \subseteq B(\mathbf{k}_0, \operatorname{diam}(K))$ and so,

$$QK \subseteq B^{p-1}(Q\mathbf{k}_0, \operatorname{diam}(QK)) = B^{p-1}(Q\mathbf{k}_0, \operatorname{diam}(K))$$

where B^{p-1} refers to the ball taken with respect to the usual norm in $\mathbb{R}^{p-1}$. Every point of K_ε is within ε of some point of K and so it follows that every point of QK_ε is within ε of some point of QK. Therefore,

$$QK_\varepsilon \subseteq B^{p-1}(Q\mathbf{k}_0, \operatorname{diam}(QK) + \varepsilon) \times (-\varepsilon, \varepsilon),$$

To see this, let $\mathbf{x} \in QK_\varepsilon$. Then there exists $\mathbf{k} \in QK$ such that $|\mathbf{k}-\mathbf{x}| < \varepsilon$. Therefore, $|(x_1,\cdots,x_{p-1}) - (k_1,\cdots,k_{p-1})| < \varepsilon$ and $|x_p - k_p| < \varepsilon$ and so $\mathbf{x}$ is contained in the set on the right in the above inclusion because $k_p = 0$. However, the measure of the set on the right is smaller than

$$[2(\operatorname{diam}(QK)+\varepsilon)]^{p-1}(2\varepsilon) = 2^p[(\operatorname{diam}(K)+\varepsilon)]^{p-1}\varepsilon.$$

This proves the lemma. ■

Note this is a very sloppy estimate. You can certainly do much better but this estimate is sufficient to prove Sard's lemma which follows.

Definition 9.9.7 *If T, S are two nonempty sets in a normed vector space,*

$$\operatorname{dist}(S,T) \equiv \inf\{||\mathbf{s}-\mathbf{t}|| : \mathbf{s} \in S, \mathbf{t} \in T\}.$$

Lemma 9.9.8 *Let $\mathbf{h}$ be a C^1 function defined on an open set, $U \subseteq \mathbb{R}^p$ and let K be a compact subset of U. Then if $\varepsilon > 0$ is given, there exists $r_1 > 0$ such that if $|\mathbf{v}| \le r_1$, then for all $\mathbf{x} \in K$,*

$$|\mathbf{h}(\mathbf{x}+\mathbf{v}) - \mathbf{h}(\mathbf{x}) - D\mathbf{h}(\mathbf{x})\mathbf{v}| < \varepsilon|\mathbf{v}|.$$

Proof: Let $0 < \delta < \operatorname{dist}(K, U^C)$. Such a positive number exists because if there exists a sequence of points in $K, \{\mathbf{k}_k\}$ and points in $U^C, \{\mathbf{s}_k\}$ such that $|\mathbf{k}_k - \mathbf{s}_k| \to 0$, then you could take a subsequence, still denoted by k such that $\mathbf{k}_k \to \mathbf{k} \in K$ and then $\mathbf{s}_k \to \mathbf{k}$ also. But U^C is closed so $\mathbf{k} \in K \cap U^C$, a contradiction. Then for $|\mathbf{v}| < \delta$ it follows that for every $\mathbf{x} \in K$,

$$\mathbf{x}+t\mathbf{v} \in U$$

and

$$\begin{aligned}\frac{|\mathbf{h}(\mathbf{x}+\mathbf{v}) - \mathbf{h}(\mathbf{x}) - D\mathbf{h}(\mathbf{x})\mathbf{v}|}{|\mathbf{v}|} &\le \frac{\left|\int_0^1 D\mathbf{h}(\mathbf{x}+t\mathbf{v})\mathbf{v}dt - D\mathbf{h}(\mathbf{x})\mathbf{v}\right|}{|\mathbf{v}|} \\ &\le \frac{\int_0^1 |D\mathbf{h}(\mathbf{x}+t\mathbf{v})\mathbf{v} - D\mathbf{h}(\mathbf{x})\mathbf{v}|\,dt}{|\mathbf{v}|}.\end{aligned}$$

The integral in the above involves integrating componentwise. Thus $t \to D\mathbf{h}(\mathbf{x}+t\mathbf{v})\mathbf{v}$ is a function having values in $\mathbb{R}^p$

$$\begin{pmatrix} Dh_1(\mathbf{x}+t\mathbf{v})\mathbf{v} \\ \vdots \\ Dh_p(\mathbf{x}+t\mathbf{v})\mathbf{v} \end{pmatrix}$$

and the integral is defined by

$$\begin{pmatrix} \int_0^1 Dh_1(\mathbf{x}+t\mathbf{v})\mathbf{v}dt \\ \vdots \\ \int_0^1 Dh_p(\mathbf{x}+t\mathbf{v})\mathbf{v}dt \end{pmatrix}$$

Now from uniform continuity of $D\mathbf{h}$ on the compact set, $\{\mathbf{x} : \operatorname{dist}(\mathbf{x},K) \le \delta\}$ it follows there exists $r_1 < \delta$ such that if $|\mathbf{v}| \le r_1$, then $||D\mathbf{h}(\mathbf{x}+t\mathbf{v}) - D\mathbf{h}(\mathbf{x})|| < \varepsilon$ for every $\mathbf{x} \in K$. From the above formula, it follows that if $|\mathbf{v}| \le r_1$,

$$\begin{aligned}\frac{|\mathbf{h}(\mathbf{x}+\mathbf{v}) - \mathbf{h}(\mathbf{x}) - D\mathbf{h}(\mathbf{x})\mathbf{v}|}{|\mathbf{v}|} &\le \frac{\int_0^1 |D\mathbf{h}(\mathbf{x}+t\mathbf{v})\mathbf{v} - D\mathbf{h}(\mathbf{x})\mathbf{v}|\,dt}{|\mathbf{v}|} \\ &< \frac{\int_0^1 \varepsilon|\mathbf{v}|\,dt}{|\mathbf{v}|} = \varepsilon.\end{aligned}$$

This proves the lemma. ■

The following is Sard's lemma. In the proof, it does not matter which norm you use in defining balls.

Lemma 9.9.9 *(Sard) Let U be an open set in $\mathbb{R}^p$ and let $\mathbf{h}: U \to \mathbb{R}^p$ be C^1. Let*

$$Z \equiv \{\mathbf{x} \in U : \det D\mathbf{h}(\mathbf{x}) = 0\}.$$

Then $m_p(\mathbf{h}(Z)) = 0$.

Proof: Let $\{U_k\}_{k=1}^{\infty}$ be an increasing sequence of open sets whose closures are compact and whose union equals U and let $Z_k \equiv Z \cap \overline{U_k}$. To obtain such a sequence, let

$$U_k = \left\{\mathbf{x} \in U : \operatorname{dist}\left(\mathbf{x}, U^C\right) < \frac{1}{k}\right\} \cap B(\mathbf{0}, k).$$

First it is shown that $\mathbf{h}(Z_k)$ has measure zero. Let W be an open set contained in U_{k+1} which contains Z_k and satisfies

$$m_p(Z_k) + \varepsilon > m_p(W)$$

where here and elsewhere, $\varepsilon < 1$. Let

$$r = \operatorname{dist}\left(\overline{U_k}, U_{k+1}^C\right)$$

and let $r_1 > 0$ be a constant as in Lemma 9.9.8 such that whenever $\mathbf{x} \in \overline{U_k}$ and $0 < |\mathbf{v}| \leq r_1$,

$$|\mathbf{h}(\mathbf{x}+\mathbf{v}) - \mathbf{h}(\mathbf{x}) - D\mathbf{h}(\mathbf{x})\mathbf{v}| < \varepsilon |\mathbf{v}|. \tag{9.14}$$

Now the closures of balls which are contained in W and which have the property that their diameters are less than r_1 yield a Vitali covering of W. Therefore, by Corollary 9.7.6 there is a disjoint sequence of these closed balls, $\left\{\widetilde{B}_i\right\}$ such that

$$W = \cup_{i=1}^{\infty} \widetilde{B}_i \cup N$$

where N is a set of measure zero. Denote by $\{B_i\}$ those closed balls in this sequence which have nonempty intersection with Z_k, let d_i be the diameter of B_i, and let $\mathbf{z}_i$ be a point in $B_i \cap Z_k$. Since $\mathbf{z}_i \in Z_k$, it follows $D\mathbf{h}(\mathbf{z}_i) B(\mathbf{0}, d_i) = D_i$ where D_i is contained in a subspace, V which has dimension $p-1$ and the diameter of D_i is no larger than $2C_k d_i$ where

$$C_k \geq \max\{||D\mathbf{h}(\mathbf{x})|| : \mathbf{x} \in Z_k\}$$

Then by 9.14, if $\mathbf{z} \in B_i$,

$$\mathbf{h}(\mathbf{z}) - \mathbf{h}(\mathbf{z}_i) \in D_i + B(\mathbf{0}, \varepsilon d_i) \subseteq \overline{D_i} + B(\mathbf{0}, \varepsilon d_i).$$

Thus

$$\mathbf{h}(B_i) \subseteq \mathbf{h}(\mathbf{z}_i) + \overline{D_i} + B(\mathbf{0}, \varepsilon d_i)$$

By Lemma 9.9.6

$$\begin{aligned} m_p(\mathbf{h}(B_i)) &\leq 2^p (2C_k d_i + \varepsilon d_i)^{p-1} \varepsilon d_i \\ &\leq d_i^p \left(2^p [2C_k + \varepsilon]^{p-1}\right) \varepsilon \\ &\leq C_{p,k} m_p(B_i) \varepsilon. \end{aligned}$$

Therefore, by Lemma 9.8.1

$$\begin{aligned} m_p(\mathbf{h}(Z_k)) &\leq m_p(W) = \sum_i m_p(\mathbf{h}(B_i)) \leq C_{p,k}\varepsilon \sum_i m_p(B_i) \\ &\leq \varepsilon C_{p,k} m_p(W) \leq \varepsilon C_{p,k}(m_p(Z_k) + \varepsilon) \end{aligned}$$

Since ε is arbitrary, this shows $m_p(\mathbf{h}(Z_k)) = 0$ and so $0 = \lim_{k\to\infty} m_p(\mathbf{h}(Z_k)) = m_p(\mathbf{h}(Z))$. ■

With this important lemma, here is a generalization of Theorem 9.9.4.

Theorem 9.9.10 *Let U be an open set and let $\mathbf{h}$ be a $1-1$, $C^1(U)$ function with values in $\mathbb{R}^p$. Then if g is a nonnegative Lebesgue measurable function,*

$$\int_{\mathbf{h}(U)} g(\mathbf{y})\, dm_p = \int_U g(\mathbf{h}(\mathbf{x}))\, |\det(D\mathbf{h}(\mathbf{x}))|\, dm_p. \tag{9.15}$$

Proof: Let $Z = \{\mathbf{x} : \det(D\mathbf{h}(\mathbf{x})) = 0\}$, a closed set. Then by the inverse function theorem, $\mathbf{h}^{-1}$ is C^1 on $\mathbf{h}(U \setminus Z)$ and $\mathbf{h}(U \setminus Z)$ is an open set. Therefore, from Lemma 9.9.9, $\mathbf{h}(Z)$ has measure zero and so by Theorem 9.9.4,

$$\begin{aligned} \int_{\mathbf{h}(U)} g(\mathbf{y})\, dm_p &= \int_{\mathbf{h}(U\setminus Z)} g(\mathbf{y})\, dm_p = \int_{U\setminus Z} g(\mathbf{h}(\mathbf{x}))\, |\det(D\mathbf{h}(\mathbf{x}))|\, dm_p \\ &= \int_U g(\mathbf{h}(\mathbf{x}))\, |\det(D\mathbf{h}(\mathbf{x}))|\, dm_p. \end{aligned}$$

This proves the theorem. ■

Of course the next generalization considers the case when $\mathbf{h}$ is not even one to one.

9.10 Change Of Variables For Mappings Which Are Not One To One

Now suppose $\mathbf{h}$ is only C^1, not necessarily one to one. For

$$U_+ \equiv \{\mathbf{x} \in U : |\det D\mathbf{h}(x)| > 0\}$$

and Z the set where $|\det D\mathbf{h}(\mathbf{x})| = 0$, Lemma 9.9.9 implies $m_p(\mathbf{h}(Z)) = 0$. For $\mathbf{x} \in U_+$, the inverse function theorem implies there exists an open set $B_{\mathbf{x}} \subseteq U_+$, such that $\mathbf{h}$ is one to one on $B_{\mathbf{x}}$.

Let $\{B_i\}$ be a countable subset of $\{B_{\mathbf{x}}\}_{\mathbf{x}\in U_+}$ such that $U_+ = \cup_{i=1}^\infty B_i$. Let $E_1 = B_1$. If $E_1, \cdots, E_k$ have been chosen, $E_{k+1} = B_{k+1} \setminus \cup_{i=1}^k E_i$. Thus

$$\cup_{i=1}^\infty E_i = U_+, \ \ \mathbf{h} \text{ is one to one on } E_i, \ \ E_i \cap E_j = \emptyset,$$

and each E_i is a Borel set contained in the open set B_i. Now define

$$n(\mathbf{y}) \equiv \sum_{i=1}^\infty \mathcal{X}_{\mathbf{h}(E_i)}(\mathbf{y}) + \mathcal{X}_{\mathbf{h}(Z)}(\mathbf{y}).$$

The set, $\mathbf{h}(E_i), \mathbf{h}(Z)$ are measurable by Lemma 9.8.3. Thus $n(\cdot)$ is measurable.

Lemma 9.10.1 *Let $F \subseteq \mathbf{h}(U)$ be measurable. Then*

$$\int_{\mathbf{h}(U)} n(\mathbf{y}) \mathcal{X}_F(\mathbf{y}) dm_p = \int_U \mathcal{X}_F(\mathbf{h}(\mathbf{x}))|\det D\mathbf{h}(\mathbf{x})| dm_p.$$

Proof: Using Lemma 9.9.9 and the Monotone Convergence Theorem

$$\begin{aligned}
\int_{\mathbf{h}(U)} n(\mathbf{y})\mathcal{X}_F(\mathbf{y})dm_p &= \int_{\mathbf{h}(U)} \left(\sum_{i=1}^{\infty} \mathcal{X}_{\mathbf{h}(E_i)}(\mathbf{y}) + \overbrace{\mathcal{X}_{\mathbf{h}(Z)}(\mathbf{y})}^{m_p(\mathbf{h}(Z))=0} \right) \mathcal{X}_F(\mathbf{y})dm_p \\
&= \sum_{i=1}^{\infty} \int_{\mathbf{h}(U)} \mathcal{X}_{\mathbf{h}(E_i)}(\mathbf{y})\mathcal{X}_F(\mathbf{y})dm_p \\
&= \sum_{i=1}^{\infty} \int_{\mathbf{h}(B_i)} \mathcal{X}_{\mathbf{h}(E_i)}(\mathbf{y})\mathcal{X}_F(\mathbf{y})dm_p \\
&= \sum_{i=1}^{\infty} \int_{B_i} \mathcal{X}_{E_i}(\mathbf{x})\mathcal{X}_F(\mathbf{h}(\mathbf{x}))|\det D\mathbf{h}(\mathbf{x})|dm_p \\
&= \sum_{i=1}^{\infty} \int_{U} \mathcal{X}_{E_i}(\mathbf{x})\mathcal{X}_F(\mathbf{h}(\mathbf{x}))|\det D\mathbf{h}(\mathbf{x})|dm_p \\
&= \int_{U} \sum_{i=1}^{\infty} \mathcal{X}_{E_i}(\mathbf{x})\mathcal{X}_F(\mathbf{h}(\mathbf{x}))|\det D\mathbf{h}(\mathbf{x})|dm_p
\end{aligned}$$

$$= \int_{U_+} \mathcal{X}_F(\mathbf{h}(\mathbf{x}))|\det D\mathbf{h}(\mathbf{x})|dm_p = \int_{U} \mathcal{X}_F(\mathbf{h}(\mathbf{x}))|\det D\mathbf{h}(\mathbf{x})|dm_p.$$

This proves the lemma. ■

Definition 9.10.2 *For $\mathbf{y} \in \mathbf{h}(U)$, define a function, $\#$, according to the formula*

$$\#(\mathbf{y}) \equiv \textit{number of elements in } \mathbf{h}^{-1}(\mathbf{y}).$$

Observe that

$$\#(\mathbf{y}) = n(\mathbf{y}) \quad \text{a.e.} \tag{9.16}$$

because $n(\mathbf{y}) = \#(\mathbf{y})$ if $\mathbf{y} \notin \mathbf{h}(Z)$, a set of measure 0. Therefore, $\#$ is a measurable function because of completeness of Lebesgue measure.

Theorem 9.10.3 *Let $g \geq 0$, g measurable, and let $\mathbf{h}$ be $C^1(U)$. Then*

$$\int_{\mathbf{h}(U)} \#(\mathbf{y})g(\mathbf{y})dm_p = \int_{U} g(\mathbf{h}(\mathbf{x}))|\det D\mathbf{h}(\mathbf{x})|dm_p. \tag{9.17}$$

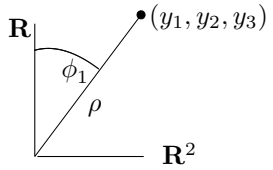
Proof: From 9.16 and Lemma 9.10.1, 9.17 holds for all g, a nonnegative simple function. Approximating an arbitrary measurable nonnegative function, g, with an increasing pointwise convergent sequence of simple functions and using the monotone convergence theorem, yields 9.17 for an arbitrary nonnegative measurable function, g. This proves the theorem. ■

9.11 Spherical Coordinates In p Dimensions

Sometimes there is a need to deal with spherical coordinates in more than three dimensions. In this section, this concept is defined and formulas are derived for these coordinate systems. Recall polar coordinates are of the form

$$\begin{aligned}
y_1 &= \rho\cos\theta \\
y_2 &= \rho\sin\theta
\end{aligned}$$

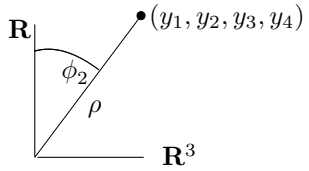
where $\rho > 0$ and $\theta \in \mathbb{R}$. Thus these transformation equations are not one to one but they are one to one on $(0,\infty)\times[0,2\pi)$. Here I am writing ρ in place of r to emphasize a pattern which is about to emerge. I will consider polar coordinates as spherical coordinates in two dimensions. I will also simply refer to such coordinate systems as polar coordinates regardless of the dimension. This is also the reason I am writing y_1 and y_2 instead of the more usual x and y. Now consider what happens when you go to three dimensions. The situation is depicted in the following picture.



From this picture, you see that $y_3 = \rho\cos\phi_1$. Also the distance between (y_1, y_2) and $(0,0)$ is $\rho\sin(\phi_1)$. Therefore, using polar coordinates to write (y_1, y_2) in terms of θ and this distance,

$$\begin{aligned} y_1 &= \rho\sin\phi_1\cos\theta, \\ y_2 &= \rho\sin\phi_1\sin\theta, \\ y_3 &= \rho\cos\phi_1. \end{aligned}$$

where $\phi_1 \in \mathbb{R}$ and the transformations are one to one if ϕ_1 is restricted to be in $[0,\pi]$. What was done is to replace ρ with $\rho\sin\phi_1$ and then to add in $y_3 = \rho\cos\phi_1$. Having done this, there is no reason to stop with three dimensions. Consider the following picture:



From this picture, you see that $y_4 = \rho\cos\phi_2$. Also the distance between (y_1, y_2, y_3) and $(0,0,0)$ is $\rho\sin(\phi_2)$. Therefore, using polar coordinates to write (y_1, y_2, y_3) in terms of θ, ϕ_1, and this distance,

$$\begin{aligned} y_1 &= \rho\sin\phi_2\sin\phi_1\cos\theta, \\ y_2 &= \rho\sin\phi_2\sin\phi_1\sin\theta, \\ y_3 &= \rho\sin\phi_2\cos\phi_1, \\ y_4 &= \rho\cos\phi_2 \end{aligned}$$

where $\phi_2 \in \mathbb{R}$ and the transformations will be one to one if

$$\phi_2, \phi_1 \in (0,\pi), \theta \in (0,2\pi), \rho \in (0,\infty).$$

Continuing this way, given spherical coordinates in $\mathbb{R}^p$, to get the spherical coordinates in $\mathbb{R}^{p+1}$, you let $y_{p+1} = \rho\cos\phi_{p-1}$ and then replace every occurance of ρ with $\rho\sin\phi_{p-1}$ to obtain $y_1 \cdots y_p$ in terms of $\phi_1, \phi_2, \cdots, \phi_{p-1}$,$\theta$, and ρ.

It is always the case that ρ measures the distance from the point in $\mathbb{R}^p$ to the origin in $\mathbb{R}^p$, $\mathbf{0}$. Each $\phi_i \in \mathbb{R}$ and the transformations will be one to one if each $\phi_i \in (0,\pi)$, and $\theta \in (0,2\pi)$. Denote by $\mathbf{h}_p\left(\rho, \vec{\phi}, \theta\right)$ the above transformation.

It can be shown using math induction and geometric reasoning that these coordinates map $\prod_{i=1}^{p-2}(0,\pi)\times(0,2\pi)\times(0,\infty)$ one to one onto an open subset of $\mathbb{R}^p$ which is

everything except for the set of measure zero $\Psi_p(N)$ where N results from having some ϕ_i equal to 0 or π or for $\rho = 0$ or for θ equal to either 2π or 0. Each of these are sets of Lebesgue measure zero and so their union is also a set of measure zero. You can see that $\mathbf{h}_p\left(\prod_{i=1}^{p-2}(0,\pi)\times(0,2\pi)\times(0,\infty)\right)$ omits the union of the coordinate axes except for maybe one of them. This is not important to the integral because it is just a set of measure zero.

Theorem 9.11.1 *Let $\mathbf{y} = \mathbf{h}_p\left(\vec{\phi},\theta,\rho\right)$ be the spherical coordinate transformations in $\mathbb{R}^p$. Then letting $A = \prod_{i=1}^{p-2}(0,\pi)\times(0,2\pi)$, it follows $\mathbf{h}$ maps $A\times(0,\infty)$ one to one onto all of $\mathbb{R}^p$ except a set of measure zero given by $\mathbf{h}_p(N)$ where N is the set of measure zero*

$$\left(\bar{A}\times[0,\infty)\right)\setminus(A\times(0,\infty))$$

Also $\left|\det D\mathbf{h}_p\left(\vec{\phi},\theta,\rho\right)\right|$ will always be of the form

$$\left|\det D\mathbf{h}_p\left(\vec{\phi},\theta,\rho\right)\right| = \rho^{p-1}\Phi\left(\vec{\phi},\theta\right). \tag{9.18}$$

where Φ is a continuous function of $\vec{\phi}$ and θ.[1] *Then if f is nonnegative and Lebesgue measurable,*

$$\int_{\mathbb{R}^p} f(\mathbf{y})\,dm_p = \int_{\mathbf{h}_p(A)} f(\mathbf{y})\,dm_p = \int_A f\left(\mathbf{h}_p\left(\vec{\phi},\theta,\rho\right)\right)\rho^{p-1}\Phi\left(\vec{\phi},\theta\right)dm_p \tag{9.19}$$

Furthermore whenever f is Borel measurable and nonnegative, one can apply Fubini's theorem and write

$$\int_{\mathbb{R}^p} f(\mathbf{y})\,dy = \int_0^\infty \rho^{p-1}\int_A f\left(\mathbf{h}\left(\vec{\phi},\theta,\rho\right)\right)\Phi\left(\vec{\phi},\theta\right)d\vec{\phi}d\theta d\rho \tag{9.20}$$

where here $d\vec{\phi}\,d\theta$ denotes dm_{p-1} on A. The same formulas hold if $f\in L^1(\mathbb{R}^p)$.

Proof: Formula 9.18 is obvious from the definition of the spherical coordinates because in the matrix of the derivative, there will be a ρ in $p-1$ columns. The first claim is also clear from the definition and math induction or from the geometry of the above description. It remains to verify 9.19 and 9.20. It is clear $\mathbf{h}_p$ maps $\bar{A}\times[0,\infty)$ onto $\mathbb{R}^p$. Since $\mathbf{h}_p$ is differentiable, it maps sets of measure zero to sets of measure zero. Then

$$\mathbb{R}^p = \mathbf{h}_p(N\cup A\times(0,\infty)) = \mathbf{h}_p(N)\cup\mathbf{h}_p(A\times(0,\infty)),$$

the union of a set of measure zero with $\mathbf{h}_p(A\times(0,\infty))$. Therefore, from the change of variables formula,

$$\int_{\mathbb{R}^p} f(\mathbf{y})\,dm_p = \int_{\mathbf{h}_p(A\times(0,\infty))} f(\mathbf{y})\,dm_p = \int_{A\times(0,\infty)} f\left(\mathbf{h}_p\left(\vec{\phi},\theta,\rho\right)\right)\rho^{p-1}\Phi\left(\vec{\phi},\theta\right)dm_p$$

which proves 9.19. This formula continues to hold if f is in $L^1(\mathbb{R}^p)$. Finally, if $f\geq 0$ or in $L^1(\mathbb{R}^n)$ and is Borel measurable, then it is $\mathcal{F}^p$ measurable as well. Recall that $\mathcal{F}^p$ includes the smallest σ algebra which contains products of open intervals. Hence $\mathcal{F}^p$ includes the Borel sets $\mathcal{B}(\mathbb{R}^p)$. Thus from the definition of m_p

$$\int_{A\times(0,\infty)} f\left(\mathbf{h}_p\left(\vec{\phi},\theta,\rho\right)\right)\rho^{p-1}\Phi\left(\vec{\phi},\theta\right)dm_p$$

[1] Actually it is only a function of the first but this is not important in what follows.

$$\begin{aligned} &= \int_{(0,\infty)} \int_A f\left(\mathbf{h}_p\left(\vec{\phi},\theta,\rho\right)\right) \rho^{p-1} \Phi\left(\vec{\phi},\theta\right) dm_{p-1} dm \\ &= \int_{(0,\infty)} \rho^{p-1} \int_A f\left(\mathbf{h}_p\left(\vec{\phi},\theta,\rho\right)\right) \Phi\left(\vec{\phi},\theta\right) dm_{p-1} dm \end{aligned}$$

Now the claim about $f \in L^1$ follows routinely from considering the positive and negative parts of the real and imaginary parts of f in the usual way. ■

Note that the above equals

$$\int_{\bar{A}\times[0,\infty)} f\left(\mathbf{h}_p\left(\vec{\phi},\theta,\rho\right)\right) \rho^{p-1} \Phi\left(\vec{\phi},\theta\right) dm_p$$

and the iterated integral is also equal to

$$\int_{[0,\infty)} \rho^{p-1} \int_{\bar{A}} f\left(\mathbf{h}_p\left(\vec{\phi},\theta,\rho\right)\right) \Phi\left(\vec{\phi},\theta\right) dm_{p-1} dm$$

because the difference is just a set of measure zero.

Notation 9.11.2 *Often this is written differently. Note that from the spherical coordinate formulas,* $f\left(\mathbf{h}\left(\vec{\phi},\theta,\rho\right)\right) = f(\rho\boldsymbol{\omega})$ *where* $|\boldsymbol{\omega}| = 1$*. Letting* S^{p-1} *denote the unit sphere,* $\{\boldsymbol{\omega} \in \mathbb{R}^p : |\boldsymbol{\omega}| = 1\}$*, the inside integral in the above formula is sometimes written as*

$$\int_{S^{p-1}} f(\rho\boldsymbol{\omega})\, d\sigma$$

where σ *is a measure on* S^{p-1}*. See [27] for another description of this measure. It isn't an important issue here. Either 9.20 or the formula*

$$\int_0^\infty \rho^{p-1} \left(\int_{S^{p-1}} f(\rho\boldsymbol{\omega})\, d\sigma \right) d\rho$$

will be referred to as polar coordinates and is very useful in establishing estimates. Here $\sigma\left(S^{p-1}\right) \equiv \int_A \Phi\left(\vec{\phi},\theta\right) dm_{p-1}$*.*

Example 9.11.3 *For what values of* s *is the integral* $\int_{B(\mathbf{0},R)} \left(1+|\mathbf{x}|^2\right)^s dy$ *bounded independent of* R*? Here* $B(\mathbf{0},R)$ *is the ball,* $\{\mathbf{x} \in \mathbb{R}^p : |\mathbf{x}| \le R\}$*.*

I think you can see immediately that s must be negative but exactly how negative? It turns out it depends on p and using polar coordinates, you can find just exactly what is needed. From the polar coordinates formula above,

$$\begin{aligned} \int_{B(\mathbf{0},R)} \left(1+|\mathbf{x}|^2\right)^s dy &= \int_0^R \int_{S^{p-1}} \left(1+\rho^2\right)^s \rho^{p-1} d\sigma d\rho \\ &= C_p \int_0^R \left(1+\rho^2\right)^s \rho^{p-1} d\rho \end{aligned}$$

Now the very hard problem has been reduced to considering an easy one variable problem of finding when

$$\int_0^R \rho^{p-1} \left(1+\rho^2\right)^s d\rho$$

is bounded independent of R. You need $2s + (p-1) < -1$ so you need $s < -p/2$.

9.12 Brouwer Fixed Point Theorem

The Brouwer fixed point theorem is one of the most significant theorems in mathematics. There exist relatively easy proofs of this important theorem. The proof I am giving here is the one given in Evans [14]. I think it is one of the shortest and easiest proofs of this important theorem. It is based on the following lemma which is an interesting result about cofactors of a matrix.

Recall that for A an $p \times p$ matrix, $\operatorname{cof}(A)_{ij}$ is the determinant of the matrix which results from deleting the i^{th} row and the j^{th} column and multiplying by $(-1)^{i+j}$. In the proof and in what follows, I am using $D\mathbf{g}$ to equal the matrix of the linear transformation $D\mathbf{g}$ taken with respect to the usual basis on $\mathbb{R}^p$. Thus

$$D\mathbf{g}(\mathbf{x}) = \sum_{ij} (D\mathbf{g})_{ij} \mathbf{e}_i \mathbf{e}_j$$

and recall that $(D\mathbf{g})_{ij} = \partial g_i / \partial x_j$ where $\mathbf{g} = \sum_i g_i \mathbf{e}_i$.

Lemma 9.12.1 *Let $\mathbf{g} : U \to \mathbb{R}^p$ be C^2 where U is an open subset of $\mathbb{R}^p$. Then*

$$\sum_{j=1}^{p} \operatorname{cof}(D\mathbf{g})_{ij,j} = 0,$$

where here $(D\mathbf{g})_{ij} \equiv g_{i,j} \equiv \frac{\partial g_i}{\partial x_j}$. Also, $\operatorname{cof}(D\mathbf{g})_{ij} = \frac{\partial \det(D\mathbf{g})}{\partial g_{i,j}}$.

Proof: From the cofactor expansion theorem,

$$\det(D\mathbf{g}) = \sum_{i=1}^{p} g_{i,j} \operatorname{cof}(D\mathbf{g})_{ij}$$

and so

$$\frac{\partial \det(D\mathbf{g})}{\partial g_{i,j}} = \operatorname{cof}(D\mathbf{g})_{ij} \tag{9.21}$$

which shows the last claim of the lemma. Also

$$\delta_{kj} \det(D\mathbf{g}) = \sum_{i} g_{i,k} (\operatorname{cof}(D\mathbf{g}))_{ij} \tag{9.22}$$

because if $k \neq j$ this is just the cofactor expansion of the determinant of a matrix in which the k^{th} and j^{th} columns are equal. Differentiate 9.22 with respect to x_j and sum on j. This yields

$$\sum_{r,s,j} \delta_{kj} \frac{\partial (\det D\mathbf{g})}{\partial g_{r,s}} g_{r,sj} = \sum_{ij} g_{i,kj} (\operatorname{cof}(D\mathbf{g}))_{ij} + \sum_{ij} g_{i,k} \operatorname{cof}(D\mathbf{g})_{ij,j}.$$

Hence, using $\delta_{kj} = 0$ if $j \neq k$ and 9.21,

$$\sum_{rs} (\operatorname{cof}(D\mathbf{g}))_{rs} g_{r,sk} = \sum_{rs} g_{r,ks} (\operatorname{cof}(D\mathbf{g}))_{rs} + \sum_{ij} g_{i,k} \operatorname{cof}(D\mathbf{g})_{ij,j}.$$

Subtracting the first sum on the right from both sides and using the equality of mixed partials,

$$\sum_{i} g_{i,k} \left(\sum_{j} (\operatorname{cof}(D\mathbf{g}))_{ij,j} \right) = 0.$$

If $\det(g_{i,k}) \neq 0$ so that $(g_{i,k})$ is invertible, this shows $\sum_j (\operatorname{cof}(D\mathbf{g}))_{ij,j} = 0$. If $\det(D\mathbf{g}) = 0$, let

$$\mathbf{g}_k(\mathbf{x}) = \mathbf{g}(\mathbf{x}) + \varepsilon_k \mathbf{x}$$

where $\varepsilon_k \to 0$ and $\det(D\mathbf{g} + \varepsilon_k I) \equiv \det(D\mathbf{g}_k) \neq 0$. Then

$$\sum_j (\operatorname{cof}(D\mathbf{g}))_{ij,j} = \lim_{k\to\infty} \sum_j (\operatorname{cof}(D\mathbf{g}_k))_{ij,j} = 0$$

and This proves the lemma. ■

Definition 9.12.2 *Let* $\mathbf{h}$ *be a function defined on an open set,* $U \subseteq \mathbb{R}^p$*. Then* $\mathbf{h} \in C^k(\overline{U})$ *if there exists a function* $\mathbf{g}$ *defined on an open set,* W *containng* $\overline{U}$ *such that* $\mathbf{g} = \mathbf{h}$ *on* U *and* $\mathbf{g}$ *is* $C^k(W)$*.*

In the following lemma, you could use any norm in defining the balls and everything would work the same but I have in mind the usual norm.

Lemma 9.12.3 *There does not exist* $\mathbf{h} \in C^2\left(\overline{B(\mathbf{0},R)}\right)$ *such that* $\mathbf{h} : \overline{B(\mathbf{0},R)} \to \partial B(\mathbf{0},R)$ *which also has the property that* $\mathbf{h}(\mathbf{x}) = \mathbf{x}$ *for all* $\mathbf{x} \in \partial B(\mathbf{0},R)$*. Such a function is called a retraction.*

Proof: Suppose such an $\mathbf{h}$ exists. Let $\lambda \in [0,1]$ and let $\mathbf{p}_\lambda(\mathbf{x}) \equiv \mathbf{x} + \lambda(\mathbf{h}(\mathbf{x}) - \mathbf{x})$. This function, $\mathbf{p}_\lambda$ is called a homotopy of the identity map and the retraction, $\mathbf{h}$. Let

$$I(\lambda) \equiv \int_{B(\mathbf{0},R)} \det(D\mathbf{p}_\lambda(\mathbf{x}))\, dx.$$

Then using the dominated convergence theorem,

$$\begin{aligned} I'(\lambda) &= \int_{B(\mathbf{0},R)} \sum_{i.j} \frac{\partial \det(D\mathbf{p}_\lambda(\mathbf{x}))}{\partial p_{\lambda i,j}} \frac{\partial p_{\lambda ij}(\mathbf{x})}{\partial \lambda} dx \\ &= \int_{B(\mathbf{0},R)} \sum_i \sum_j \frac{\partial \det(D\mathbf{p}_\lambda(\mathbf{x}))}{\partial p_{\lambda i,j}} (h_i(\mathbf{x}) - x_i)_{,j}\, dx \\ &= \int_{B(\mathbf{0},R)} \sum_i \sum_j \operatorname{cof}(D\mathbf{p}_\lambda(\mathbf{x}))_{ij} (h_i(\mathbf{x}) - x_i)_{,j}\, dx \end{aligned}$$

Now by assumption, $h_i(\mathbf{x}) = x_i$ on $\partial B(\mathbf{0},R)$ and so one can form iterated integrals and integrate by parts in each of the one dimensional integrals to obtain

$$I'(\lambda) = -\sum_i \int_{B(\mathbf{0},R)} \sum_j \operatorname{cof}(D\mathbf{p}_\lambda(\mathbf{x}))_{ij,j} (h_i(\mathbf{x}) - x_i)\, dx = 0.$$

Therefore, $I(\lambda)$ equals a constant. However,

$$I(0) = m_p(B(\mathbf{0},R)) > 0$$

but

$$I(1) = \int_{B(\mathbf{0},1)} \det(D\mathbf{h}(\mathbf{x}))\, dm_p = \int_{\partial B(\mathbf{0},1)} \#(\mathbf{y})\, dm_p = 0$$

because from polar coordinates or other elementary reasoning, $m_p(\partial B(\mathbf{0},1)) = 0$. This proves the lemma. ■

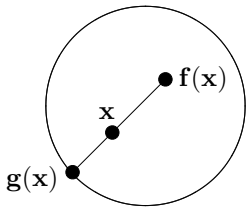
The following is the Brouwer fixed point theorem for C^2 maps.

Lemma 9.12.4 *If* $\mathbf{h} \in C^2\left(\overline{B(\mathbf{0},R)}\right)$ *and* $\mathbf{h} : \overline{B(\mathbf{0},R)} \to \overline{B(\mathbf{0},R)}$, *then* $\mathbf{h}$ *has a fixed point,* $\mathbf{x}$ *such that* $\mathbf{h}(\mathbf{x}) = \mathbf{x}$.

Proof: Suppose the lemma is not true. Then for all $\mathbf{x}$, $|\mathbf{x} - \mathbf{h}(\mathbf{x})| \neq 0$. Then define

$$\mathbf{g}(\mathbf{x}) = \mathbf{h}(\mathbf{x}) + \frac{\mathbf{x} - \mathbf{h}(\mathbf{x})}{|\mathbf{x} - \mathbf{h}(\mathbf{x})|} t(\mathbf{x})$$

where $t(\mathbf{x})$ is nonnegative and is chosen such that $\mathbf{g}(\mathbf{x}) \in \partial B(\mathbf{0},R)$. This mapping is illustrated in the following picture.



If $\mathbf{x} \to t(\mathbf{x})$ is C^2 near $\overline{B(\mathbf{0},R)}$, it will follow $\mathbf{g}$ is a C^2 retraction onto $\partial B(\mathbf{0},R)$ contrary to Lemma 9.12.3. Now $t(\mathbf{x})$ is the nonnegative solution, t to

$$H(\mathbf{x},\mathbf{t}) = |\mathbf{h}(\mathbf{x})|^2 + 2\left(\mathbf{h}(\mathbf{x}), \frac{\mathbf{x} - \mathbf{h}(\mathbf{x})}{|\mathbf{x} - \mathbf{h}(\mathbf{x})|}\right) t + t^2 = R^2 \tag{9.23}$$

Then

$$H_t(\mathbf{x},t) = 2\left(\mathbf{h}(\mathbf{x}), \frac{\mathbf{x} - \mathbf{h}(\mathbf{x})}{|\mathbf{x} - \mathbf{h}(\mathbf{x})|}\right) + 2t.$$

If this is nonzero for all $\mathbf{x}$ near $\overline{B(\mathbf{0},R)}$, it follows from the implicit function theorem that t is a C^2 function of $\mathbf{x}$. From 9.23

$$\begin{aligned} 2t &= -2\left(\mathbf{h}(\mathbf{x}), \frac{\mathbf{x} - \mathbf{h}(\mathbf{x})}{|\mathbf{x} - \mathbf{h}(\mathbf{x})|}\right) \\ &\pm\sqrt{4\left(\mathbf{h}(\mathbf{x}), \frac{\mathbf{x} - \mathbf{h}(\mathbf{x})}{|\mathbf{x} - \mathbf{h}(\mathbf{x})|}\right)^2 - 4\left(|\mathbf{h}(\mathbf{x})|^2 - R^2\right)} \end{aligned}$$

and so

$$\begin{aligned} H_t(\mathbf{x},t) &= 2t + 2\left(\mathbf{h}(\mathbf{x}), \frac{\mathbf{x} - \mathbf{h}(\mathbf{x})}{|\mathbf{x} - \mathbf{h}(\mathbf{x})|}\right) \\ &= \pm\sqrt{4\left(R^2 - |\mathbf{h}(\mathbf{x})|^2\right) + 4\left(\mathbf{h}(\mathbf{x}), \frac{\mathbf{x} - \mathbf{h}(\mathbf{x})}{|\mathbf{x} - \mathbf{h}(\mathbf{x})|}\right)^2} \end{aligned}$$

If $|\mathbf{h}(\mathbf{x})| < R$, this is nonzero. If $|\mathbf{h}(\mathbf{x})| = R$, then it is still nonzero unless

$$(\mathbf{h}(\mathbf{x}), \mathbf{x} - \mathbf{h}(\mathbf{x})) = 0.$$

But this cannot happen because the angle between $\mathbf{h}(\mathbf{x})$ and $\mathbf{x} - \mathbf{h}(\mathbf{x})$ cannot be $\pi/2$. Alternatively, if the above equals zero, you would need

$$(\mathbf{h}(\mathbf{x}), \mathbf{x}) = |\mathbf{h}(\mathbf{x})|^2 = R^2$$

which cannot happen unless $\mathbf{x} = \mathbf{h}(\mathbf{x})$ which is assumed not to happen. Therefore, $\mathbf{x} \to t(\mathbf{x})$ is C^2 near $\overline{B(\mathbf{0},R)}$ and so $\mathbf{g}(\mathbf{x})$ given above contradicts Lemma 9.12.3. This proves the lemma. ■

Now it is easy to prove the Brouwer fixed point theorem.

Theorem 9.12.5 *Let* $\mathbf{f} : \overline{B(\mathbf{0},R)} \to \overline{B(\mathbf{0},R)}$ *be continuous. Then* $\mathbf{f}$ *has a fixed point.*

Proof: If this is not so, there exists $\varepsilon > 0$ such that for all $\mathbf{x} \in \overline{B(\mathbf{0},R)}$,

$$|\mathbf{x} - \mathbf{f}(\mathbf{x})| > \varepsilon.$$

By the Weierstrass approximation theorem, there exists $\mathbf{h}$, a polynomial such that

$$\max\left\{|\mathbf{h}(\mathbf{x}) - \mathbf{f}(\mathbf{x})| : \mathbf{x} \in \overline{B(\mathbf{0},R)}\right\} < \frac{\varepsilon}{2}.$$

Then for all $\mathbf{x} \in \overline{B(\mathbf{0},R)}$,

$$|\mathbf{x} - \mathbf{h}(\mathbf{x})| \geq |\mathbf{x} - \mathbf{f}(\mathbf{x})| - |\mathbf{h}(\mathbf{x}) - \mathbf{f}(\mathbf{x})| > \varepsilon - \frac{\varepsilon}{2} = \frac{\varepsilon}{2}$$

contradicting Lemma 9.12.4. This proves the theorem. ■

9.13 Exercises

1. Recall the definition of $f_{\mathbf{y}}$. Prove that if $f \in L^1(\mathbb{R}^p)$, then

$$\lim_{\mathbf{y}\to\mathbf{0}} \int_{\mathbb{R}^p} |f - f_{\mathbf{y}}|\, dm_p = 0$$

   This is known as continuity of translation. **Hint:** Use the theorem about being able to approximate an arbitrary function in $L^1(\mathbb{R}^p)$ with a function in $C_c(\mathbb{R}^p)$.

2. Show that if $a, b \geq 0$ and if $p, q > 0$ such that

$$\frac{1}{p} + \frac{1}{q} = 1$$

   then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

   Hint: You might consider for fixed $a \geq 0$, the function $h(b) \equiv \frac{a^p}{p} + \frac{b^q}{q} - ab$ and find its minimum.

3. In the context of the previous problem, prove Holder's inequality. If f, g measurable functions, then

$$\int |f|\,|g|\, d\mu \leq \left(\int |f|^p\, d\mu\right)^{1/p} \left(\int |g|^q\, d\mu\right)^{1/q}$$

   Hint: If either of the factors on the right equals 0, explain why there is nothing to show. Now let $a = |f| / \left(\int |f|^p\, d\mu\right)^{1/p}$ and $b = |g| / \left(\int |g|^q\, d\mu\right)^{1/q}$. Apply the inequality of the previous problem.

4. Let E be a Lebesgue measurable set in $\mathbb{R}$. Suppose $m(E) > 0$. Consider the set

$$E - E = \{x - y : x \in E, y \in E\}.$$

   Show that $E - E$ contains an interval. **Hint:** Let

$$f(x) = \int \mathcal{X}_E(t)\mathcal{X}_E(x+t)dt.$$

   Explain why f is continuous at 0 and $f(0) > 0$ and use continuity of translation in L^1.

5. If $f \in L^1(\mathbb{R}^p)$, show there exists $g \in L^1(\mathbb{R}^p)$ such that g is also Borel measurable such that $g(\mathbf{x}) = f(\mathbf{x})$ for a.e. $\mathbf{x}$.

6. Suppose $f, g \in L^1(\mathbb{R}^p)$. Define $f * g(\mathbf{x})$ by

$$\int f(\mathbf{x}-\mathbf{y})\, g(\mathbf{y})\, dm_p(y).$$

Show this makes sense for a.e. $\mathbf{x}$ and that in fact for a.e. $\mathbf{x}$

$$\int |f(\mathbf{x}-\mathbf{y})|\, |g(\mathbf{y})|\, dm_p(y)$$

Next show

$$\int |f * g(\mathbf{x})|\, dm_p(\mathbf{x}) \leq \int |f|\, dm_p \int |g|\, dm_p.$$

Hint: Use Problem 5. Show first there is no problem if f, g are Borel measurable. The reason for this is that you can use Fubini's theorem to write

$$\begin{aligned} & \int\int |f(\mathbf{x}-\mathbf{y})|\, |g(\mathbf{y})|\, dm_p(y)\, dm_p(x) \\ = & \int\int |f(\mathbf{x}-\mathbf{y})|\, |g(\mathbf{y})|\, dm_p(x)\, dm_p(y) \\ = & \int |f(\mathbf{z})|\, dm_p \int |g(\mathbf{y})|\, dm_p. \end{aligned}$$

Explain. Then explain why if f and g are replaced by functions which are equal to f and g a.e. but are Borel measurable, the convolution is unchanged.

7. In the situation of Problem 6 Show $\mathbf{x} \to f * g(\mathbf{x})$ is continuous whenever g is also bounded. **Hint:** Use Problem 1.

8. Let $f : [0, \infty) \to \mathbb{R}$ be in $L^1(\mathbb{R}, m)$. The Laplace transform is given by $\widehat{f}(x) = \int_0^\infty e^{-xt} f(t) dt$. Let f, g be in $L^1(\mathbb{R}, m)$, and let $h(x) = \int_0^x f(x-t) g(t) dt$. Show $h \in L^1$, and $\widehat{h} = \widehat{f}\widehat{g}$.

9. Suppose A is covered by a finite collection of Balls, $\mathcal{F}$. Show that then there exists a disjoint collection of these balls, $\{B_i\}_{i=1}^p$, such that $A \subseteq \cup_{i=1}^p \widehat{B}_i$ where $\widehat{B}_i$ has the same center as B_i but 3 times the radius. **Hint:** Since the collection of balls is finite, they can be arranged in order of decreasing radius.

10. Let f be a function defined on an interval, (a, b). The Dini derivates are defined as

$$\begin{aligned} D_+ f(x) &\equiv \liminf_{h \to 0+} \frac{f(x+h) - f(x)}{h}, \\ D^+ f(x) &\equiv \limsup_{h \to 0+} \frac{f(x+h) - f(x)}{h} \end{aligned}$$

$$\begin{aligned} D_- f(x) &\equiv \liminf_{h \to 0+} \frac{f(x) - f(x-h)}{h}, \\ D^- f(x) &\equiv \limsup_{h \to 0+} \frac{f(x) - f(x-h)}{h}. \end{aligned}$$

Suppose f is continuous on (a, b) and for all $x \in (a, b)$, $D_+ f(x) \geq 0$. Show that then f is increasing on (a, b). **Hint:** Consider the function, $H(x) \equiv f(x)(d-c) -$

$x(f(d)-f(c))$ where $a<c<d<b$. Thus $H(c)=H(d)$. Also it is easy to see that H cannot be constant if $f(d)<f(c)$ due to the assumption that $D_+f(x)\geq 0$. If there exists $x_1\in(a,b)$ where $H(x_1)>H(c)$, then let $x_0\in(c,d)$ be the point where the maximum of f occurs. Consider $D_+f(x_0)$. If, on the other hand, $H(x)<H(c)$ for all $x\in(c,d)$, then consider $D_+H(c)$.

11. ↑ Suppose in the situation of the above problem we only know

$$D_+f(x)\geq 0 \text{ a.e.}$$

Does the conclusion still follow? What if we only know $D_+f(x)\geq 0$ for every x outside a countable set? **Hint:** In the case of $D_+f(x)\geq 0$,consider the bad function in the exercises for the chapter on the construction of measures which was based on the Cantor set. In the case where $D_+f(x)\geq 0$ for all but countably many x, by replacing $f(x)$ with $\widetilde{f}(x)\equiv f(x)+\varepsilon x$, consider the situation where $D_+\widetilde{f}(x)>0$ for all but countably many x. If in this situation, $\widetilde{f}(c)>\widetilde{f}(d)$ for some $c<d$, and $y\in\left(\widetilde{f}(d),\widetilde{f}(c)\right)$,let

$$z\equiv\sup\left\{x\in[c,d]:\widetilde{f}(x)>y_0\right\}.$$

Show that $\widetilde{f}(z)=y_0$ and $D_+\widetilde{f}(z)\leq 0$. Conclude that if $\widetilde{f}$ fails to be increasing, then $D_+\widetilde{f}(z)\leq 0$ for uncountably many points, z. Now draw a conclusion about f.

12. ↑ Let $f:[a,b]\to\mathbb{R}$ be increasing. Show

$$m\left(\overbrace{\left[D^+f(x)>q>p>D_+f(x)\right]}^{N_{pq}}\right)=0 \tag{9.24}$$

and conclude that aside from a set of measure zero, $D^+f(x)=D_+f(x)$. Similar reasoning will show $D^-f(x)=D_-f(x)$ a.e. and $D^+f(x)=D_-f(x)$ a.e. and so off some set of measure zero, we have

$$D_-f(x)=D^-f(x)=D^+f(x)=D_+f(x)$$

which implies the derivative exists and equals this common value. **Hint**: To show 9.24, let U be an open set containing N_{pq} such that $\overline{m}(N_{pq})+\varepsilon>m(U)$. For each $x\in N_{pq}$ there exist $y>x$ arbitrarily close to x such that

$$f(y)-f(x)<p(y-x).$$

Thus the set of such intervals, $\{[x,y]\}$ which are contained in U constitutes a Vitali cover of N_{pq}. Let $\{[x_i,y_i]\}$ be disjoint and

$$\overline{m}(N_{pq}\setminus\cup_i[x_i,y_i])=0.$$

Now let $V\equiv\cup_i(x_i,y_i)$. Then also we have

$$\overline{m}\left(N_{pq}\setminus\overbrace{\cup_i(x_i,y_i)}^{=V}\right)=0.$$

and so $\overline{m}(N_{pq} \cap V) = \overline{m}(N_{pq})$. For each $x \in N_{pq} \cap V$, there exist $y > x$ arbitrarily close to x such that

$$f(y) - f(x) > q(y - x).$$

Thus the set of such intervals, $\{[x', y']\}$ which are contained in V is a Vitali cover of $N_{pq} \cap V$. Let $\{[x_i', y_i']\}$ be disjoint and

$$\overline{m}(N_{pq} \cap V \setminus \cup_i [x_i', y_i']) = 0.$$

Then verify the following:

$$\begin{aligned} \sum_i f(y_i') - f(x_i') &> q \sum_i (y_i' - x_i') \geq q\overline{m}(N_{pq} \cap V) = q\overline{m}(N_{pq}) \\ &\geq p\overline{m}(N_{pq}) > p(m(U) - \varepsilon) \geq p \sum_i (y_i - x_i) - p\varepsilon \\ &\geq \sum_i (f(y_i) - f(x_i)) - p\varepsilon \geq \sum_i f(y_i') - f(x_i') - p\varepsilon \end{aligned}$$

and therefore, $(q - p)\overline{m}(N_{pq}) \leq p\varepsilon$. Since $\varepsilon > 0$ is arbitrary, this proves that there is a right derivative a.e. A similar argument does the other cases.

13. Suppose f is a function in $L^1(\mathbb{R})$ and f is infinitely differentiable. Does it follow that $f' \in L^1(\mathbb{R})$? **Hint:** What if $\phi \in C_c^\infty(0,1)$ and $f(x) = \phi(2^p(x - p))$ for $x \in (p, p+1)$, $f(x) = 0$ if $x < 0$?

14. For a function $f \in L^1(\mathbb{R}^p)$, the Fourier transform, Ff is given by

$$Ff(\mathbf{t}) \equiv \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^p} e^{-i\mathbf{t}\cdot\mathbf{x}} f(\mathbf{x})\, dx$$

and the so called inverse Fourier transform, $F^{-1}f$ is defined by

$$Ff(\mathbf{t}) \equiv \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^p} e^{i\mathbf{t}\cdot\mathbf{x}} f(\mathbf{x})\, dx$$

Show that if $f \in L^1(\mathbb{R}^p)$, then $\lim_{|\mathbf{x}|\to\infty} Ff(\mathbf{x}) = 0$. **Hint:** You might try to show this first for $f \in C_c^\infty(\mathbb{R}^p)$.

15. Prove Lemma 9.8.1 which says a C^1 function maps a set of measure zero to a set of measure zero using Theorem 9.10.3.

16. For this problem define $\int_a^\infty f(t)\, dt \equiv \lim_{r\to\infty} \int_a^r f(t)\, dt$. Note this coincides with the Lebesgue integral when $f \in L^1(a, \infty)$. Show

    (a) $\int_0^\infty \frac{\sin(u)}{u} du = \frac{\pi}{2}$

    (b) $\lim_{r\to\infty} \int_\delta^\infty \frac{\sin(ru)}{u} du = 0$ whenever $\delta > 0$.

    (c) If $f \in L^1(\mathbb{R})$, then $\lim_{r\to\infty} \int_{\mathbb{R}} \sin(ru) f(u)\, du = 0$.

    Hint: For the first two, use $\frac{1}{u} = \int_0^\infty e^{-ut} dt$ and apply Fubini's theorem to $\int_0^R \sin u \int_{\mathbb{R}} e^{-ut} dt du$. For the last part, first establish it for $f \in C_c^\infty(\mathbb{R})$ and then use the density of this set in $L^1(\mathbb{R})$ to obtain the result. This is called the Riemann Lebesgue lemma.

17. ↑Suppose that $g \in L^1(\mathbb{R})$ and that at some $x > 0$, g is locally Holder continuous from the right and from the left. This means

$$\lim_{r \to 0+} g(x+r) \equiv g(x+)$$

exists,

$$\lim_{r \to 0+} g(x-r) \equiv g(x-)$$

exists and there exist constants $K, \delta > 0$ and $r \in (0,1]$ such that for $|x-y| < \delta$,

$$|g(x+) - g(y)| < K|x-y|^r$$

for $y > x$ and

$$|g(x-) - g(y)| < K|x-y|^r$$

for $y < x$. Show that under these conditions,

$$\lim_{r \to \infty} \frac{2}{\pi} \int_0^\infty \frac{\sin(ur)}{u} \left(\frac{g(x-u) + g(x+u)}{2} \right) du$$

$$= \frac{g(x+) + g(x-)}{2}.$$

18. ↑ Let $g \in L^1(\mathbb{R})$ and suppose g is locally Holder continuous from the right and from the left at x. Show that then

$$\lim_{R \to \infty} \frac{1}{2\pi} \int_{-R}^{R} e^{ixt} \int_{-\infty}^{\infty} e^{-ity} g(y)\, dy dt = \frac{g(x+) + g(x-)}{2}.$$

This is very interesting. This shows $F^{-1}(Fg)(x) = \frac{g(x+)+g(x-)}{2}$, the midpoint of the jump in g at the point, x provided Fg is in L^1. **Hint:** Show the left side of the above equation reduces to

$$\frac{2}{\pi} \int_0^\infty \frac{\sin(ur)}{u} \left(\frac{g(x-u) + g(x+u)}{2} \right) du$$

and then use Problem 17 to obtain the result.

19. ↑ A measurable function g defined on $(0, \infty)$ has exponential growth if $|g(t)| \leq Ce^{\eta t}$ for some η. For $\mathrm{Re}(s) > \eta$, define the Laplace Transform by

$$Lg(s) \equiv \int_0^\infty e^{-su} g(u)\, du.$$

Assume that g has exponential growth as above and is Holder continuous from the right and from the left at t. Pick $\gamma > \eta$. Show that

$$\lim_{R \to \infty} \frac{1}{2\pi} \int_{-R}^{R} e^{\gamma t} e^{iyt} Lg(\gamma + iy)\, dy = \frac{g(t+) + g(t-)}{2}.$$

This formula is sometimes written in the form

$$\frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} e^{st} Lg(s)\, ds$$

and is called the complex inversion integral for Laplace transforms. It can be used to find inverse Laplace transforms. **Hint:**

$$\frac{1}{2\pi} \int_{-R}^{R} e^{\gamma t} e^{iyt} Lg(\gamma + iy)\, dy =$$

$$\frac{1}{2\pi}\int_{-R}^{R} e^{\gamma t}e^{iyt}\int_0^\infty e^{-(\gamma+iy)u}g(u)\,dudy.$$

Now use Fubini's theorem and do the integral from $-R$ to R to get this equal to

$$\frac{e^{\gamma t}}{\pi}\int_{-\infty}^{\infty} e^{-\gamma u}\overline{g}(u)\,\frac{\sin(R(t-u))}{t-u}du$$

where $\overline{g}$ is the zero extension of g off $[0,\infty)$. Then this equals

$$\frac{e^{\gamma t}}{\pi}\int_{-\infty}^{\infty} e^{-\gamma(t-u)}\overline{g}(t-u)\,\frac{\sin(Ru)}{u}du$$

which equals

$$\frac{2e^{\gamma t}}{\pi}\int_0^\infty \frac{\overline{g}(t-u)\,e^{-\gamma(t-u)}+\overline{g}(t+u)\,e^{-\gamma(t+u)}}{2}\,\frac{\sin(Ru)}{u}du$$

and then apply the result of Problem 17.

20. Let K be a nonempty closed and convex subset of $\mathbb{R}^p$. Recall K is convex means that if $\mathbf{x},\mathbf{y}\in K$, then for all $t\in[0,1]$, $t\mathbf{x}+(1-t)\mathbf{y}\in K$. Show that if $\mathbf{x}\in\mathbb{R}^p$ there exists a unique $\mathbf{z}\in K$ such that

$$|\mathbf{x}-\mathbf{z}| = \min\{|\mathbf{x}-\mathbf{y}| : \mathbf{y}\in K\}.$$

This $\mathbf{z}$ will be denoted as $P\mathbf{x}$. **Hint:** First note you do not know K is compact. Establish the parallelogram identity if you have not already done so,

$$|\mathbf{u}-\mathbf{v}|^2+|\mathbf{u}+\mathbf{v}|^2 = 2|\mathbf{u}|^2+2|\mathbf{v}|^2.$$

Then let $\{\mathbf{z}_k\}$ be a minimizing sequence,

$$\lim_{k\to\infty}|\mathbf{z}_k-\mathbf{x}|^2 = \inf\{|\mathbf{x}-\mathbf{y}| : \mathbf{y}\in K\}\equiv\lambda.$$

Now using convexity, explain why

$$\left|\frac{\mathbf{z}_k-\mathbf{z}_m}{2}\right|^2+\left|\mathbf{x}-\frac{\mathbf{z}_k+\mathbf{z}_m}{2}\right|^2 = 2\left|\frac{\mathbf{x}-\mathbf{z}_k}{2}\right|^2+2\left|\frac{\mathbf{x}-\mathbf{z}_m}{2}\right|^2$$

and then use this to argue $\{\mathbf{z}_k\}$ is a Cauchy sequence. Then if $\mathbf{z}_i$ works for $i=1,2$, consider $(\mathbf{z}_1+\mathbf{z}_2)/2$ to get a contradiction.

21. In Problem 20 show that $P\mathbf{x}$ satisfies the following variational inequality.

$$(\mathbf{x}-P\mathbf{x})\cdot(\mathbf{y}-P\mathbf{x})\le 0$$

for all $\mathbf{y}\in K$. Then show that $|P\mathbf{x}_1-P\mathbf{x}_2|\le|\mathbf{x}_1-\mathbf{x}_2|$. **Hint:** For the first part note that if $\mathbf{y}\in K$, the function $t\to|\mathbf{x}-(P\mathbf{x}+t(\mathbf{y}-P\mathbf{x}))|^2$ achieves its minimum on $[0,1]$ at $t=0$. For the second part,

$$(\mathbf{x}_1-P\mathbf{x}_1)\cdot(P\mathbf{x}_2-P\mathbf{x}_1)\le 0,\ (\mathbf{x}_2-P\mathbf{x}_2)\cdot(P\mathbf{x}_1-P\mathbf{x}_2)\le 0.$$

Explain why

$$(\mathbf{x}_2-P\mathbf{x}_2-(\mathbf{x}_1-P\mathbf{x}_1))\cdot(P\mathbf{x}_2-P\mathbf{x}_1)\ge 0$$

and then use a some manipulations and the Cauchy Schwarz inequality to get the desired inequality.

22. Establish the Brouwer fixed point theorem for any convex compact set in $\mathbb{R}^p$. **Hint:** If K is a compact and convex set, let R be large enough that the closed ball, $D(\mathbf{0},R) \supseteq K$. Let P be the projection onto K as in Problem 21 above. If $\mathbf{f}$ is a continuous map from K to K, consider $\mathbf{f}\circ P$. You want to show $\mathbf{f}$ has a fixed point in K.

23. In the situation of the implicit function theorem, suppose $\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0) = \mathbf{0}$ and assume $\mathbf{f}$ is C^1. Show that for $(\mathbf{x},\mathbf{y}) \in B(\mathbf{x}_0,\delta) \times B(\mathbf{y}_0,r)$ where δ, r are small enough, the mapping

$$\mathbf{x} \to T_{\mathbf{y}}(\mathbf{x}) \equiv \mathbf{x} - D_1\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)^{-1}\mathbf{f}(\mathbf{x},\mathbf{y})$$

is continuous and maps $\overline{B(\mathbf{x}_0,\delta)}$ to $\overline{B(\mathbf{x}_0,\delta/2)} \subseteq \overline{B(\mathbf{x}_0,\delta)}$. Apply the Brouwer fixed point theorem to obtain a shorter proof of the implicit function theorem.

24. Here is a really interesting little theorem which depends on the Brouwer fixed point theorem. It plays a prominent role in the treatment of the change of variables formula in Rudin's book, [35] and is useful in other contexts as well. The idea is that if a continuous function mapping a ball in $\mathbb{R}^k$ to $\mathbb{R}^k$ doesn't move any point very much, then the image of the ball must contain a slightly smaller ball.

Lemma: Let $B = B(\mathbf{0},r)$, a ball in $\mathbb{R}^k$ and let $\mathbf{F} : \overline{B} \to \mathbb{R}^k$ be continuous and suppose for some $\varepsilon < 1$,

$$|\mathbf{F}(\mathbf{v}) - \mathbf{v}| < \varepsilon r \tag{9.25}$$

for all $\mathbf{v} \in \overline{B}$. Then

$$\mathbf{F}(B) \supseteq B(\mathbf{0}, r(1-\varepsilon)).$$

Hint: Suppose $\mathbf{a} \in B(\mathbf{0}, r(1-\varepsilon)) \setminus \mathbf{F}(B)$ so it didn't work. First explain why $\mathbf{a} \neq \mathbf{F}(\mathbf{v})$ for all $\mathbf{v} \in \overline{B}$. Now letting $\mathbf{G} : \overline{B} \to \overline{B}$, be defined by $\mathbf{G}(\mathbf{v}) \equiv \frac{r(\mathbf{a}-\mathbf{F}(\mathbf{v}))}{|\mathbf{a}-\mathbf{F}(\mathbf{v})|}$, it follows $\mathbf{G}$ is continuous. Then by the Brouwer fixed point theorem, $\mathbf{G}(\mathbf{v}) = \mathbf{v}$ for some $\mathbf{v} \in \overline{B}$. Explain why $|\mathbf{v}| = r$. Then take the inner product with $\mathbf{v}$ and explain the following steps.

$$\begin{aligned}
(\mathbf{G}(\mathbf{v}),\mathbf{v}) &= |\mathbf{v}|^2 = r^2 = \frac{r}{|\mathbf{a}-\mathbf{F}(\mathbf{v})|}(\mathbf{a}-\mathbf{F}(\mathbf{v}),\mathbf{v}) \\
&= \frac{r}{|\mathbf{a}-\mathbf{F}(\mathbf{v})|}(\mathbf{a}-\mathbf{v}+\mathbf{v}-\mathbf{F}(\mathbf{v}),\mathbf{v}) \\
&= \frac{r}{|\mathbf{a}-\mathbf{F}(\mathbf{v})|}[(\mathbf{a}-\mathbf{v},\mathbf{v}) + (\mathbf{v}-\mathbf{F}(\mathbf{v}),\mathbf{v})] \\
&= \frac{r}{|\mathbf{a}-\mathbf{F}(\mathbf{v})|}\left[(\mathbf{a},\mathbf{v}) - |\mathbf{v}|^2 + (\mathbf{v}-\mathbf{F}(\mathbf{v}),\mathbf{v})\right] \\
&\leq \frac{r}{|\mathbf{a}-\mathbf{F}(\mathbf{v})|}\left[r^2(1-\varepsilon) - r^2 + r^2\varepsilon\right] = 0.
\end{aligned}$$

25. Using Problem 24 establish the following interesting result. Suppose $\mathbf{f} : U \to \mathbb{R}^p$ is differentiable. Let

$$S = \overline{\{\mathbf{x} \in U : \det D\mathbf{f}(\mathbf{x}) = 0\}}.$$

Show $\mathbf{f}(U \setminus S)$ is an open set.

26. Let K be a closed, bounded and convex set in $\mathbb{R}^p$ and let $\mathbf{f} : K \to \mathbb{R}^p$ be continuous and let $\mathbf{y} \in \mathbb{R}^p$. Show using the Brouwer fixed point theorem there exists a point $\mathbf{x} \in K$ such that $P(\mathbf{y} - \mathbf{f}(\mathbf{x}) + \mathbf{x}) = \mathbf{x}$. Next show that $(\mathbf{y} - \mathbf{f}(\mathbf{x}), \mathbf{z} - \mathbf{x}) \leq 0$ for all $\mathbf{z} \in K$. The existence of this $\mathbf{x}$ is known as Browder's lemma and it has

great significance in the study of certain types of nolinear operators. Now suppose $\mathbf{f} : \mathbb{R}^p \to \mathbb{R}^p$ is continuous and satisfies

$$\lim_{|\mathbf{x}|\to\infty} \frac{(\mathbf{f}(\mathbf{x}), \mathbf{x})}{|\mathbf{x}|} = \infty.$$

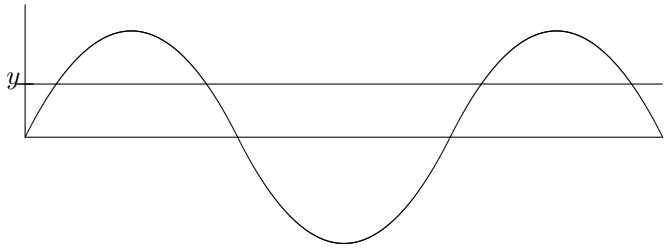
Show using Browder's lemma that $\mathbf{f}$ is onto.

Chapter 10

Degree Theory, An Introduction

This chapter is on the Brouwer degree, a very useful concept with numerous and important applications. The degree can be used to prove some difficult theorems in topology such as the Brouwer fixed point theorem, the Jordan separation theorem, and the invariance of domain theorem. It also is used in bifurcation theory and many other areas in which it is an essential tool. This is an advanced calculus course so the degree will be developed for $\mathbb{R}^n$. When this is understood, it is not too difficult to extend to versions of the degree which hold in Banach space. There is more on degree theory in the book by Deimling [9] and much of the presentation here follows this reference.

To give you an idea what the degree is about, consider a real valued C^1 function defined on an interval, I, and let $y \in f(I)$ be such that $f'(x) \neq 0$ for all $x \in f^{-1}(y)$. In this case the degree is the sum of the signs of $f'(x)$ for $x \in f^{-1}(y)$, written as $d(f, I, y)$.



In the above picture, $d(f, I, y)$ is 0 because there are two places where the sign is 1 and two where it is -1.

The amazing thing about this is the number you obtain in this simple manner is a specialization of something which is defined for continuous functions and which has nothing to do with differentiability.

There are many ways to obtain the Brouwer degree. The method I will use here is due to Heinz [23] and appeared in 1959. It involves first studying the degree for functions in C^2 and establishing all its most important topological properties with the aid of an integral. Then when this is done, it is very easy to extend to general continuous functions.

When you have the topological degree, you can get all sorts of amazing theorems like the invariance of domain theorem and others.

10.1 Preliminary Results

In this chapter Ω will refer to a bounded open set.

Definition 10.1.1 *For Ω a bounded open set, denote by $C\left(\overline{\Omega}\right)$ the set of functions which are continuous on $\overline{\Omega}$ and by $C^m\left(\overline{\Omega}\right), m \leq \infty$ the space of restrictions of functions in $C_c^m\left(\mathbb{R}^n\right)$ to $\overline{\Omega}$. The norm in $C\left(\overline{\Omega}\right)$ is defined as follows.*

$$||f||_\infty = ||f||_{C\left(\overline{\Omega}\right)} \equiv \sup\left\{|f(\mathbf{x})| : \mathbf{x} \in \overline{\Omega}\right\}.$$

If the functions take values in $\mathbb{R}^n$ write $C^m\left(\overline{\Omega};\mathbb{R}^n\right)$ or $C\left(\overline{\Omega};\mathbb{R}^n\right)$ for these functions if there is no differentiability assumed. The norm on $C\left(\overline{\Omega};\mathbb{R}^n\right)$ is defined in the same way as above,

$$||\mathbf{f}||_\infty = ||\mathbf{f}||_{C\left(\overline{\Omega};\mathbb{R}^n\right)} \equiv \sup\left\{|\mathbf{f}(\mathbf{x})| : \mathbf{x} \in \overline{\Omega}\right\}.$$

Also, $C\left(\Omega;\mathbb{R}^n\right)$ consists of functions which are continuous on Ω that have values in $\mathbb{R}^n$ and $C^m\left(\Omega;\mathbb{R}^n\right)$ denotes the functions which have m continuous derivatives defined on Ω.

Theorem 10.1.2 *Let Ω be a bounded open set in $\mathbb{R}^n$ and let $f \in C\left(\overline{\Omega}\right)$. Then there exists $g \in C^\infty\left(\overline{\Omega}\right)$ with $||g-f||_{C\left(\overline{\Omega}\right)} < \varepsilon$. In fact, g can be assumed to equal a polynomial for all $\mathbf{x} \in \Omega$.*

Proof: This follows immediately from the Weierstrass approximation theorem. Pick a polynomial, p such that $||p-f||_{C\left(\overline{\Omega}\right)} < \varepsilon$. Now $p \notin C^\infty\left(\overline{\Omega}\right)$ because it does not vanish outside some compact subset of $\mathbb{R}^n$ so let g equal p multiplied by some function $\psi \in C_c^\infty\left(\mathbb{R}^n\right)$ where $\psi = 1$ on $\overline{\Omega}$. See Theorem 9.5.8. ■

Applying this result to the components of a vector valued function yields the following corollary.

Corollary 10.1.3 *If $\mathbf{f} \in C\left(\overline{\Omega};\mathbb{R}^n\right)$ for Ω a bounded subset of $\mathbb{R}^n$, then for all $\varepsilon > 0$, there exists $\mathbf{g} \in C^\infty\left(\overline{\Omega};\mathbb{R}^n\right)$ such that*

$$||\mathbf{g}-\mathbf{f}||_\infty < \varepsilon.$$

Lemma 9.12.1 on Page 253 will also play an important role in the definition of the Brouwer degree. Earlier it made possible an easy proof of the Brouwer fixed point theorem. Later in this chapter, it is used to show the definition of the degree is well defined. For convenience, here it is stated again.

Lemma 10.1.4 *Let $\mathbf{g} : U \to \mathbb{R}^n$ be C^2 where U is an open subset of $\mathbb{R}^n$. Then*

$$\sum_{j=1}^n \operatorname{cof}(D\mathbf{g})_{ij,j} = 0,$$

where here $(D\mathbf{g})_{ij} \equiv g_{i,j} \equiv \frac{\partial g_i}{\partial x_j}$. Also, $\operatorname{cof}(D\mathbf{g})_{ij} = \frac{\partial \det(D\mathbf{g})}{\partial g_{i,j}}$.

Another simple result which will be used whenever convenient is the following lemma, stated in somewhat more generality than needed.

Lemma 10.1.5 *Let K be a compact set and C a closed set in a complete normed vector space such that $K \cap C = \emptyset$. Then*

$$\operatorname{dist}(K,C) > 0.$$

Proof: Let

$$d \equiv \inf \{||k-c|| : k \in K, c \in C\}$$

Let $\{k_n\}, \{c_n\}$ be such that

$$d + \frac{1}{n} > ||k_n - c_n||.$$

Since K is compact, there is a subsequence still denoted by $\{k_n\}$ such that $k_n \to k \in K$. Then also

$$||c_n - c_m|| \leq ||c_n - k_n|| + ||k_n - k_m|| + ||c_m - k_m||$$

If $d = 0$, then as $m, n \to \infty$ it follows $||c_n - c_m|| \to 0$ and so $\{c_n\}$ is a Cauchy sequence which must converge to some $c \in C$. But then $||c-k|| = \lim_{n\to\infty} ||c_n - k_n|| = 0$ and so $c = k \in C \cap K$, a contradiction to these sets being disjoint. ■

In particular the distance between a point and a closed set is always positive if the point is not in the closed set. Of course this is obvious even without the above lemma.

10.2 Definitions And Elementary Properties

In this section, $\mathbf{f} : \overline{\Omega} \to \mathbb{R}^n$ will be a continuous map. It is always assumed that $\mathbf{f}(\partial\Omega)$ misses the point $\mathbf{y}$ where $d(\mathbf{f}, \Omega, \mathbf{y})$ is the topological degree which is being defined. Also, it is assumed Ω is a bounded open set.

Definition 10.2.1 $\mathcal{U}_{\mathbf{y}} \equiv \left\{\mathbf{f} \in C\left(\overline{\Omega}; \mathbb{R}^n\right) : \mathbf{y} \notin \mathbf{f}(\partial\Omega)\right\}$. *(Recall that* $\partial\Omega = \overline{\Omega} \setminus \Omega$*) For two functions,*

$$\mathbf{f}, \mathbf{g} \in \mathcal{U}_{\mathbf{y}},$$

$\mathbf{f} \sim \mathbf{g}$ *if there exists a continuous function,*

$$\mathbf{h} : \overline{\Omega} \times [0,1] \to \mathbb{R}^n$$

such that $\mathbf{h}(\mathbf{x},1) = \mathbf{g}(\mathbf{x})$ *and* $\mathbf{h}(\mathbf{x},0) = \mathbf{f}(\mathbf{x})$ *and* $\mathbf{x} \to \mathbf{h}(\mathbf{x},t) \in \mathcal{U}_{\mathbf{y}}$ *for all* $t \in [0,1]$ $(\mathbf{y} \notin \mathbf{h}(\partial\Omega, t))$. *This function,* $\mathbf{h}$*, is called a homotopy and* $\mathbf{f}$ *and* $\mathbf{g}$ *are homotopic.*

Definition 10.2.2 *For* W *an open set in* $\mathbb{R}^n$ *and* $\mathbf{g} \in C^1(W; \mathbb{R}^n)$ $\mathbf{y}$ *is called a regular value of* $\mathbf{g}$ *if whenever* $\mathbf{x} \in \mathbf{g}^{-1}(\mathbf{y})$, $\det(D\mathbf{g}(\mathbf{x})) \neq 0$. *Note that if* $\mathbf{g}^{-1}(\mathbf{y}) = \emptyset$, *it follows that* $\mathbf{y}$ *is a regular value from this definition. Denote by* $S_{\mathbf{g}}$ *the set of singular values of* $\mathbf{g}$*, those* $\mathbf{y}$ *such that* $\det(D\mathbf{g}(\mathbf{x})) = 0$ *for some* $\mathbf{x} \in \mathbf{g}^{-1}(\mathbf{y})$.

Lemma 10.2.3 *The relation* $\sim$ *is an equivalence relation and, denoting by* $[\mathbf{f}]$ *the equivalence class determined by* $\mathbf{f}$*, it follows that* $[\mathbf{f}]$ *is an open subset of*

$$\mathcal{U}_{\mathbf{y}} \equiv \left\{\mathbf{f} \in C\left(\overline{\Omega}; \mathbb{R}^n\right) : \mathbf{y} \notin \mathbf{f}(\partial\Omega)\right\}.$$

Furthermore, $\mathcal{U}_{\mathbf{y}}$ *is an open set in* $C\left(\overline{\Omega}; \mathbb{R}^n\right)$ *and if* $\mathbf{f} \in \mathcal{U}_{\mathbf{y}}$ *and* $\varepsilon > 0$*, there exists* $\mathbf{g} \in [\mathbf{f}] \cap C^2\left(\overline{\Omega}; \mathbb{R}^n\right)$ *for which* $\mathbf{y}$ *is a regular value of* $\mathbf{g}$ *and* $||\mathbf{f} - \mathbf{g}|| < \varepsilon$.

Proof: In showing that $\sim$ is an equivalence relation, it is easy to verify that $\mathbf{f} \sim \mathbf{f}$ and that if $\mathbf{f} \sim \mathbf{g}$, then $\mathbf{g} \sim \mathbf{f}$. To verify the transitive property for an equivalence relation, suppose $\mathbf{f} \sim \mathbf{g}$ and $\mathbf{g} \sim \mathbf{k}$, with the homotopy for $\mathbf{f}$ and $\mathbf{g}$, the function, $\mathbf{h}_1$ and the homotopy for $\mathbf{g}$ and $\mathbf{k}$, the function $\mathbf{h}_2$. Thus $\mathbf{h}_1(\mathbf{x},0) = \mathbf{f}(\mathbf{x})$, $\mathbf{h}_1(\mathbf{x},1) = \mathbf{g}(\mathbf{x})$ and $\mathbf{h}_2(\mathbf{x},0) = \mathbf{g}(\mathbf{x})$, $\mathbf{h}_2(\mathbf{x},1) = \mathbf{k}(\mathbf{x})$. Then define a homotopy of $\mathbf{f}$ and $\mathbf{k}$ as follows.

$$\mathbf{h}(\mathbf{x},t) \equiv \begin{cases} \mathbf{h}_1(\mathbf{x},2t) & \text{if } t \in \left[0, \frac{1}{2}\right] \\ \mathbf{h}_2(\mathbf{x},2t-1) & \text{if } t \in \left[\frac{1}{2}, 1\right] \end{cases}.$$

It is obvious that $\mathcal{U}_{\mathbf{y}}$ is an open subset of $C\left(\overline{\Omega};\mathbb{R}^n\right)$. If $\mathbf{g}\in\mathcal{U}_y$ then $\mathbf{y}\notin\mathbf{g}\left(\partial\Omega\right)$ a compact set. Hence if $\mathbf{f}$ is close enough to $\mathbf{g}$, the same is true of $\mathbf{f}$.

Next consider the claim that $[\mathbf{f}]$ is also an open set. If $\mathbf{f}\in\mathcal{U}_{\mathbf{y}}$, There exists $\delta>0$ such that $B\left(\mathbf{y},2\delta\right)\cap\mathbf{f}\left(\partial\Omega\right)=\emptyset$. Let $\mathbf{f}_1\in C\left(\overline{\Omega};\mathbb{R}^n\right)$ with $\left\|\mathbf{f}_1-\mathbf{f}\right\|_\infty<\delta$. Then if $t\in[0,1]$, and $\mathbf{x}\in\partial\Omega$

$$\left|\mathbf{f}\left(\mathbf{x}\right)+t\left(\mathbf{f}_1\left(\mathbf{x}\right)-\mathbf{f}\left(\mathbf{x}\right)\right)-\mathbf{y}\right|\geq\left|\mathbf{f}\left(\mathbf{x}\right)-\mathbf{y}\right|-t\left\|\mathbf{f}-\mathbf{f}_1\right\|_\infty>2\delta-t\delta>0.$$

Therefore, $B\left(\mathbf{f},\delta\right)\subseteq[\mathbf{f}]$ because if $\mathbf{f}_1\in B\left(\mathbf{f},\delta\right)$, this shows that, letting $\mathbf{h}\left(\mathbf{x},t\right)\equiv\mathbf{f}\left(\mathbf{x}\right)+t\left(\mathbf{f}_1\left(\mathbf{x}\right)-\mathbf{f}\left(\mathbf{x}\right)\right)$, $\mathbf{f}_1\sim\mathbf{f}$.

It remains to verify the last assertion of the lemma. Since $[\mathbf{f}]$ is an open set, it follows from Theorem 10.1.2 there exists $\mathbf{g}\in[\mathbf{f}]\cap C^2\left(\overline{\Omega};\mathbb{R}^n\right)$ and $\left\|\mathbf{g}-\mathbf{f}\right\|_\infty<\varepsilon/2$. If $\mathbf{y}$ is a regular value of $\mathbf{g}$, leave $\mathbf{g}$ unchanged. The desired function has been found. In the other case, let δ be small enough that $B\left(\mathbf{y},2\delta\right)\cap\mathbf{g}\left(\partial\Omega\right)=\emptyset$. Next let

$$S\equiv\left\{\mathbf{x}\in\overline{\Omega}:\det D\mathbf{g}\left(\mathbf{x}\right)=0\right\}$$

By Sard's lemma, Lemma 9.9.9 on Page 247, $\mathbf{g}\left(S\right)$ is a set of measure zero and so in particular contains no open ball and so there exist regular values of $\mathbf{g}$ arbitrarily close to $\mathbf{y}$. Let $\widetilde{\mathbf{y}}$ be one of these regular values, $|\mathbf{y}-\widetilde{\mathbf{y}}|<\varepsilon/2$, and consider

$$\mathbf{g}_1\left(\mathbf{x}\right)\equiv\mathbf{g}\left(\mathbf{x}\right)+\mathbf{y}-\widetilde{\mathbf{y}}.$$

It follows $\mathbf{g}_1\left(\mathbf{x}\right)=\mathbf{y}$ if and only if $\mathbf{g}\left(\mathbf{x}\right)=\widetilde{\mathbf{y}}$ and so, since $D\mathbf{g}\left(\mathbf{x}\right)=D\mathbf{g}_1\left(\mathbf{x}\right)$, $\mathbf{y}$ is a regular value of $\mathbf{g}_1$. Then for $t\in[0,1]$ and $\mathbf{x}\in\partial\Omega$,

$$\left|\mathbf{g}\left(\mathbf{x}\right)+t\left(\mathbf{g}_1\left(\mathbf{x}\right)-\mathbf{g}\left(\mathbf{x}\right)\right)-\mathbf{y}\right|\geq\left|\mathbf{g}\left(\mathbf{x}\right)-\mathbf{y}\right|-t\left|\mathbf{y}-\widetilde{\mathbf{y}}\right|>2\delta-t\delta\geq\delta>0.$$

provided $|\mathbf{y}-\widetilde{\mathbf{y}}|$ is small enough. It follows $\mathbf{g}_1\sim\mathbf{g}$ and so $\mathbf{g}_1\sim\mathbf{f}$. Also provided $|\mathbf{y}-\widetilde{\mathbf{y}}|$ is small enough,

$$\begin{aligned}\left\|\mathbf{f}-\mathbf{g}_1\right\| &\leq \left\|\mathbf{f}-\mathbf{g}\right\|+\left\|\mathbf{g}-\mathbf{g}_1\right\| \\ &< \varepsilon/2+\varepsilon/2=\varepsilon.\ \blacksquare\end{aligned}$$

The main conclusion of this lemma is that for $\mathbf{f}\in\mathcal{U}_{\mathbf{y}}$, there always exists a function $\mathbf{g}$ of $C^2\left(\overline{\Omega};\mathbb{R}^n\right)$ which is uniformly close to $\mathbf{f}$, homotopic to $\mathbf{f}$ and also such that $\mathbf{y}$ is a regular value of $\mathbf{g}$.

10.2.1 The Degree For $C^2\left(\overline{\Omega};\mathbb{R}^n\right)$

Here I will give a definition of the degree which works for all functions in $C^2\left(\overline{\Omega};\mathbb{R}^n\right)$.

Definition 10.2.4 *Let $\mathbf{g}\in C^2\left(\overline{\Omega};\mathbb{R}^n\right)\cap\mathcal{U}_{\mathbf{y}}$ where Ω is a bounded open set. Also let ϕ_ε be a mollifier.*

$$\phi_\varepsilon\in C_c^\infty\left(B\left(\mathbf{0},\varepsilon\right)\right),\phi_\varepsilon\geq0,\ \int\phi_\varepsilon dx=1.$$

Then

$$d\left(\mathbf{g},\Omega,\mathbf{y}\right)\equiv\lim_{\varepsilon\to0}\int_\Omega\phi_\varepsilon\left(\mathbf{g}\left(\mathbf{x}\right)-\mathbf{y}\right)\det D\mathbf{g}\left(\mathbf{x}\right)dx$$

It is necessary to show that this limit exists.

Lemma 10.2.5 *The above definition is well defined. In particular the limit exists. In fact*

$$\int_\Omega \phi_\varepsilon (\mathbf{g}(\mathbf{x}) - \mathbf{y}) \det D\mathbf{g}(\mathbf{x})\, dx$$

does not depend on ε whenenver ε is small enough. If $\mathbf{y}$ is a regular value for $\mathbf{g}$ then for all ε small enough,

$$\int_\Omega \phi_\varepsilon (\mathbf{g}(\mathbf{x}) - \mathbf{y}) \det D\mathbf{g}(\mathbf{x})\, dx \equiv$$

$$\sum \left\{ \operatorname{sgn}(\det D\mathbf{g}(\mathbf{x})) : \mathbf{x} \in \mathbf{g}^{-1}(\mathbf{y}) \right\} \tag{10.1}$$

If $\mathbf{f}, \mathbf{g}$ are two functions in $C^2\left(\overline{\Omega}; \mathbb{R}^n\right)$ such that for all $t \in [0,1]$,

$$\mathbf{y} \notin (t\mathbf{f} + (1-t)\mathbf{g})(\partial\Omega) \tag{10.2}$$

then for each $\varepsilon > 0$,

$$\begin{aligned} & \int_\Omega \phi_\varepsilon (\mathbf{f}(\mathbf{x}) - \mathbf{y}) \det D\mathbf{f}(\mathbf{x})\, dx \\ = & \int_\Omega \phi_\varepsilon (\mathbf{g}(\mathbf{x}) - \mathbf{y}) \det D\mathbf{g}(\mathbf{x})\, dx \end{aligned} \tag{10.3}$$

If $\mathbf{g}, \mathbf{f} \in \mathcal{U}_\mathbf{y} \cap C^2\left(\overline{\Omega}; \mathbb{R}^n\right)$, and 10.2 holds, then

$$d(\mathbf{f}, \Omega, \mathbf{y}) = d(\mathbf{g}, \Omega, \mathbf{y})$$

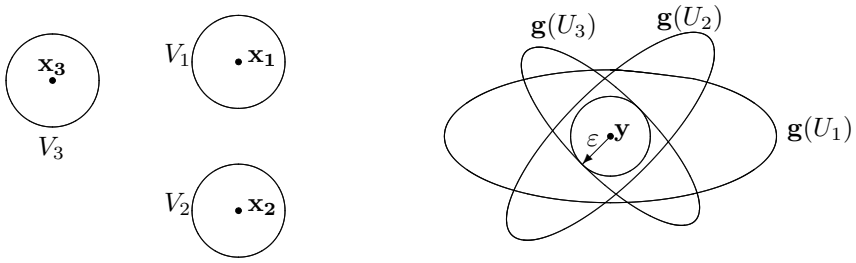
If $\operatorname{dist}(\mathbf{y}, \mathbf{g}(\partial\Omega)) > 5\delta$ and $\mathbf{y}_1 \in B(\mathbf{y}, \delta)$, then $d(\mathbf{g}, \Omega, \mathbf{y}) = d(\mathbf{g}, \Omega, \mathbf{y}_1)$. Also, the appropriate integrals are equal.

Proof: If $\mathbf{y}$ is not a value of $\mathbf{g}$ then there is not much to show. For small enough ε, you will get 0 in the integral.

The case where y is a regular value

First consider the case where $\mathbf{y}$ is a regular value of $\mathbf{g}$. I will show that in this case, the integral expression is eventually constant for small $\varepsilon > 0$ and equals the right side of 10.1. I claim the right side of this equation is actually a finite sum. This follows from the inverse function theorem because $\mathbf{g}^{-1}(\mathbf{y})$ is a closed, hence compact subset of Ω due to the assumption that $\mathbf{y} \notin \mathbf{g}(\partial\Omega)$. If $\mathbf{g}^{-1}(\mathbf{y})$ had infinitely many points in it, there would exist a sequence of distinct points $\{\mathbf{x}_k\} \subseteq \mathbf{g}^{-1}(\mathbf{y})$. Since Ω is bounded, some subsequence $\{\mathbf{x}_{k_l}\}$ would converge to a limit point $\mathbf{x}_\infty$. By continuity of $\mathbf{g}$, it follows $\mathbf{x}_\infty \in \mathbf{g}^{-1}(\mathbf{y})$ also and so $\mathbf{x}_\infty \in \Omega$. Therefore, since $\mathbf{y}$ is a regular value, there is an open set, $U_{\mathbf{x}_\infty}$, containing $\mathbf{x}_\infty$ such that $\mathbf{g}$ is one to one on this open set contradicting the assertion that $\lim_{l\to\infty} \mathbf{x}_{k_l} = \mathbf{x}_\infty$. Therefore, this set is finite and so the sum is well defined.

Thus the right side of 10.1 is finite when $\mathbf{y}$ is a regular value. Next I need to show the left side of this equation is eventually constant and equals the right side. By what was just shown, there are finitely many points, $\{\mathbf{x}_i\}_{i=1}^m = \mathbf{g}^{-1}(\mathbf{y})$. By the inverse function theorem, there exist disjoint open sets U_i with $\mathbf{x}_i \in U_i$, such that $\mathbf{g}$ is one to one on U_i with $\det(D\mathbf{g}(\mathbf{x}))$ having constant sign on U_i and $\mathbf{g}(U_i)$ is an open set containing $\mathbf{y}$. Then let ε be small enough that $B(\mathbf{y}, \varepsilon) \subseteq \cap_{i=1}^m \mathbf{g}(U_i)$ and let $V_i \equiv \mathbf{g}^{-1}(B(\mathbf{y}, \varepsilon)) \cap U_i$.



Therefore, for any ε this small,

$$\int_\Omega \phi_\varepsilon(\mathbf{g}(\mathbf{x})-\mathbf{y})\det D\mathbf{g}(\mathbf{x})\,dx = \sum_{i=1}^m \int_{V_i} \phi_\varepsilon(\mathbf{g}(\mathbf{x})-\mathbf{y})\det D\mathbf{g}(\mathbf{x})\,dx$$

The reason for this is as follows. The integrand on the left is nonzero only if $\mathbf{g}(\mathbf{x})-\mathbf{y} \in B(\mathbf{0},\varepsilon)$ which occurs only if $\mathbf{g}(\mathbf{x}) \in B(\mathbf{y},\varepsilon)$ which is the same as $\mathbf{x} \in \mathbf{g}^{-1}(B(\mathbf{y},\varepsilon))$. Therefore, the integrand is nonzero only if $\mathbf{x}$ is contained in exactly one of the disjoint sets, V_i. Now using the change of variables theorem, $(\mathbf{z}=\mathbf{g}(\mathbf{x})-\mathbf{y}, \mathbf{g}^{-1}(\mathbf{y}+\mathbf{z})=\mathbf{x}.)$

$$= \sum_{i=1}^m \int_{\mathbf{g}(V_i)-\mathbf{y}} \phi_\varepsilon(\mathbf{z}) \det D\mathbf{g}\left(\mathbf{g}^{-1}(\mathbf{y}+\mathbf{z})\right) \left|\det D\mathbf{g}^{-1}(\mathbf{y}+\mathbf{z})\right| dz$$

By the chain rule, $I = D\mathbf{g}\left(\mathbf{g}^{-1}(\mathbf{y}+\mathbf{z})\right) D\mathbf{g}^{-1}(\mathbf{y}+\mathbf{z})$ and so

$$\begin{aligned} & \det D\mathbf{g}\left(\mathbf{g}^{-1}(\mathbf{y}+\mathbf{z})\right) \left|\det D\mathbf{g}^{-1}(\mathbf{y}+\mathbf{z})\right| \\ = \quad & \operatorname{sgn}\left(\det D\mathbf{g}\left(\mathbf{g}^{-1}(\mathbf{y}+\mathbf{z})\right)\right) \cdot \end{aligned}$$

$$\begin{aligned} & \left|\det D\mathbf{g}\left(\mathbf{g}^{-1}(\mathbf{y}+\mathbf{z})\right)\right| \left|\det D\mathbf{g}^{-1}(\mathbf{y}+\mathbf{z})\right| \\ = \quad & \operatorname{sgn}\left(\det D\mathbf{g}\left(\mathbf{g}^{-1}(\mathbf{y}+\mathbf{z})\right)\right) \end{aligned}$$

$$= \operatorname{sgn}(\det D\mathbf{g}(\mathbf{x})) = \operatorname{sgn}(\det D\mathbf{g}(\mathbf{x}_i)).$$

Therefore, this reduces to

$$\sum_{i=1}^m \operatorname{sgn}(\det D\mathbf{g}(\mathbf{x}_i)) \int_{\mathbf{g}(V_i)-\mathbf{y}} \phi_\varepsilon(\mathbf{z})\,dz =$$

$$\sum_{i=1}^m \operatorname{sgn}(\det D\mathbf{g}(\mathbf{x}_i)) \int_{B(\mathbf{0},\varepsilon)} \phi_\varepsilon(\mathbf{z})\,dz = \sum_{i=1}^m \operatorname{sgn}(\det D\mathbf{g}(\mathbf{x}_i)).$$

In case $\mathbf{g}^{-1}(\mathbf{y}) = \emptyset$, there exists $\varepsilon > 0$ such that $\mathbf{g}\left(\overline{\Omega}\right) \cap B(\mathbf{y},\varepsilon) = \emptyset$ and so for ε this small,

$$\int_\Omega \phi_\varepsilon(\mathbf{g}(\mathbf{x})-\mathbf{y})\det D\mathbf{g}(\mathbf{x})\,dx = 0.$$

Showing the integral is constant for small ε

With this done it is necessary to show that the integral in the definition of the degree is constant for small enough ε even if $\mathbf{y}$ is not a regular value. To do this, I will first show that if 10.2 holds, then 10.3 holds. This particular part of the argument is the

trick which makes surprising things happen. This is where the fact the functions are twice continuously differentiable is used. Suppose then that $\mathbf{f}, \mathbf{g}$ satisfy 10.2. Also let $\varepsilon > 0$ be such that for all $t \in [0,1]$,

$$B(\mathbf{y}, \varepsilon) \cap (\mathbf{f} + t(\mathbf{g} - \mathbf{f}))(\partial\Omega) = \emptyset \tag{10.4}$$

Define for $t \in [0,1]$,

$$H(t) \equiv \int_\Omega \phi_\varepsilon(\mathbf{f} - \mathbf{y} + t(\mathbf{g} - \mathbf{f})) \det(D(\mathbf{f} + t(\mathbf{g} - \mathbf{f}))) \, dx.$$

Then if $t \in (0,1)$,

$$\begin{aligned} H'(t) = \int_\Omega \sum_\alpha \phi_{\varepsilon,\alpha}(\mathbf{f}(\mathbf{x}) - \mathbf{y} + t(\mathbf{g}(\mathbf{x}) - \mathbf{f}(\mathbf{x}))) \cdot \\ (g_\alpha(\mathbf{x}) - f_\alpha(\mathbf{x})) \det D(\mathbf{f} + t(\mathbf{g} - \mathbf{f})) \, dx \\ + \int_\Omega \phi_\varepsilon(\mathbf{f} - \mathbf{y} + t(\mathbf{g} - \mathbf{f})) \cdot \\ \sum_{\alpha,j} \det D(\mathbf{f} + t(\mathbf{g} - \mathbf{f}))_{,\alpha j} (g_\alpha - f_\alpha)_{,j} \, dx \equiv \mathbf{A} + \mathbf{B}. \end{aligned}$$

In this formula, the function det is considered as a function of the n^2 entries in the $n \times n$ matrix and the $,\alpha j$ represents the derivative with respect to the αj^{th} entry. Now as in the proof of Lemma 9.12.1 on Page 253,

$$\det D(\mathbf{f} + t(\mathbf{g} - \mathbf{f}))_{,\alpha j} = (\operatorname{cof} D(\mathbf{f} + t(\mathbf{g} - \mathbf{f})))_{\alpha j}$$

and so

$$\begin{aligned} \mathbf{B} = \int_\Omega \sum_\alpha \sum_j \phi_\varepsilon(\mathbf{f} - \mathbf{y} + t(\mathbf{g} - \mathbf{f})) \cdot \\ (\operatorname{cof} \ D(\mathbf{f} + t(\mathbf{g} - \mathbf{f})))_{\alpha j} (g_\alpha - f_\alpha)_{,j} \, dx. \end{aligned}$$

By hypothesis

$$\begin{aligned} \mathbf{x} \ \to \ & \phi_\varepsilon(\mathbf{f}(\mathbf{x}) - \mathbf{y} + t(\mathbf{g}(\mathbf{x}) - \mathbf{f}(\mathbf{x}))) \cdot \\ & (\operatorname{cof} \ D(\mathbf{f}(\mathbf{x}) + t(\mathbf{g}(\mathbf{x}) - \mathbf{f}(\mathbf{x}))))_{\alpha j} \end{aligned}$$

is in $C_c^1(\Omega)$ because if $\mathbf{x} \in \partial\Omega$, it follows by 10.4 that for all $t \in [0,1]$

$$\mathbf{f}(\mathbf{x}) - \mathbf{y} + t(\mathbf{g}(\mathbf{x}) - \mathbf{f}(\mathbf{x})) \notin B(\mathbf{0}, \varepsilon)$$

and so $\phi_\varepsilon(\mathbf{f}(\mathbf{x}) - \mathbf{y} + t(\mathbf{g}(\mathbf{x}) - \mathbf{f}(\mathbf{x}))) = 0$. Thus it equals 0 on $\partial\Omega$. Therefore, integrate by parts and write

$$\begin{aligned} \mathbf{B} = -\int_\Omega \sum_\alpha \sum_j \frac{\partial}{\partial x_j} (\phi_\varepsilon(\mathbf{f} - \mathbf{y} + t(\mathbf{g} - \mathbf{f}))) \cdot \\ (\operatorname{cof} \ D(\mathbf{f} + t(\mathbf{g} - \mathbf{f})))_{\alpha j} (g_\alpha - f_\alpha) \, dx + \\ - \int_\Omega \sum_\alpha \sum_j \phi_\varepsilon(\mathbf{f} - \mathbf{y} + t(\mathbf{g} - \mathbf{f})) \\ \cdot (\operatorname{cof} D(\mathbf{f} + t(\mathbf{g} - \mathbf{f})))_{\alpha j,j} (g_\alpha - f_\alpha) \, dx. \end{aligned}$$

The second term equals zero by Lemma 10.1.4. Simplifying the first term yields

$$\mathbf{B} = -\int_\Omega \sum_\alpha \sum_j \sum_\beta \phi_{\varepsilon,\beta}\left(\mathbf{f}-\mathbf{y}+t\left(\mathbf{g}-\mathbf{f}\right)\right)\cdot$$

$$\left(f_{\beta,j}+t\left(g_{\beta,j}-f_{\beta,j}\right)\right)\left(\operatorname{cof} D\left(\mathbf{f}+t\left(\mathbf{g}-\mathbf{f}\right)\right)\right)_{\alpha j}\left(g_\alpha - f_\alpha\right)dx$$

Now the sum on j is the dot product of the β^{th} row with the α^{th} row of the cofactor matrix which equals zero unless $\beta=\alpha$ because it would be a cofactor expansion of a matrix with two equal rows. When $\beta=\alpha$, the sum on j reduces to $\det\left(D\left(\mathbf{f}+t\left(\mathbf{g}-\mathbf{f}\right)\right)\right)$. Thus it reduces to

$$\begin{aligned} &= -\int_\Omega \sum_\alpha \sum_\beta \phi_{\varepsilon,\beta}\left(\mathbf{f}-\mathbf{y}+t\left(\mathbf{g}-\mathbf{f}\right)\right)\delta_{\beta\alpha}\cdot \\ &\quad \det\left(D\left(\mathbf{f}+t\left(\mathbf{g}-\mathbf{f}\right)\right)\right)\left(g_\alpha-f_\alpha\right)dx \\ &= -\int_\Omega \sum_\alpha \phi_{\varepsilon,\alpha}\left(\mathbf{f}-\mathbf{y}+t\left(\mathbf{g}-\mathbf{f}\right)\right) \\ &\quad \cdot\det\left(D\left(\mathbf{f}+t\left(\mathbf{g}-\mathbf{f}\right)\right)\right)\left(g_\alpha-f_\alpha\right)dx = -\mathbf{A}. \end{aligned}$$

Therefore, $H'(t)=0$ and so H is a constant.

Now let $\mathbf{g}\in\mathcal{U}_{\mathbf{y}}\cap C^2\left(\overline{\Omega};\mathbb{R}^n\right)$. Say $\operatorname{dist}\left(\mathbf{y},\mathbf{g}\left(\partial\Omega\right)\right)\geq 5\delta$ where $\mathbf{y}$ is a value, maybe not regular. By Sard's lemma, Lemma 9.9.9 there exists a regular value $\mathbf{y}_1$ of $\mathbf{g}$ in $B\left(\mathbf{y},\delta\right)$. Thus $\operatorname{dist}\left(\mathbf{y}_1,\mathbf{g}\left(\partial\Omega\right)\right)\geq 4\delta$. This is because, by this lemma, the set of points which are not regular values has measure zero so this set of points must have empty interior.

$$\mathbf{g}_1\left(\mathbf{x}\right)\equiv\mathbf{g}\left(\mathbf{x}\right)+\mathbf{y}-\mathbf{y}_1$$

Suppose for some $\mathbf{x}\in\partial\Omega$,

$$\left(\left(1-t\right)\mathbf{g}_1+t\mathbf{g}\right)\left(\mathbf{x}\right)=\left(1-t\right)\mathbf{g}\left(\mathbf{x}\right)+\left(1-t\right)\left(\mathbf{y}-\mathbf{y}_1\right)+t\mathbf{g}\left(\mathbf{x}\right)=\mathbf{y}$$

Then

$$\mathbf{g}\left(\mathbf{x}\right)+\left(1-t\right)\left(\mathbf{y}-\mathbf{y}_1\right)=\mathbf{y}$$

which cannot occur because $\left|\left(1-t\right)\left(\mathbf{y}-\mathbf{y}_1\right)\right|<\delta$ and $\mathbf{y}$ is at least 5δ away from all points of $\mathbf{g}\left(\partial\Omega\right)$. Thus

$$\mathbf{y}\notin\left(\left(1-t\right)\mathbf{g}_1+t\mathbf{g}\right)\left(\partial\Omega\right)\equiv\left(\mathbf{g}_1+t\left(\mathbf{g}-\mathbf{g}_1\right)\right)\left(\partial\Omega\right)\text{ for all }t\in[0,1]$$

whenever $\mathbf{y}_1$ is this close to $\mathbf{y}$. Then $\mathbf{g}_1\left(\mathbf{x}\right)=\mathbf{y}$ if and only if $\mathbf{g}\left(\mathbf{x}\right)=\mathbf{y}_1$ which is a regular value. Note also $D\left(\mathbf{g}\left(\mathbf{x}\right)\right)=D\left(\mathbf{g}_1\left(\mathbf{x}\right)\right)$. Then from what was just shown, letting $\mathbf{f}=\mathbf{g}$ and $\mathbf{g}=\mathbf{g}_1$ in the above and using $\mathbf{g}-\mathbf{y}_1=\mathbf{g}_1-\mathbf{y}$, for ε small enough that $B\left(\mathbf{y},\varepsilon\right)$ has empty intersection with $\left(\mathbf{g}+t\left(\mathbf{g}_1-\mathbf{g}\right)\right)\left(\partial\Omega\right)$ for all $t\in[0,1]$,

$$\begin{aligned} H\left(t\right) &\equiv \int_\Omega \phi_\varepsilon\left(\mathbf{g}-\mathbf{y}+t\left(\mathbf{g}_1-\mathbf{g}\right)\right)\det\left(D\left(\mathbf{g}+t\left(\mathbf{g}_1-\mathbf{g}\right)\right)\right)dx \\ &= \int_\Omega \phi_\varepsilon\left(\mathbf{g}-\mathbf{y}+t\left(\mathbf{y}-\mathbf{y}_1\right)\right)\det\left(D\left(\mathbf{g}+t\left(\mathbf{y}-\mathbf{y}_1\right)\right)\right)dx \end{aligned}$$

is constant for $t\in[0,1]$. Hence,

$$\begin{aligned} &\int_\Omega \phi_\varepsilon\left(\mathbf{g}\left(\mathbf{x}\right)-\mathbf{y}_1\right)\det\left(D\left(\mathbf{g}\left(\mathbf{x}\right)\right)\right)dx \Big|^{t=1}_{t=0} \\ &= \int_\Omega \phi_\varepsilon\left(\mathbf{g}\left(\mathbf{x}\right)-\mathbf{y}\right)\det\left(D\left(\mathbf{g}\left(\mathbf{x}\right)\right)\right)dx \end{aligned}$$

Since $\mathbf{y}_1$ is a regular value of $\mathbf{g}$ it follows from the first part of the argument that the first integral in the above is eventually constant for small enough ε. It follows the last integral is also eventually constant for small enough ε. This proves the claim about the limit existing and in fact being constant for small ε.

The last claim follows right away from the above. Suppose 10.2 holds. Then choosing ε small enough, it follows $d(\mathbf{f},\Omega,\mathbf{y}) = d(\mathbf{g},\Omega,\mathbf{y})$ because the two integrals defining the degree for small ε are equal.

Ignoring the question whether $\mathbf{y}_1$ is a regular value, this shows also that if $\operatorname{dist}(\mathbf{y},\mathbf{g}(\partial\Omega)) > 5\delta$ and if $\mathbf{y}_1 \in B(\mathbf{y},\delta)$, then for all ε small enough,

$$\int_\Omega \phi_\varepsilon(\mathbf{g}(\mathbf{x}) - \mathbf{y}_1)\det(D(\mathbf{g}(\mathbf{x})))\,dx = \int_\Omega \phi_\varepsilon(\mathbf{g}(\mathbf{x}) - \mathbf{y})\det(D(\mathbf{g}(\mathbf{x})))\,dx$$

Thus

$$d(\mathbf{g},\Omega,\mathbf{y}_1) = d(\mathbf{g},\Omega,\mathbf{y}) \ \blacksquare$$

The next theorem is on homotopy invariance.

Theorem 10.2.6 *Let $\mathbf{h} : \bar{\Omega}\times[0,1] \to \mathbb{R}^n$ be such that for each t, $\mathbf{h}(\cdot,t) \in C^2(\overline{\Omega};\mathbb{R}^n)$ and $t \to \mathbf{y}(t)$ is continuous such that for each t, $\mathbf{y}(t) \notin \mathbf{h}(\partial\Omega,t)$. Then*

$$d(\mathbf{h}(\cdot,t),\Omega,\mathbf{y}(t)) \text{ is constant}$$

When $\mathbf{y} \notin \mathbf{f}(\partial\Omega)$ and $\mathbf{f} \in C^2(\overline{\Omega};\mathbb{R}^n)$ and $\mathbf{y}$ is a regular value of $\mathbf{g}$ with $\mathbf{f} \sim \mathbf{g}$,

$$d(\mathbf{f},\Omega,\mathbf{y}) = \sum \left\{ \operatorname{sgn}(\det D\mathbf{g}(\mathbf{x})) : \mathbf{x} \in \mathbf{g}^{-1}(\mathbf{y}) \right\}.$$

The degree is an integer. Also

$$\mathbf{y} \to d(\mathbf{f},\Omega,\mathbf{y})$$

is continuous on $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$ and $\mathbf{y} \to d(\mathbf{f},\Omega,\mathbf{y})$ is constant on every connected component of $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$.

Proof: First of all, there exists a single $\delta > 0$ such that for all t,

$$B(\mathbf{y}(t),6\delta) \cap \mathbf{h}(\partial\Omega,t) = \emptyset \tag{10.5}$$

If this is not so, there exists $\mathbf{x}_n \in \partial\Omega$, t_n, such that $|\mathbf{y}(t_n) - \mathbf{h}(\mathbf{x}_n,t_n)| < 1/n$. Then by compactness, there is a subsequence, still denoted with subscript n such that passing to a limit, we can have $\mathbf{x}_n \to \mathbf{x} \in \partial\Omega$ and $t_n \to t \in [0,1]$. Then by continuity, $\mathbf{y}(t) = \mathbf{h}(\mathbf{x},t)$ contrary to the assumption that for all $t \in [0,1]$, $\mathbf{y}(t) \notin \mathbf{h}(\partial\Omega,t)$.

Now let $0 = t_0 < t_1 < \cdots < t_m = 1$. Let these be close enough together that

$$||\mathbf{h}(\cdot,s) - \mathbf{h}(\cdot,t)||_\infty < \delta, \text{ for all } t,s \in [t_{i-1},t_i] \tag{10.6}$$

$$|\mathbf{y}(t) - \mathbf{y}(s)| < \delta, \text{ for all } t,s \in [t_{i-1},t_i] \tag{10.7}$$

For $s,t \in [t_{i-1},t_i]$, it follows from 10.6

$$B(\mathbf{y}(t),5\delta) \cap \mathbf{h}(\partial\Omega,s) = \emptyset \tag{10.8}$$

By 10.8, 10.5 and Lemma 10.2.5, it follows that

$$d(\mathbf{h}(\cdot,t),\Omega,\mathbf{y}(t)) = d(\mathbf{h}(\cdot,s),\Omega,\mathbf{y}(t))$$

Now from 10.8 and Lemma 10.2.5,

$$d(\mathbf{h}(\cdot,s),\Omega,\mathbf{y}(t)) = d(\mathbf{h}(\cdot,s),\Omega,\mathbf{y}(s))$$

Then the above two equations say that

$$d(\mathbf{h}(\cdot,t),\Omega,\mathbf{y}(t)) = d(\mathbf{h}(\cdot,s),\Omega,\mathbf{y}(t)) = d(\mathbf{h}(\cdot,s),\Omega,\mathbf{y}(s))$$

showing that $t \to d(\mathbf{h}(\cdot,t),\Omega,\mathbf{y}(t))$ is constant on $[t_{i-1},t_i]$. Since this is true for each of these intervals, it follows that this is true on $[0,1]$.

The second assertion follows from Lemma 10.2.5. Finally consider the claim the degree is an integer. This is obvious if $\mathbf{y}$ is a regular point. If $\mathbf{y}$ is not a regular point, let

$$\mathbf{g}_1(\mathbf{x}) \equiv \mathbf{g}(\mathbf{x}) + \mathbf{y} - \mathbf{y}_1$$

where $\mathbf{y}_1$ is a regular point of $\mathbf{g}$ and $|\mathbf{y}-\mathbf{y}_1|$ is so small that

$$\mathbf{y} \notin (t\mathbf{g}_1 + (1-t)\mathbf{g})(\partial\Omega).$$

From Lemma 10.2.5

$$d(\mathbf{g}_1,\Omega,\mathbf{y}) = d(\mathbf{g},\Omega,\mathbf{y}).$$

But since $\mathbf{g}_1 - \mathbf{y} = \mathbf{g} - \mathbf{y}_1$ and $\det D\mathbf{g}(\mathbf{x}) = \det D\mathbf{g}_1(\mathbf{x})$,

$$\begin{aligned} d(\mathbf{g}_1,\Omega,\mathbf{y}) &= \lim_{\varepsilon\to 0}\int_\Omega \phi_\varepsilon(\mathbf{g}_1(\mathbf{x})-\mathbf{y})\det D\mathbf{g}(\mathbf{x})\,dx \\ &= \lim_{\varepsilon\to 0}\int_\Omega \phi_\varepsilon(\mathbf{g}(\mathbf{x})-\mathbf{y}_1)\det D\mathbf{g}(\mathbf{x})\,dx \end{aligned}$$

which by Lemma 10.2.5 equals $\sum\{\operatorname{sgn}(\det D\mathbf{g}(\mathbf{x})) : \mathbf{x} \in \mathbf{g}^{-1}(\mathbf{y}_1)\}$, an integer.

What about the continuity assertion and being constant on connected components? Being constant on connected components follows right away if it can be shown that $\mathbf{y} \to d(\mathbf{f},\Omega,\mathbf{y})$ is continuous. So let $\mathbf{y} \notin \mathbf{f}(\partial\Omega)$. Thus for some $\delta > 0$, $B(\mathbf{y},\delta) \cap \mathbf{f}(\partial\Omega) = \emptyset$. Then if $\hat{\mathbf{y}} \in B(\mathbf{y},\delta/5)$, it follows from Lemma 10.2.5 that

$$d(\mathbf{f},\Omega,\mathbf{y}) = d(\mathbf{f},\Omega,\hat{\mathbf{y}})$$

and so this function is continuous. In fact it is locally constant with integer values. Since it has integer values, it follows from Corollary 5.3.15 on Page 95 that this function must be constant on every connected component. ■

10.2.2 Definition Of The Degree For Continuous Functions

With the above results, it is now possible to extend the definition of the degree to continuous functions which have no differentiability. It is desired to preserve the homotopy invariance. This requires the following definition.

Definition 10.2.7 *Let* $\mathbf{y} \in \mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$ *where* $\mathbf{f} \in C(\bar{\Omega};\mathbb{R}^n)$ *Then*

$$d(\mathbf{f},\Omega,\mathbf{y}) \equiv d(\mathbf{g},\Omega,\mathbf{y})$$

where $\mathbf{y} \notin \mathbf{g}(\partial\Omega)$, $\mathbf{g} \in C^2(\overline{\Omega};\mathbb{R}^n)$ *and* $\mathbf{f} \sim \mathbf{g}$.

Theorem 10.2.8 *The definition of the degree given in Definition 10.2.7 is well defined, equals an integer, and satisfies the following properties. In what follows,* $I(\mathbf{x}) = \mathbf{x}$.

1. $d(I,\Omega,\mathbf{y}) = 1$ *if* $\mathbf{y} \in \Omega$.
2. *If* $\Omega_i \subseteq \Omega, \Omega_i$ *open, and* $\Omega_1 \cap \Omega_2 = \emptyset$ *and if* $\mathbf{y} \notin \mathbf{f}\left(\overline{\Omega} \setminus (\Omega_1 \cup \Omega_2)\right)$, *then* $d(\mathbf{f},\Omega_1,\mathbf{y}) + d(\mathbf{f},\Omega_2,\mathbf{y}) = d(\mathbf{f},\Omega,\mathbf{y})$.

3. *If* $\mathbf{y} \notin \mathbf{f}\left(\overline{\Omega} \setminus \Omega_1\right)$ *and* Ω_1 *is an open subset of* Ω*, then*

$$d(\mathbf{f}, \Omega, \mathbf{y}) = d(\mathbf{f}, \Omega_1, \mathbf{y}).$$

4. *For* $\mathbf{y} \in \mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$*, if* $d(\mathbf{f}, \Omega, \mathbf{y}) \neq 0$ *then* $\mathbf{f}^{-1}(\mathbf{y}) \cap \Omega \neq \emptyset$.

5. *If* $t \to \mathbf{y}(t)$ *is continuous* $\mathbf{h} : \bar{\Omega} \times [0,1] \to \mathbb{R}^n$ *is continuous and if* $\mathbf{y}(t) \notin \mathbf{h}(\partial\Omega, t)$ *for all* t*, then* $t \to d(\mathbf{h}(\cdot, t), \Omega, \mathbf{y}(t))$ *is constant.*

6. $d(\cdot, \Omega, \mathbf{y})$ *is defined and constant on*

$$\left\{\mathbf{g} \in C\left(\overline{\Omega}; \mathbb{R}^n\right) : ||\mathbf{g} - \mathbf{f}||_\infty < r\right\}$$

where $r = \operatorname{dist}(\mathbf{y}, \mathbf{f}(\partial\Omega))$.

7. $d(\mathbf{f}, \Omega, \cdot)$ *is constant on every connected component of* $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$.

8. $d(\mathbf{g}, \Omega, \mathbf{y}) = d(\mathbf{f}, \Omega, \mathbf{y})$ *if* $\mathbf{g}|_{\partial\Omega} = \mathbf{f}|_{\partial\Omega}$.

9. *If* $\operatorname{dist}(y, f(\partial\Omega)) \geq \delta$ *and* $|z - y| < \delta$*, then* $d(f, \Omega, y) = d(f, \Omega, z)$.

Proof: First it is necessary to show the definition is well defined. There are two parts to this. First I need to show there exists $\mathbf{g}$ with the desired properties and then I need to show that it doesn't matter which $\mathbf{g}$ I happen to pick. The first part is easy. Let δ be small enough that

$$\overline{B(\mathbf{y}, \delta)} \cap \mathbf{f}(\partial\Omega) = \emptyset.$$

Then by Lemma 10.2.3 there exists $\mathbf{g} \in C^2\left(\overline{\Omega}; \mathbb{R}^n\right)$ such that $||\mathbf{g} - \mathbf{f}||_\infty < \delta$. It follows that for $t \in [0,1]$,

$$\mathbf{y} \notin (t\mathbf{g} + (1-t)\mathbf{f})(\partial\Omega)$$

and so $\mathbf{g} \sim \mathbf{f}$. The reason is that if $\mathbf{x} \in \partial\Omega$,

$$|t\mathbf{g}(\mathbf{x}) + (1-t)\mathbf{f}(\mathbf{x}) - \mathbf{y}| \geq |\mathbf{f}(\mathbf{x}) - \mathbf{y}| - t|\mathbf{g}(\mathbf{x}) - \mathbf{f}(\mathbf{x})| > \delta - \delta = 0$$

Now consider the second part. Suppose $\mathbf{g} \sim \mathbf{f}$ and $\mathbf{g}_1 \sim \mathbf{f}$. Then by Lemma 10.2.3 again

$$\mathbf{g} \sim \mathbf{g}_1$$

Thus there is a function $\mathbf{h} : \bar{\Omega} \times [0,1] \to \mathbb{R}^n$ such that $\mathbf{h}(\mathbf{x}, 0) = \mathbf{g}(\mathbf{x})$ and $\mathbf{h}(\mathbf{x}, 1) = \mathbf{g}_1(\mathbf{x})$. The difficulty is that it is only known that this function is continuous. It is not known that $\mathbf{h}(\cdot, t)$ is $C^2\left(\bar{\Omega}; \mathbb{R}^n\right)$. Let ψ_ε be a mollifier. Thus it is infinitely differentiable, has support in $B(\mathbf{0}, \varepsilon)$ and $\int_{\mathbb{R}^n} \psi_\varepsilon(\mathbf{x})\, dx = 1$. Then define

$$\mathbf{h}_\varepsilon(\mathbf{x}, t) \equiv \mathbf{h}(\cdot, t) * \psi_\varepsilon(\mathbf{x}) \equiv \int_{\mathbb{R}^n} \mathbf{h}(\mathbf{x} - \mathbf{y}, t)\, \psi_\varepsilon(\mathbf{y})\, dy.$$

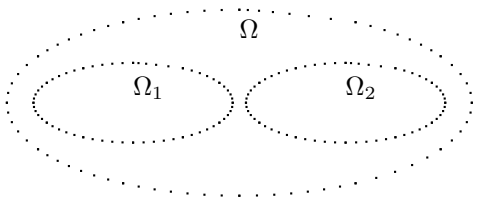
Then as $\varepsilon \to 0$, the convergence is uniform on $\bar{\Omega} \times [0,1]$. Now just as in the first part of the proof of Theorem 10.2.6 there exists $\delta > 0$ such that $B(\mathbf{y}, 6\delta) \cap \mathbf{h}(\partial\Omega, t) = \emptyset$ for all t. Hence, by uniform convergence, for ε small enough, $B(\mathbf{y}, 5\delta) \cap \mathbf{h}_\varepsilon(\partial\Omega, t) = \emptyset$ for all t. Then by Theorem 10.2.6 it follows $d(\mathbf{h}_\varepsilon(\cdot, 0), \Omega, \mathbf{y}) = d(\mathbf{h}_\varepsilon(\cdot, 1), \Omega, \mathbf{y})$. But the same theorem or Lemma 10.2.5,

$$\begin{aligned} d(\mathbf{h}_\varepsilon(\cdot, 1), \Omega, \mathbf{y}) &= d\left(\overset{\mathbf{g}_1}{\mathbf{h}}(\cdot, 1), \Omega, \mathbf{y}\right) \equiv d(\mathbf{g}_1, \Omega, \mathbf{y}) \\ d(\mathbf{h}_\varepsilon(\cdot, 0), \Omega, \mathbf{y}) &= d\left(\overset{\mathbf{g}}{\mathbf{h}}(\cdot, 0), \Omega, \mathbf{y}\right) \equiv d(\mathbf{g}, \Omega, \mathbf{y}) \end{aligned}$$

Thus the definition is well defined because $d(\mathbf{g}_1,\Omega,\mathbf{y}) = d(\mathbf{g},\Omega,\mathbf{y})$.

Now consider the properties. The first one is obvious from Theorem 10.2.6 since $\mathbf{y}$ is a regular point of I.

Consider the second property.



The assumption implies

$$\mathbf{y} \notin \mathbf{f}(\partial\Omega) \cup \mathbf{f}(\partial\Omega_1) \cup \mathbf{f}(\partial\Omega_2)$$

Recall that $\mathbf{y} \notin \mathbf{f}\left(\overline{\Omega} \setminus (\Omega_1 \cup \Omega_2)\right)$. Let $\mathbf{g} \in C^2\left(\overline{\Omega};\mathbb{R}^n\right)$ such that $||\mathbf{f}-\mathbf{g}||_\infty$ is small enough that

$$\mathbf{y} \notin \mathbf{g}\left(\overline{\Omega} \setminus (\Omega_1 \cup \Omega_2)\right) \tag{10.9}$$

and also small enough that

$$\begin{array}{rcl} \mathbf{y} & \notin & (t\mathbf{g}+(1-t)\mathbf{f})(\partial\Omega),\ \mathbf{y} \notin (t\mathbf{g}+(1-t)\mathbf{f})(\partial\Omega_1) \\ \mathbf{y} & \notin & (t\mathbf{g}+(1-t)\mathbf{f})(\partial\Omega_2) \end{array} \tag{10.10}$$

for all $t \in [0,1]$. Then it follows from Lemma 10.2.5, for all ε small enough,

$$d(\mathbf{g},\Omega,\mathbf{y}) = \int_\Omega \phi_\varepsilon(\mathbf{g}(\mathbf{x})-\mathbf{y})\, D\mathbf{g}(\mathbf{x})\, dx$$

From 10.9 there is a positive distance between the compact set

$$\mathbf{g}\left(\overline{\Omega} \setminus (\Omega_1 \cup \Omega_2)\right)$$

and $\mathbf{y}$. Therefore, making ε still smaller if necessary,

$$\phi_\varepsilon(\mathbf{g}(\mathbf{x})-\mathbf{y}) = 0 \text{ if } \mathbf{x} \notin \Omega_1 \cup \Omega_2$$

Therefore, using the definition of the degree and 10.10,

$$d(\mathbf{f},\Omega,\mathbf{y}) = d(\mathbf{g},\Omega,\mathbf{y}) = \lim_{\varepsilon\to 0}\int_\Omega \phi_\varepsilon(\mathbf{g}(\mathbf{x})-\mathbf{y})\, D\mathbf{g}(\mathbf{x})\, dx$$

$$\begin{array}{rcl} & = & \lim_{\varepsilon\to 0}\left(\int_{\Omega_1} \phi_\varepsilon(\mathbf{g}(\mathbf{x})-\mathbf{y})\, D\mathbf{g}(\mathbf{x})\, dx + \right. \\ & & \left. \int_{\Omega_2} \phi_\varepsilon(\mathbf{g}(\mathbf{x})-\mathbf{y})\, D\mathbf{g}(\mathbf{x})\, dx\right) \\ & = & d(\mathbf{g},\Omega_1,\mathbf{y}) + d(\mathbf{g},\Omega_2,\mathbf{y}) \\ & = & d(\mathbf{f},\Omega_1,\mathbf{y}) + d(\mathbf{f},\Omega_2,\mathbf{y}) \end{array}$$

This proves the second property.

Consider the third. This really follows from the second property. You can take $\Omega_2 = \emptyset$. I leave the details to you. To be more careful, you can modify the proof of Property 2 slightly.

The fourth property is very important because it can be used to deduce the existence of solutions to a nonlinear equation. Suppose $\mathbf{f}^{-1}(\mathbf{y}) \cap \Omega = \emptyset$. I will show this requires $d(\mathbf{f}, \Omega, \mathbf{y}) = 0$. It is assumed $\mathbf{y} \notin \mathbf{f}(\partial\Omega)$ and so if $\mathbf{f}^{-1}(\mathbf{y}) \cap \Omega = \emptyset$, then $\mathbf{y} \notin \mathbf{f}(\overline{\Omega})$. Choosing $\mathbf{g} \in C^2(\overline{\Omega}; \mathbb{R}^n)$ such that $||\mathbf{f} - \mathbf{g}||_\infty$ is sufficiently small, it can be assumed

$$\mathbf{y} \notin \mathbf{g}(\overline{\Omega}),\ \mathbf{y} \notin ((1-t)\mathbf{f} + t\mathbf{g})(\partial\Omega) \text{ for all } t \in [0,1].$$

Then it follows from the definition of the degree

$$d(\mathbf{f}, \Omega, \mathbf{y}) = d(\mathbf{g}, \Omega, \mathbf{y}) \equiv \lim_{\varepsilon \to 0} \int_\Omega \phi_\varepsilon(\mathbf{g}(\mathbf{x}) - \mathbf{y}) D\mathbf{g}(\mathbf{x})\, dx = 0$$

because eventually ε is smaller than the distance from $\mathbf{y}$ to $\mathbf{g}(\overline{\Omega})$ and so $\phi_\varepsilon(\mathbf{g}(\mathbf{x}) - \mathbf{y}) = 0$ for all $\mathbf{x} \in \Omega$.

Consider the fifth property. As in Theorem 10.2.6, there is a $\delta > 0$ such that $B(\mathbf{y}(t), 6\delta) \cap \mathbf{h}(\partial\Omega, t)$ for all t. As in showing the definition is well defined, let ψ_ε be a mollifier and let $\mathbf{h}_\varepsilon(\mathbf{x}, t) \equiv \mathbf{h}(\cdot, t) * \psi_\varepsilon(\mathbf{x})$. Then by the uniform convergence, whenever ε is sufficiently small, $B(\mathbf{y}(t), 5\delta) \cap \mathbf{h}_\varepsilon(\partial\Omega, t) = \emptyset$ because for all t,

$$\|\mathbf{h}(\cdot, t) - \mathbf{h}_\varepsilon(\cdot, t)\|_\infty < \delta.$$

Therefore, $\mathbf{h}(\cdot, t) \sim \mathbf{h}_\varepsilon(\cdot, t)$ for all t since $\mathbf{y}(t) \notin (1-\lambda)\mathbf{h}(\partial\Omega, t) + \lambda\mathbf{h}_\varepsilon(\partial\Omega, t), \lambda \in [0,1]$. To see this, let $\mathbf{x} \in \partial\Omega$

$$\begin{aligned} & |(1-\lambda)\mathbf{h}(\mathbf{x}, t) + \lambda\mathbf{h}_\varepsilon(\mathbf{x}, t) - \mathbf{y}(t)| \\ \geq\ & |\mathbf{h}(\mathbf{x}, t) - \mathbf{y}(t)| - \lambda|\mathbf{h}(\mathbf{x}, t) - \mathbf{h}_\varepsilon(\mathbf{x}, t)| \\ \geq\ & 6\delta - \lambda\delta \geq 5\delta > 0 \end{aligned}$$

Then from the definition of the degree above, which was shown above to be well defined, it follows that for all t,

$$d(\mathbf{h}(\cdot, t), \Omega, \mathbf{y}(t)) = d(\mathbf{h}_\varepsilon(\cdot, t), \Omega, \mathbf{y}(t))$$

and the expression on the right is constant in t.

Consider the sixth property. Just consider $\mathbf{h}(\mathbf{x}, t) = t\mathbf{g}(\mathbf{x}) + (1-t)\mathbf{f}(\mathbf{x})$. Then note that $\mathbf{y} \notin \mathbf{h}(\partial\Omega, t)$ and use property 5.

The seventh claim is done already for the case where $\mathbf{f} \in C^2(\overline{\Omega}; \mathbb{R}^n)$ in Theorem 10.2.6. It remains to verify this for the case where $\mathbf{f}$ is only continuous. This will be done by showing $\mathbf{y} \to d(\mathbf{f},\Omega,\mathbf{y})$ is continuous. Let $\mathbf{y}_0 \in \mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$ and let δ be small enough that

$$B(\mathbf{y}_0, 4\delta) \cap \mathbf{f}(\partial\Omega) = \emptyset.$$

Now let $\mathbf{g} \in C^2(\overline{\Omega}; \mathbb{R}^n)$ such that $||\mathbf{g} - \mathbf{f}||_\infty < \delta$. Then for $\mathbf{x} \in \partial\Omega, t \in [0,1]$, and $\mathbf{y} \in B(\mathbf{y}_0, \delta)$,

$$|(t\mathbf{g} + (1-t)\mathbf{f})(\mathbf{x}) - \mathbf{y}| \geq |\mathbf{f}(\mathbf{x}) - \mathbf{y}| - t|\mathbf{g}(\mathbf{x}) - \mathbf{f}(\mathbf{x})|$$

$$\begin{aligned} \geq\ & |\mathbf{f}(\mathbf{x}) - \mathbf{y}_0| - |\mathbf{y}_0 - \mathbf{y}| - ||\mathbf{g} - \mathbf{f}||_\infty \\ \geq\ & 4\delta - \delta - \delta > 0. \end{aligned}$$

Therefore, for all such $\mathbf{y} \in B(\mathbf{y}_0, \delta)$

$$d(\mathbf{f}, \Omega, \mathbf{y}) = d(\mathbf{g}, \Omega, \mathbf{y})$$

and it was shown in Theorem 10.2.6 that $\mathbf{y} \to d(\mathbf{g}, \Omega, \mathbf{y})$ is continuous. In particular $d(\mathbf{f}, \Omega, \cdot)$ is continuous at $\mathbf{y}_0$. Since $\mathbf{y}_0$ was arbitrary, this shows $\mathbf{y} \to d(\mathbf{f}, \Omega, \mathbf{y})$ is

continuous. Therefore, since it has integer values, this function is constant on every connected component of $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$ by Corollary 5.3.15.

Consider the eighth claim about the degree in which $\mathbf{f} = \mathbf{g}$ on $\partial\Omega$. This one is easy because for

$$\mathbf{y} \in \mathbb{R}^n \setminus \mathbf{f}(\partial\Omega) = \mathbb{R}^n \setminus \mathbf{g}(\partial\Omega),$$

and $\mathbf{x} \in \partial\Omega$,

$$t\mathbf{f}(\mathbf{x}) + (1-t)\mathbf{g}(\mathbf{x}) - \mathbf{y} = \mathbf{f}(\mathbf{x}) - \mathbf{y} \neq \mathbf{0}$$

for all $t \in [0,1]$ and so by the fifth claim, $d(\mathbf{f}, \Omega, \mathbf{y}) = d(\mathbf{g}, \Omega, \mathbf{y})$

Finally, consider the last claim. Let $\mathbf{y}(t) \equiv (1-t)\mathbf{y} + t\mathbf{z}$. Then for $\mathbf{x} \in \partial\Omega$

$$\begin{aligned} |(1-t)\mathbf{y} + t\mathbf{z} - \mathbf{f}(\mathbf{x})| &= |\mathbf{y} - \mathbf{f}(\mathbf{x}) + t(\mathbf{z} - \mathbf{y})| \\ &\geq \delta - t|\mathbf{z} - \mathbf{y}| > \delta - \delta = 0 \end{aligned}$$

Then by the fifth property, $d(\mathbf{f}, \Omega, (1-t)\mathbf{y} + t\mathbf{z})$ is constant. When $t = 0$ you get $d(\mathbf{f}, \Omega, \mathbf{y})$ and when $t = 1$ you get $d(\mathbf{f}, \Omega, \mathbf{z})$. ■

10.3 Borsuk's Theorem

In this section is an important theorem which can be used to verify that $d(\mathbf{f}, \Omega, \mathbf{y}) \neq 0$. This is significant because when this is known, it follows from Theorem 10.2.8 that $\mathbf{f}^{-1}(\mathbf{y}) \neq \emptyset$. In other words there exists $\mathbf{x} \in \Omega$ such that $\mathbf{f}(\mathbf{x}) = \mathbf{y}$.

Definition 10.3.1 *A bounded open set, Ω is symmetric if $-\Omega = \Omega$. A continuous function, $\mathbf{f} : \overline{\Omega} \to \mathbb{R}^n$ is odd if $\mathbf{f}(-\mathbf{x}) = -\mathbf{f}(\mathbf{x})$.*

Suppose Ω is symmetric and $\mathbf{g} \in C^2\left(\overline{\Omega}; \mathbb{R}^n\right)$ is an odd map for which $\mathbf{0}$ is a regular value. Then the chain rule implies $D\mathbf{g}(-\mathbf{x}) = D\mathbf{g}(\mathbf{x})$ and so $d(\mathbf{g}, \Omega, \mathbf{0})$ must equal an odd integer because if $\mathbf{x} \in \mathbf{g}^{-1}(\mathbf{0})$, it follows that $-\mathbf{x} \in \mathbf{g}^{-1}(\mathbf{0})$ also and since $D\mathbf{g}(-\mathbf{x}) = D\mathbf{g}(\mathbf{x})$, it follows the overall contribution to the degree from $\mathbf{x}$ and $-\mathbf{x}$ must be an even integer. Also $\mathbf{0} \in \mathbf{g}^{-1}(\mathbf{0})$ and so the degree equals an even integer added to $\operatorname{sgn}(\det D\mathbf{g}(\mathbf{0}))$, an odd integer, either -1 or 1. It seems reasonable to expect that something like this would hold for an arbitrary continuous odd function defined on symmetric Ω. In fact this is the case and this is next. The following lemma is the key result used. This approach is due to Gromes [20]. See also Deimling [9] which is where I found this argument.

The idea is to start with a smooth odd map and approximate it with a smooth odd map which also has $\mathbf{0}$ a regular value.

Lemma 10.3.2 *Let $\mathbf{g} \in C^2\left(\overline{\Omega}; \mathbb{R}^n\right)$ be an odd map. Then for every $\varepsilon > 0$, there exists $\mathbf{h} \in C^2\left(\overline{\Omega}; \mathbb{R}^n\right)$ such that $\mathbf{h}$ is also an odd map, $||\mathbf{h} - \mathbf{g}||_\infty < \varepsilon$, and $\mathbf{0}$ is a regular value of $\mathbf{h}$. Here Ω is a symmetric bounded open set. In addition, $d(\mathbf{g}, \Omega, \mathbf{0})$ is an odd integer.*

Proof: In this argument $\eta > 0$ will be a small positive number and C will be a constant which depends only on the diameter of Ω. Let $\mathbf{h}_0(\mathbf{x}) = \mathbf{g}(\mathbf{x}) + \eta\mathbf{x}$ where η is chosen such that $\det D\mathbf{h}_0(\mathbf{0}) \neq 0$. Now let $\Omega_i \equiv \{\mathbf{x} \in \Omega : x_i \neq 0\}$. In other words, leave out the plane $x_i = 0$ from Ω in order to obtain Ω_i. A succession of modifications is about to take place on $\Omega_1, \Omega_1 \cup \Omega_2$, etc. Finally a function will be obtained on $\cup_{j=1}^n \Omega_j$ which is everything except $\mathbf{0}$.

Define $\mathbf{h}_1(\mathbf{x}) \equiv \mathbf{h}_0(\mathbf{x}) - \mathbf{y}^1 x_1^3$ where $\left|\mathbf{y}^1\right| < \eta$ and $\mathbf{y}^1 = \left(y_1^1, \cdots, y_n^1\right)$ is a regular value of the function, $\mathbf{x} \to \frac{\mathbf{h}_0(\mathbf{x})}{x_1^3}$ for $\mathbf{x} \in \Omega_1$. The existence of $\mathbf{y}^1$ follows from Sard's lemma

because this function is in $C^2(\Omega_1;\mathbb{R}^n)$. Thus $\mathbf{h}_1(\mathbf{x})=\mathbf{0}$ if and only if $\mathbf{y}^1=\frac{\mathbf{h}_0(\mathbf{x})}{x_1^3}$. Since $\mathbf{y}^1$ is a regular value, it follows that for such $\mathbf{x}$,

$$\det\left(\frac{h_{0i,j}(\mathbf{x})\,x_1^3-\frac{\partial}{\partial x_j}\left(x_1^3\right)h_{0i}(\mathbf{x})}{x_1^6}\right)=$$

$$\det\left(\frac{h_{0i,j}(\mathbf{x})\,x_1^3-\frac{\partial}{\partial x_j}\left(x_1^3\right)y_i^1x_1^3}{x_1^6}\right)\neq 0$$

implying that

$$\det\left(h_{0i,j}(\mathbf{x})-\frac{\partial}{\partial x_j}\left(x_1^3\right)y_i^1\right)=\det\left(D\mathbf{h}_1(\mathbf{x})\right)\neq 0.$$

This shows $\mathbf{0}$ is a regular value of $\mathbf{h}_1$ on the set Ω_1 and it is clear $\mathbf{h}_1$ is an odd map in $C^2\left(\overline{\Omega};\mathbb{R}^n\right)$ and $||\mathbf{h}_1-\mathbf{g}||_\infty\leq C\eta$ where C depends only on the diameter of Ω.

Now suppose for some k such that $1\leq k<n$ there exists an odd mapping $\mathbf{h}_k$ in $C^2\left(\overline{\Omega};\mathbb{R}^n\right)$ such that $\mathbf{0}$ is a regular value of $\mathbf{h}_k$ on $\cup_{i=1}^k\Omega_i$ and $||\mathbf{h}_k-\mathbf{g}||_\infty\leq C\eta$. Sard's theorem implies there exists $\mathbf{y}^{k+1}$ a regular value of the function $\mathbf{x}\to\mathbf{h}_k(\mathbf{x})/x_{k+1}^3$ defined on Ω_{k+1}such that $\left|\left|\mathbf{y}^{k+1}\right|\right|<\eta$ and let $\mathbf{h}_{k+1}(\mathbf{x})\equiv\mathbf{h}_k(\mathbf{x})-\mathbf{y}^{k+1}x_{k+1}^3$. As before, $\mathbf{h}_{k+1}(\mathbf{x})=\mathbf{0}$ if and only if $\mathbf{h}_k(\mathbf{x})/x_{k+1}^3=\mathbf{y}^{k+1}$, a regular value of $\mathbf{x}\to\mathbf{h}_k(\mathbf{x})/x_{k+1}^3$. Consider such $\mathbf{x}$ for which $\mathbf{h}_{k+1}(\mathbf{x})=\mathbf{0}$. First suppose $\mathbf{x}\in\Omega_{k+1}$. Then

$$\det\left(\frac{h_{ki,j}(\mathbf{x})\,x_{k+1}^3-\frac{\partial}{\partial x_j}\left(x_{k+1}^3\right)y_i^{k+1}x_{k+1}^3}{x_{k+1}^6}\right)\neq 0$$

which implies that whenever $\mathbf{h}_{k+1}(\mathbf{x})=\mathbf{0}$ and $\mathbf{x}\in\Omega_{k+1}$,

$$\det\left(h_{ki,j}(\mathbf{x})-\frac{\partial}{\partial x_j}\left(x_{k+1}^3\right)y_i^{k+1}\right)=\det\left(D\mathbf{h}_{k+1}(\mathbf{x})\right)\neq 0. \tag{10.11}$$

However, if $\mathbf{x}\in\cup_{i=1}^k\Omega_k$ but $\mathbf{x}\notin\Omega_{k+1}$, then $x_{k+1}=0$ and so the left side of 10.11 reduces to $\det\left(h_{ki,j}(\mathbf{x})\right)$ which is not zero because $\mathbf{0}$ is assumed a regular value of $\mathbf{h}_k$. Therefore, $\mathbf{0}$ is a regular value for $\mathbf{h}_{k+1}$ on $\cup_{i=1}^{k+1}\Omega_k$. (For $\mathbf{x}\in\cup_{i=1}^{k+1}\Omega_k$, either $\mathbf{x}\in\Omega_{k+1}$ or $\mathbf{x}\notin\Omega_{k+1}$. If $\mathbf{x}\in\Omega_{k+1}$ $\mathbf{0}$ is a regular value by the construction above. In the other case, $\mathbf{0}$ is a regular value by the induction hypothesis.) Also $\mathbf{h}_{k+1}$ is odd and in $C^2\left(\overline{\Omega};\mathbb{R}^n\right)$, and $||\mathbf{h}_{k+1}-\mathbf{g}||_\infty\leq C\eta$.

Let $\mathbf{h}\equiv\mathbf{h}_n$. Then $\mathbf{0}$ is a regular value of $\mathbf{h}$ for $\mathbf{x}\in\cup_{j=1}^n\Omega_j$. The point of Ω which is not in $\cup_{j=1}^n\Omega_j$ is $\mathbf{0}$. If $\mathbf{x}=\mathbf{0}$, then from the construction, $D\mathbf{h}(\mathbf{0})=D\mathbf{h}_0(\mathbf{0})$ and so $\mathbf{0}$ is a regular value of $\mathbf{h}$ for $\mathbf{x}\in\Omega$. By choosing η small enough, it follows $||\mathbf{h}-\mathbf{g}||_\infty<\varepsilon$.

For the last part, let $3\delta=\operatorname{dist}\left(\mathbf{g}(\partial\Omega),\mathbf{0}\right)$ and let $\mathbf{h}$ be as described above with $\|\mathbf{h}-\mathbf{g}\|_\infty<\delta$. Then $\mathbf{0}\notin\left(t\mathbf{h}+(1-t)\mathbf{g}\right)(\partial\Omega)$ and so by the homotopy invariance of the degree, $t\to d\left(t\mathbf{h}+(1-t)\mathbf{g},\Omega,\mathbf{0}\right)$ is constant for $t\in[0,1]$. Therefore,

$$d(\mathbf{g},\Omega,\mathbf{0})=d(\mathbf{h},\Omega,\mathbf{0})$$

So what is $d(\mathbf{h},\Omega,\mathbf{0})$? Since $\mathbf{0}$ is a regular value and $\mathbf{h}$ is odd, $\mathbf{h}^{-1}(\mathbf{0})=\{\mathbf{x}_1,\cdots,\mathbf{x}_r,-\mathbf{x}_1,\cdots,-\mathbf{x}_r,\mathbf{0}\}$. So consider $D\mathbf{h}(\mathbf{x})$ and $D\mathbf{h}(-\mathbf{x})$.

$$\begin{aligned}D\mathbf{h}(-\mathbf{x})\,\mathbf{u}+\mathbf{o}(\mathbf{u}) &= \mathbf{h}(-\mathbf{x}+\mathbf{u})-\mathbf{h}(-\mathbf{x})\\ &= -\mathbf{h}(\mathbf{x}+(-\mathbf{u}))+\mathbf{h}(\mathbf{x})\\ &= -\left(D\mathbf{h}(\mathbf{x})(-\mathbf{u})\right)+\mathbf{o}(-\mathbf{u})\\ &= D\mathbf{h}(\mathbf{x})(\mathbf{u})+\mathbf{o}(\mathbf{u})\end{aligned}$$

Hence $D\mathbf{h}(\mathbf{x}) = D\mathbf{h}(-\mathbf{x})$ and so the determinants of these two are the same. It follows that

$$\begin{aligned} d(\mathbf{h},\Omega,\mathbf{0}) &= \sum_{i=1}^{r} \operatorname{sgn}(\det(D\mathbf{h}(\mathbf{x}_i))) + \sum_{i=1}^{r} \operatorname{sgn}(\det(D\mathbf{h}(-\mathbf{x}_i))) + \operatorname{sgn}(\det(D\mathbf{h}(\mathbf{0}))) \\ &= 2m \pm 1 \text{ some integer } m \end{aligned}$$

an odd integer. ■

Theorem 10.3.3 *(Borsuk) Let $\mathbf{f} \in C\left(\overline{\Omega};\mathbb{R}^n\right)$ be odd and let Ω be symmetric with $\mathbf{0} \notin \mathbf{f}(\partial\Omega)$. Then $d(\mathbf{f},\Omega,\mathbf{0})$ equals an odd integer.*

Proof: Let ψ_n be a mollifier which is symmetric, $\psi(-\mathbf{x}) = \psi(\mathbf{x})$. Also recall that $\mathbf{f}$ is the restriction to $\overline{\Omega}$ of a continuous function, still denoted as $\mathbf{f}$ which is defined on all of $\mathbb{R}^n$. Let $\mathbf{g}$ be the odd part of this function. That is,

$$\mathbf{g}(\mathbf{x}) \equiv \frac{1}{2}(\mathbf{f}(\mathbf{x}) - \mathbf{f}(-\mathbf{x}))$$

Since $\mathbf{f}$ is odd, $\mathbf{g} = \mathbf{f}$ on $\overline{\Omega}$. Then

$$\mathbf{g}_n(-\mathbf{x}) \equiv \mathbf{g} * \psi_n(-\mathbf{x}) = \int_{\mathbb{R}^n} \mathbf{g}(-\mathbf{x}-\mathbf{y})\,\psi_n(\mathbf{y})\,dy$$

$$= -\int_{\mathbb{R}^n} \mathbf{g}(\mathbf{x}+\mathbf{y})\,\psi_n(\mathbf{y})\,dy = -\int_{\mathbb{R}^n} \mathbf{g}(\mathbf{x}-(-\mathbf{y}))\,\psi_n(-\mathbf{y})\,dy = -\mathbf{g}_n(\mathbf{x})$$

Thus $\mathbf{g}_n$ is odd and is infinitely differentiable. Let $3\delta = \operatorname{dist}(\mathbf{f}(\partial\Omega),\mathbf{0})$ and let n be large enough that $\|\mathbf{g}_n - \mathbf{f}\|_\infty < \delta$. Then $\mathbf{0} \notin (t\mathbf{g}_n + (1-t)\mathbf{f})(\partial\Omega)$ for $t \in [0,1]$ and so by homotopy invariance,

$$d(\mathbf{f},\Omega,\mathbf{0}) = d(\mathbf{g},\Omega,\mathbf{0}) = d(\mathbf{g}_n,\Omega,\mathbf{0})$$

and by Lemma 10.3.2 this is an odd integer. ■

10.4 Applications

With these theorems it is possible to give easy proofs of some very important and difficult theorems.

Definition 10.4.1 *If $\mathbf{f} : U \subseteq \mathbb{R}^n \to \mathbb{R}^n$ where U is an open set. Then $\mathbf{f}$ is locally one to one if for every $\mathbf{x} \in U$, there exists $\delta > 0$ such that $\mathbf{f}$ is one to one on $B(\mathbf{x},\delta)$.*

As a first application, consider the invariance of domain theorem. This result says that a one to one continuous map takes open sets to open sets. It is an amazing result which is essential to understand if you wish to study manifolds. In fact, the following theorem only requires $\mathbf{f}$ to be locally one to one. First here is a lemma which has the main idea.

Lemma 10.4.2 *Let $\mathbf{g} : \overline{B(\mathbf{0},r)} \to \mathbb{R}^n$ be one to one and continuous where here $B(\mathbf{0},r)$ is the ball centered at $\mathbf{0}$ of radius r in $\mathbb{R}^n$. Then there exists $\delta > 0$ such that*

$$\mathbf{g}(\mathbf{0}) + B(\mathbf{0},\delta) \subseteq \mathbf{g}(B(\mathbf{0},r)).$$

The symbol on the left means: $\{\mathbf{g}(\mathbf{0}) + \mathbf{x} : \mathbf{x} \in B(\mathbf{0},\delta)\}$.

Proof: For $t \in [0,1]$, let

$$\mathbf{h}(\mathbf{x},t) \equiv \mathbf{g}\left(\frac{\mathbf{x}}{1+t}\right) - \mathbf{g}\left(\frac{-t\mathbf{x}}{1+t}\right)$$

Then for $\mathbf{x} \in \partial B(\mathbf{0},r)$, $\mathbf{h}(\mathbf{x},t) \neq \mathbf{0}$ because if this were so, the fact $\mathbf{g}$ is one to one implies

$$\frac{\mathbf{x}}{1+t} = \frac{-t\mathbf{x}}{1+t}$$

and this requires $\mathbf{x} = \mathbf{0}$ which is not the case since $\|\mathbf{x}\| = r$. Then $d(\mathbf{h}(\cdot,t), B(\mathbf{0},r), \mathbf{0})$ is constant. Hence it is an odd integer for all t thanks to Borsuk's theorem, because $\mathbf{h}(\cdot,1)$ is odd. Now let $B(\mathbf{0},\delta)$ be such that $B(\mathbf{0},\delta) \cap \mathbf{h}(\partial\Omega,0) = \emptyset$. Then $d(\mathbf{h}(\cdot,0), B(\mathbf{0},r), \mathbf{0}) = d(\mathbf{h}(\cdot,0), B(\mathbf{0},r), \mathbf{z})$ because the degree is constant on connected components of $\mathbb{R}^n \setminus \mathbf{h}(\partial\Omega,0)$. Hence $\mathbf{z} = \mathbf{h}(\mathbf{x},0) = \mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{0})$ for some $\mathbf{x} \in B(0,r)$. Thus

$$\mathbf{g}(B(\mathbf{0},r)) \supseteq \mathbf{g}(\mathbf{0}) + B(\mathbf{0},\delta) \ \blacksquare$$

Now with this lemma, it is easy to prove the very important invariance of domain theorem.

A function $\mathbf{f}$ is locally one to one on an open set Ω if for every $\mathbf{x}_0 \in \Omega$, there exists $B(\mathbf{x}_0,r) \subseteq \Omega$ such that $\mathbf{f}$ is one to one on $B(\mathbf{x}_0,r)$.

Theorem 10.4.3 *(invariance of domain)Let Ω be any open subset of $\mathbb{R}^n$ and let $\mathbf{f}: \Omega \to \mathbb{R}^n$ be continuous and locally one to one. Then $\mathbf{f}$ maps open subsets of Ω to open sets in $\mathbb{R}^n$.*

Proof: Let $\overline{B(\mathbf{x}_0,r)} \subseteq \Omega$ where $\mathbf{f}$ is one to one on $\overline{B(\mathbf{x}_0,r)}$. Let $\mathbf{g}$ be defined on $\overline{B(\mathbf{0},r)}$ given by

$$\mathbf{g}(\mathbf{x}) \equiv \mathbf{f}(\mathbf{x} + \mathbf{x}_0)$$

Then $\mathbf{g}$ satisfies the conditions of Lemma 10.4.2, being one to one and continuous. It follows from that lemma there exists $\delta > 0$ such that

$$\begin{aligned} \mathbf{f}(\Omega) &\supseteq \mathbf{f}(B(\mathbf{x}_0,r)) = \mathbf{f}(\mathbf{x}_0 + B(\mathbf{0},r)) \\ &= \mathbf{g}(B(\mathbf{0},r)) \supseteq \mathbf{g}(\mathbf{0}) + B(\mathbf{0},\delta) \\ &= \mathbf{f}(\mathbf{x}_0) + B(\mathbf{0},\delta) = B(\mathbf{f}(\mathbf{x}_0),\delta) \end{aligned}$$

This shows that for any $\mathbf{x}_0 \in \Omega$, $\mathbf{f}(\mathbf{x}_0)$ is an interior point of $\mathbf{f}(\Omega)$ which shows $\mathbf{f}(\Omega)$ is open. $\blacksquare$

With the above, one gets easily the following amazing result. It is something which is clear for linear maps but this is a statement about continuous maps.

Corollary 10.4.4 *If $n > m$ there does not exist a continuous one to one map from $\mathbb{R}^n$ to $\mathbb{R}^m$.*

Proof: Suppose not and let $\mathbf{f}$ be such a continuous map,

$$\mathbf{f}(\mathbf{x}) \equiv (f_1(\mathbf{x}), \cdots, f_m(\mathbf{x}))^T.$$

Then let $\mathbf{g}(\mathbf{x}) \equiv (f_1(\mathbf{x}), \cdots, f_m(\mathbf{x}), 0, \cdots, 0)^T$ where there are $n-m$ zeros added in. Then $\mathbf{g}$ is a one to one continuous map from $\mathbb{R}^n$ to $\mathbb{R}^n$ and so $\mathbf{g}(\mathbb{R}^n)$ would have to be open from the invariance of domain theorem and this is not the case. $\blacksquare$

Corollary 10.4.5 *If $\mathbf{f}$ is locally one to one and continuous, $\mathbf{f}: \mathbb{R}^n \to \mathbb{R}^n$, and*

$$\lim_{|\mathbf{x}|\to\infty} |\mathbf{f}(\mathbf{x})| = \infty,$$

then $\mathbf{f}$ maps $\mathbb{R}^n$ onto $\mathbb{R}^n$.

Proof: By the invariance of domain theorem, $\mathbf{f}(\mathbb{R}^n)$ is an open set. It is also true that $\mathbf{f}(\mathbb{R}^n)$ is a closed set. Here is why. If $\mathbf{f}(\mathbf{x}_k) \to \mathbf{y}$, the growth condition ensures that $\{\mathbf{x}_k\}$ is a bounded sequence. Taking a subsequence which converges to $\mathbf{x} \in \mathbb{R}^n$ and using the continuity of $\mathbf{f}$, it follows $\mathbf{f}(\mathbf{x}) = \mathbf{y}$. Thus $\mathbf{f}(\mathbb{R}^n)$ is both open and closed which implies $\mathbf{f}$ must be an onto map since otherwise, $\mathbb{R}^n$ would not be connected. ■

The next theorem is the famous Brouwer fixed point theorem.

Theorem 10.4.6 *(Brouwer fixed point) Let $B = \overline{B(\mathbf{0},r)} \subseteq \mathbb{R}^n$ and let $\mathbf{f} : B \to B$ be continuous. Then there exists a point $\mathbf{x} \in B$, such that $\mathbf{f}(\mathbf{x}) = \mathbf{x}$.*

Proof: Consider $\mathbf{h}(\mathbf{x},t) \equiv t\mathbf{f}(\mathbf{x}) - \mathbf{x}$ for $t \in [0,1]$. If $\|\mathbf{x}\| = r$, then could $\mathbf{h}(\mathbf{x},t) = \mathbf{0}$? If so,

$$\mathbf{0} = t\mathbf{f}(\mathbf{x}) - \mathbf{x}$$

If $t = 1$, then you would have found a fixed point. Otherwise, if $t < 1$,

$$r = \|\mathbf{x}\| = t\|\mathbf{f}(\mathbf{x})\| < \|\mathbf{f}(\mathbf{x})\| \leq r, \text{ or } r = \|\mathbf{x}\| = t0 = 0$$

a contradiction. Hence one can assume that $\mathbf{0} \notin \mathbf{h}(\partial B, t)$. Then by the homotopy invariance,

$$0 = d(\mathbf{f} - \mathrm{id}, B, \mathbf{0}) = d(-\mathrm{id}, B, \mathbf{0}) = (-1)^n \quad ■$$

It is easy to generalize this to an arbitrary closed bounded convex set in $\mathbb{R}^n$ as follows. Let K be a closed bounded convex set and let $\mathbf{f} : K \to K$ be continuous. Let P be the projection map onto K. Then P is continuous because $|P\mathbf{x} - P\mathbf{y}| \leq |\mathbf{x} - \mathbf{y}|$. Recall why this is. From the material on Hilbert space, $(\mathbf{x} - P\mathbf{x}, \mathbf{y} - P\mathbf{x}) \leq 0$ for all $\mathbf{y} \in K$. Indeed, this characterizes $P\mathbf{x}$. Therefore,

$$(\mathbf{x} - P\mathbf{x}, P\mathbf{y} - P\mathbf{x}) \leq 0, \ (\mathbf{y} - P\mathbf{y}, P\mathbf{x} - P\mathbf{y}) \leq 0 \text{ so } (\mathbf{y} - P\mathbf{y}, P\mathbf{y} - P\mathbf{x}) \geq 0$$

Hence, subtracting the first from the last,

$$(\mathbf{y} - P\mathbf{y} - (\mathbf{x} - P\mathbf{x}), P\mathbf{y} - P\mathbf{x}) \geq 0$$

consequently,

$$|\mathbf{x} - \mathbf{y}|\,|P\mathbf{y} - P\mathbf{x}| \geq (\mathbf{y} - \mathbf{x}, P\mathbf{y} - P\mathbf{x}) \geq |P\mathbf{y} - P\mathbf{x}|^2$$

and so $|P\mathbf{y} - P\mathbf{x}| \leq |\mathbf{y} - \mathbf{x}|$ as claimed.

Now let r be so large that $K \subseteq B(\mathbf{0},r)$. Then consider $\mathbf{f} \circ P$. This map takes $\overline{B(\mathbf{0},r)} \to \overline{B(\mathbf{0},r)}$. In fact it maps $\overline{B(\mathbf{0},r)}$ to K. Therefore, being the composition of continuous functions, it is continuous and so has a fixed point in $\overline{B(\mathbf{0},r)}$ denoted as $\mathbf{x}$. Hence $\mathbf{f}(P(\mathbf{x})) = \mathbf{x}$. Now, since $\mathbf{f}$ maps into K, it follows that $\mathbf{x} \in K$. Hence $P\mathbf{x} = \mathbf{x}$ and so $\mathbf{f}(\mathbf{x}) = \mathbf{x}$. This has proved the following general Brouwer fixed point theorem.

Theorem 10.4.7 *Let $\mathbf{f} : K \to K$ be continuous where K is compact and convex and nonempty. Then $\mathbf{f}$ has a fixed point.*

Definition 10.4.8 *$\mathbf{f}$ is a retraction of $\overline{B(\mathbf{0},r)}$ onto $\partial B(0,r)$ if $\mathbf{f}$ is continuous, $\mathbf{f}\left(\overline{B(\mathbf{0},r)}\right) \subseteq \partial B(\mathbf{0},r)$, and $\mathbf{f}(\mathbf{x}) = \mathbf{x}$ for all $\mathbf{x} \in \partial B(\mathbf{0},r)$.*

Theorem 10.4.9 *There does not exist a retraction of $\overline{B(\mathbf{0},r)}$ onto its boundary, $\partial B(\mathbf{0},r)$.*

Proof: Suppose $\mathbf{f}$ were such a retraction. Then for all $\mathbf{x} \in \partial B(\mathbf{0}, r)$, $\mathbf{f}(\mathbf{x}) = \mathbf{x}$ and so from the properties of the degree, the one which says if two functions agree on $\partial\Omega$, then they have the same degree,

$$1 = d(\mathrm{id}, B(\mathbf{0}, r), \mathbf{0}) = d(\mathbf{f}, B(\mathbf{0}, r), \mathbf{0})$$

which is clearly impossible because $\mathbf{f}^{-1}(\mathbf{0}) = \emptyset$ which implies $d(\mathbf{f}, B(\mathbf{0}, r), \mathbf{0}) = 0$. ■

You should now use this theorem to give another proof of the Brouwer fixed point theorem. See Page 254. You will be able to shorten the argument given there.

The proofs of the next two theorems make use of the Tietze extension theorem, Theorem 5.7.12.

Theorem 10.4.10 *Let Ω be a symmetric open set in $\mathbb{R}^n$ such that $\mathbf{0} \in \Omega$ and let $\mathbf{f} : \partial\Omega \to V$ be continuous where V is an m dimensional subspace of $\mathbb{R}^n, m < n$. Then $\mathbf{f}(-\mathbf{x}) = \mathbf{f}(\mathbf{x})$ for some $\mathbf{x} \in \partial\Omega$.*

Proof: Suppose not. Using the Tietze extension theorem, extend $\mathbf{f}$ to all of $\overline{\Omega}$, $\mathbf{f}(\overline{\Omega}) \subseteq V$. (Here the extended function is also denoted by $\mathbf{f}$.) Let $\mathbf{g}(\mathbf{x}) = \mathbf{f}(\mathbf{x}) - \mathbf{f}(-\mathbf{x})$. Then $\mathbf{0} \notin \mathbf{g}(\partial\Omega)$ and so for some $r > 0$, $B(\mathbf{0},r) \subseteq \mathbb{R}^n \setminus \mathbf{g}(\partial\Omega)$. For $\mathbf{z} \in B(\mathbf{0},r)$,

$$d(\mathbf{g}, \Omega, \mathbf{z}) = d(\mathbf{g}, \Omega, \mathbf{0}) \neq 0$$

because $B(\mathbf{0},r)$ is contained in a component of $\mathbb{R}^n \setminus \mathbf{g}(\partial\Omega)$ and Borsuk's theorem implies that $d(\mathbf{g}, \Omega, \mathbf{0}) \neq 0$ since $\mathbf{g}$ is odd. Hence

$$V \supseteq \mathbf{g}(\Omega) \supseteq B(\mathbf{0},r)$$

and this is a contradiction because V is m dimensional. ■

This theorem is called the Borsuk Ulam theorem. Note that it implies there exist two points on opposite sides of the surface of the earth which have the same atmospheric pressure and temperature, assuming the earth is symmetric and that pressure and temperature are continuous functions. The next theorem is an amusing result which is like combing hair. It gives the existence of a "cowlick".

Theorem 10.4.11 *Let n be odd and let Ω be an open bounded set in $\mathbb{R}^n$ with $\mathbf{0} \in \Omega$. Suppose $\mathbf{f} : \partial\Omega \to \mathbb{R}^n \setminus \{\mathbf{0}\}$ is continuous. Then for some $\mathbf{x} \in \partial\Omega$ and $\lambda \neq 0$, $\mathbf{f}(\mathbf{x}) = \lambda\mathbf{x}$.*

Proof: Using the Tietze extension theorem, extend $\mathbf{f}$ to all of $\overline{\Omega}$. Also denote the extended function by $\mathbf{f}$. Suppose for all $\mathbf{x} \in \partial\Omega$, $\mathbf{f}(\mathbf{x}) \neq \lambda\mathbf{x}$ for all $\lambda \in \mathbb{R}$. Then

$$\mathbf{0} \notin t\mathbf{f}(\mathbf{x}) + (1-t)\mathbf{x}, \ (\mathbf{x}, t) \in \partial\Omega \times [0,1]$$

$$\mathbf{0} \notin t\mathbf{f}(\mathbf{x}) - (1-t)\mathbf{x}, \ (\mathbf{x}, t) \in \partial\Omega \times [0,1].$$

Thus there exists a homotopy of $\mathbf{f}$ and id and a homotopy of $\mathbf{f}$ and $-\mathrm{id}$. Then by the homotopy invariance of degree,

$$d(\mathbf{f}, \Omega, \mathbf{0}) = d(\mathrm{id}, \Omega, \mathbf{0}), \ d(\mathbf{f}, \Omega, \mathbf{0}) = d(-\mathrm{id}, \Omega, \mathbf{0}).$$

But this is impossible because $d(\mathrm{id}, \Omega, \mathbf{0}) = 1$ but $d(-\mathrm{id}, \Omega, \mathbf{0}) = (-1)^n = -1$. ■

10.5 The Product Formula

This section is on the product formula for the degree which is used to prove the Jordan separation theorem. To begin with here is a lemma which is similar to an earlier result except here there are r points.

Lemma 10.5.1 *Let $\mathbf{y}_1, \cdots, \mathbf{y}_r$ be points not in $\mathbf{f}(\partial\Omega)$ and let $\delta > 0$. Then there exists $\widetilde{\mathbf{f}} \in C^2\left(\overline{\Omega}; \mathbb{R}^n\right)$ such that $\left\|\widetilde{\mathbf{f}} - \mathbf{f}\right\|_\infty < \delta$ and $\mathbf{y}_i$ is a regular value for $\widetilde{\mathbf{f}}$ for each i.*

Proof: Let $\mathbf{f}_0 \in C^2\left(\overline{\Omega}; \mathbb{R}^n\right)$, $||\mathbf{f}_0 - \mathbf{f}||_\infty < \frac{\delta}{2}$. Let $\widetilde{\mathbf{y}}_1$ be a regular value for $\mathbf{f}_0$ and $|\widetilde{\mathbf{y}}_1 - \mathbf{y}_1| < \frac{\delta}{3r}$. Let $\mathbf{f}_1(\mathbf{x}) \equiv \mathbf{f}_0(\mathbf{x}) + \mathbf{y}_1 - \widetilde{\mathbf{y}}_1$. Thus $\mathbf{y}_1$ is a regular value of $\mathbf{f}_1$ because $D\mathbf{f}_1(\mathbf{x}) = D\mathbf{f}_0(\mathbf{x})$ and if $\mathbf{f}_1(\mathbf{x}) = \mathbf{y}_1$, this is the same as having $\mathbf{f}_0(\mathbf{x}) = \widetilde{\mathbf{y}}_1$ where $\widetilde{\mathbf{y}}_1$ is a regular value of $\mathbf{f}_0$. Then also

$$\begin{aligned} ||\mathbf{f} - \mathbf{f}_1||_\infty &\leq ||\mathbf{f} - \mathbf{f}_0||_\infty + ||\mathbf{f}_0 - \mathbf{f}_1||_\infty \\ &= ||\mathbf{f} - \mathbf{f}_0||_\infty + |\widetilde{\mathbf{y}}_1 - \mathbf{y}_1| \\ &< \frac{\delta}{3r} + \frac{\delta}{2}. \end{aligned}$$

Suppose now there exists $\mathbf{f}_k \in C^2\left(\overline{\Omega}; \mathbb{R}^n\right)$ with each of the $\mathbf{y}_i$ for $i = 1, \cdots, k$ a regular value of $\mathbf{f}_k$ and

$$||\mathbf{f} - \mathbf{f}_k||_\infty < \frac{\delta}{2} + \frac{k}{r}\left(\frac{\delta}{3}\right).$$

Then letting S_k denote the singular values of $\mathbf{f}_k$, Sard's theorem implies there exists $\widetilde{\mathbf{y}}_{k+1}$ such that

$$|\widetilde{\mathbf{y}}_{k+1} - \mathbf{y}_{k+1}| < \frac{\delta}{3r}$$

and

$$\widetilde{\mathbf{y}}_{k+1} \notin S_k \cup \cup_{i=1}^k (S_k + \mathbf{y}_{k+1} - \mathbf{y}_i). \tag{10.12}$$

Let

$$\mathbf{f}_{k+1}(\mathbf{x}) \equiv \mathbf{f}_k(\mathbf{x}) + \mathbf{y}_{k+1} - \widetilde{\mathbf{y}}_{k+1}. \tag{10.13}$$

If $\mathbf{f}_{k+1}(\mathbf{x}) = \mathbf{y}_i$ for some $i \leq k$, then

$$\mathbf{f}_k(\mathbf{x}) + \mathbf{y}_{k+1} - \mathbf{y}_i = \widetilde{\mathbf{y}}_{k+1}$$

and so $\mathbf{f}_k(\mathbf{x})$ is a regular value for $\mathbf{f}_k$ since by 10.12, $\widetilde{\mathbf{y}}_{k+1} \notin S_k + \mathbf{y}_{k+1} - \mathbf{y}_i$ and so $\mathbf{f}_k(\mathbf{x}) \notin S_k$. Therefore, for $i \leq k$, $\mathbf{y}_i$ is a regular value of $\mathbf{f}_{k+1}$ since by 10.13, $D\mathbf{f}_{k+1} = D\mathbf{f}_k$. Now suppose $\mathbf{f}_{k+1}(\mathbf{x}) = \mathbf{y}_{k+1}$. Then

$$\mathbf{y}_{k+1} = \mathbf{f}_k(\mathbf{x}) + \mathbf{y}_{k+1} - \widetilde{\mathbf{y}}_{k+1}$$

so $\mathbf{f}_k(\mathbf{x}) = \widetilde{\mathbf{y}}_{k+1}$ implying that $\mathbf{f}_k(\mathbf{x}) = \widetilde{\mathbf{y}}_{k+1} \notin S_k$. Hence $\det D\mathbf{f}_{k+1}(\mathbf{x}) = \det D\mathbf{f}_k(\mathbf{x}) \neq 0$. Thus $\mathbf{y}_{k+1}$ is also a regular value of $\mathbf{f}_{k+1}$. Also,

$$\begin{aligned} ||\mathbf{f}_{k+1} - \mathbf{f}|| &\leq ||\mathbf{f}_{k+1} - \mathbf{f}_k|| + ||\mathbf{f}_k - \mathbf{f}|| \\ &\leq \frac{\delta}{3r} + \frac{\delta}{2} + \frac{k}{r}\left(\frac{\delta}{3}\right) = \frac{\delta}{2} + \frac{k+1}{r}\left(\frac{\delta}{3}\right). \end{aligned}$$

Let $\widetilde{\mathbf{f}} \equiv \mathbf{f}_r$. Then

$$\left\|\widetilde{\mathbf{f}} - \mathbf{f}\right\|_\infty < \frac{\delta}{2} + \left(\frac{\delta}{3}\right) < \delta$$

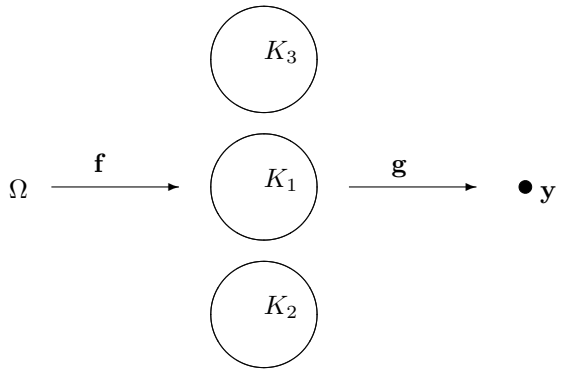
and each of the $\mathbf{y}_i$ is a regular value of $\widetilde{\mathbf{f}}$. ■

Definition 10.5.2 *Let the connected components of* $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$ *be denoted by* K_i. *From the properties of the degree listed in Theorem 10.2.8,* $d(\mathbf{f},\Omega,\cdot)$ *is constant on each of these components. Denote by* $d(\mathbf{f},\Omega,K_i)$ *the constant value on the component,* K_i.

The product formula considers the situation depicted in the following diagram in which $\mathbf{y} \notin \mathbf{g}(\mathbf{f}(\partial\Omega))$ and the K_i are the connected components of $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$.

$$\overline{\Omega} \overset{\mathbf{f}}{\to} \underset{\mathbb{R}^n\setminus\mathbf{f}(\partial\Omega)=\cup_i K_i}{\mathbf{f}\left(\overline{\Omega}\right)} \overset{\mathbf{g}}{\to} \underset{\mathbf{y}}{\mathbb{R}^n}$$

The following diagram may be helpful in remembering what it says.



Lemma 10.5.3 *Let* $\mathbf{f} \in C\left(\overline{\Omega};\mathbb{R}^n\right), \mathbf{g} \in C^2(\mathbb{R}^n,\mathbb{R}^n)$, *and* $\mathbf{y} \notin \mathbf{g}(\mathbf{f}(\partial\Omega))$. *Suppose also that* $\mathbf{y}$ *is a regular value of* $\mathbf{g}$. *Then the following product formula holds where* K_i *are the bounded components of* $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$.

$$d(\mathbf{g}\circ\mathbf{f},\Omega,\mathbf{y}) = \sum_{i=1}^{\infty} d(\mathbf{f},\Omega,K_i)\, d(\mathbf{g},K_i,\mathbf{y}).$$

All but finitely many terms in the sum are zero.

Proof: First note that if K_i is unbounded, $d(\mathbf{f},\Omega,K_i) = 0$ because there exists a point, $\mathbf{z} \in K_i$ such that $\mathbf{f}^{-1}(\mathbf{z}) = \emptyset$ due to the fact that $\mathbf{f}\left(\overline{\Omega}\right)$ is compact and is consequently bounded. Thus it makes no difference in the above formula whether the K_i are arbitrary components or only bounded components. Let $\left\{\mathbf{x}_j^i\right\}_{j=1}^{m_i}$ denote the points of $\mathbf{g}^{-1}(\mathbf{y})$ which are contained in K_i, the i^{th} bounded component of $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$. Then $m_i < \infty$ because if not, there would exist a limit point $\mathbf{x}$ for this sequence. Then $\mathbf{g}(\mathbf{x}) = \mathbf{y}$ and so $\mathbf{x} \notin \mathbf{f}(\partial\Omega)$. Thus $\det(D\mathbf{g}(\mathbf{x})) \neq 0$ and so by the inverse function theorem, $\mathbf{g}$ would be one to one on an open ball containing $\mathbf{x}$ which contradicts having $\mathbf{x}$ a limit point.

Note also that $\mathbf{g}^{-1}(\mathbf{y}) \cap \mathbf{f}\left(\overline{\Omega}\right)$ is a compact set covered by the components of $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$ because by assumption, $\mathbf{g}^{-1}(\mathbf{y}) \cap \mathbf{f}(\partial\Omega) = \emptyset$. It follows $\mathbf{g}^{-1}(\mathbf{y}) \cap \mathbf{f}\left(\overline{\Omega}\right)$ is covered by finitely many of these components. It is not in $\mathbf{f}(\partial\Omega)$.

The only terms in the above sum which are nonzero are those corresponding to K_i having nonempty intersection with $\mathbf{g}^{-1}(\mathbf{y}) \cap \mathbf{f}\left(\overline{\Omega}\right)$. The other components contribute 0 to the above sum because if $K_i \cap \mathbf{g}^{-1}(\mathbf{y}) = \emptyset$, it follows from Theorem 10.2.8 that $d(\mathbf{g},K_i,\mathbf{y}) = 0$. If K_i does not intersect $\mathbf{f}\left(\overline{\Omega}\right)$, then $d(\mathbf{f},\Omega,K_i) = 0$. Therefore, the above sum is actually a finite sum since $\mathbf{g}^{-1}(\mathbf{y}) \cap \mathbf{f}\left(\overline{\Omega}\right)$, being a compact set, is covered by finitely many of the K_i. Thus there are no convergence problems.

Let $d(\mathbf{f},\Omega,K_i) = d\left(\mathbf{f},\Omega,\mathbf{u}_j^i\right)$ where the $\left\{\mathbf{u}_j^i\right\}_{j=1}^{m_i}$ are the points in $\mathbf{g}^{-1}(\mathbf{y}) \cap K_i$. By Lemma 10.5.1, there exists $\tilde{\mathbf{f}}$ such that $\left\|\tilde{\mathbf{f}}-\mathbf{f}\right\|_\infty$ is very small and each of the $\mathbf{u}_j^i$ are regular values for $\tilde{\mathbf{f}}$. If $\left\|\tilde{\mathbf{f}}-\mathbf{f}\right\|_\infty$ is small enough, then $\left(\mathbf{f}+t\left(\widetilde{\mathbf{f}}-\mathbf{f}\right)\right)(\partial\Omega)$ does not contain any of the $\mathbf{u}_j^i$. This is so because by the definition of $\mathbf{u}_j^i$ they are in some K_i and these are connected components of $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$. Thus

$$d(\mathbf{f},\Omega,K_i) \equiv d\left(\mathbf{f},\Omega,\mathbf{u}_j^i\right) = d\left(\tilde{\mathbf{f}},\Omega,\mathbf{u}_j^i\right)$$

by the homotopy invariance of the degree, this for each $j = 1,2,\cdots,m_i$. Also if $\left\|\tilde{\mathbf{f}}-\mathbf{f}\right\|_\infty$ is small enough, one can have $\left(\mathbf{g}\circ\mathbf{f}+t\left(\mathbf{g}\circ\tilde{\mathbf{f}}-\mathbf{g}\circ\mathbf{f}\right)\right)(\partial\Omega)$ does not contain $\mathbf{y}$ for all $t\in[0,1]$. Hence by homotopy invariance again,

$$d(\mathbf{g}\circ\mathbf{f},\Omega,\mathbf{y}) = d\left(\mathbf{g}\circ\widetilde{\mathbf{f}},\Omega,\mathbf{y}\right). \tag{10.14}$$

Now $\widetilde{\mathbf{f}}^{-1}\left(\mathbf{u}_j^i\right)$ is a finite set because $\widetilde{\mathbf{f}}^{-1}\left(\mathbf{u}_j^i\right) \subseteq \Omega$, a bounded open set and $\mathbf{u}_j^i$ is a regular value. It follows from 10.14

$$d(\mathbf{g}\circ\mathbf{f},\Omega,\mathbf{y}) = d\left(\mathbf{g}\circ\widetilde{\mathbf{f}},\Omega,\mathbf{y}\right)$$

$$= \sum_{i=1}^{\infty}\sum_{j=1}^{m_i}\sum_{\mathbf{z}\in\widetilde{\mathbf{f}}^{-1}\left(\mathbf{u}_j^i\right)} \operatorname{sgn}\det D\mathbf{g}\left(\overbrace{\widetilde{\mathbf{f}}(\mathbf{z})}^{\mathbf{u}_j^i}\right)\operatorname{sgn}\det D\widetilde{\mathbf{f}}(\mathbf{z})$$

$$\begin{aligned} &= \sum_{i=1}^{\infty}\sum_{j=1}^{m_i}\operatorname{sgn}\det D\mathbf{g}\left(\mathbf{u}_j^i\right) d\left(\widetilde{\mathbf{f}},\Omega,\mathbf{x}_j^i\right) \\ &= \sum_{i=1}^{\infty} d(\mathbf{g},K_i,\mathbf{y})\, d\left(\widetilde{\mathbf{f}},\Omega,\mathbf{x}_j^i\right) \\ &= \sum_{i=1}^{\infty} d(\mathbf{g},K_i,\mathbf{y})\, d(\mathbf{f},\Omega,K_i). \ \blacksquare \end{aligned}$$

With this lemma, the following is the product formula.

Theorem 10.5.4 *(product formula) Let $\{K_i\}_{i=1}^{\infty}$ be the bounded components of $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$ for $\mathbf{f}\in C\left(\overline{\Omega};\mathbb{R}^n\right)$, let $\mathbf{g}\in C(\mathbb{R}^n,\mathbb{R}^n)$, and suppose that $\mathbf{y}\notin\mathbf{g}(\mathbf{f}(\partial\Omega))$. Then*

$$d(\mathbf{g}\circ\mathbf{f},\Omega,\mathbf{y}) = \sum_{i=1}^{\infty} d(\mathbf{g},K_i,\mathbf{y})\, d(\mathbf{f},\Omega,K_i). \tag{10.15}$$

All but finitely many terms in the sum are zero.

Proof: Let $\sup\left\{\left|\widetilde{\mathbf{g}}(\mathbf{z})-\mathbf{g}(\mathbf{z})\right| : \mathbf{z}\in\mathbf{f}\left(\overline{\Omega}\right)\right\}$ be sufficiently small that

$$\mathbf{y}\notin\left(\mathbf{g}\circ\mathbf{f}+t\left(\widetilde{\mathbf{g}}\circ\mathbf{f}-\mathbf{g}\circ\mathbf{f}\right)\right)(\partial\Omega),\ t\in[0,1]$$

$\tilde{\mathbf{g}}$ being $C^2(\mathbb{R}^n,\mathbb{R}^n)$ with $\mathbf{y}$ a regular point of $\tilde{\mathbf{g}}$. It follows that

$$d(\mathbf{g}\circ\mathbf{f},\Omega,\mathbf{y}) = d\left(\widetilde{\mathbf{g}}\circ\mathbf{f},\Omega,\mathbf{y}\right). \tag{10.16}$$

Now also, the K_i are the open components of $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$ and so $\partial K_i \subseteq \mathbf{f}(\partial\Omega)$ (if $\mathbf{x} \in \partial K_i$, then if $\mathbf{x} \notin \mathbf{f}(\partial\Omega)$, it would be in a ball contained in one of the K_j and so could not be in ∂K_i.) and so if $\mathbf{z} \in \partial K_i$, then $\mathbf{g}(\mathbf{z}) \in \mathbf{g}(\mathbf{f}(\partial\Omega))$. Consequently, for $t \in [0,1]$,

$$\mathbf{y} \notin (\mathbf{g} + t(\widetilde{\mathbf{g}} - \mathbf{g}))(\partial K_i)$$

($\mathbf{y}$ is not in the larger set $(\mathbf{g} \circ \mathbf{f} + t(\widetilde{\mathbf{g}} \circ \mathbf{f} - \mathbf{g} \circ \mathbf{f}))(\partial\Omega)$) which shows that, by homotopy invariance,

$$d(\mathbf{g}, K_i, \mathbf{y}) = d(\widetilde{\mathbf{g}}, K_i, \mathbf{y}). \tag{10.17}$$

Therefore, by Lemma 10.5.3,

$$d(\mathbf{g} \circ \mathbf{f}, \Omega, \mathbf{y}) = d(\widetilde{\mathbf{g}} \circ \mathbf{f}, \Omega, \mathbf{y}) = \sum_{i=1}^{\infty} d(\widetilde{\mathbf{g}}, K_i, \mathbf{y})\, d(\mathbf{f}, \Omega, K_i)$$

$$= \sum_{i=1}^{\infty} d(\mathbf{g}, K_i, \mathbf{y})\, d(\mathbf{f}, \Omega, K_i)$$

and the sum has only finitely many non zero terms. ■

Note there are no convergence problems because these sums are actually finite sums because, as in the previous lemma, $\mathbf{g}^{-1}(\mathbf{y}) \cap \mathbf{f}(\overline{\Omega})$ is a compact set covered by the components of $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$ and so it is covered by finitely many of these components. For the other components, $d(\mathbf{f}, \Omega, K_i) = 0$ or else $d(\mathbf{g}, K_i, \mathbf{y}) = 0$.

The following theorem is the Jordan separation theorem, a major result. A homeomorphism is a function which is one to one onto and continuous having continuous inverse. Before the theorem, here is a helpful lemma.

Lemma 10.5.5 *Let Ω be a bounded open set in $\mathbb{R}^n$, $\mathbf{f} \in C(\overline{\Omega}; \mathbb{R}^n)$, and suppose $\{\Omega_i\}_{i=1}^{\infty}$ are disjoint open sets contained in Ω such that*

$$\mathbf{y} \notin \mathbf{f}\left(\overline{\Omega} \setminus \cup_{j=1}^{\infty} \Omega_j\right)$$

Then

$$d(\mathbf{f}, \Omega, \mathbf{y}) = \sum_{j=1}^{\infty} d(\mathbf{f}, \Omega_j, \mathbf{y})$$

where the sum has all but finitely many terms equal to 0.

Proof: By assumption, the compact set $\mathbf{f}^{-1}(\mathbf{y}) \equiv \left\{\mathbf{x} \in \bar{\Omega} : \mathbf{f}(\mathbf{x}) = \mathbf{y}\right\}$ has empty intersection with

$$\overline{\Omega} \setminus \cup_{j=1}^{\infty} \Omega_j$$

and so this compact set is covered by finitely many of the Ω_j, say $\{\Omega_1, \cdots, \Omega_{n-1}\}$ and

$$\mathbf{y} \notin \mathbf{f}\left(\cup_{j=n}^{\infty} \Omega_j\right).$$

By Theorem 10.2.8 and letting $O = \cup_{j=n}^{\infty} \Omega_j$,

$$d(\mathbf{f}, \Omega, \mathbf{y}) = \sum_{j=1}^{n-1} d(\mathbf{f}, \Omega_j, \mathbf{y}) + d(\mathbf{f}, O, \mathbf{y}) = \sum_{j=1}^{\infty} d(\mathbf{f}, \Omega_j, \mathbf{y})$$

because $d(\mathbf{f}, O, \mathbf{y}) = 0$ as is $d(\mathbf{f}, \Omega_j, \mathbf{y})$ for every $j \geq n$. ■

Lemma 10.5.6 *Define ∂U to be those points $\mathbf{x}$ with the property that for every $r > 0$, $B(\mathbf{x}, r)$ contains points of U and points of U^C. Then for U an open set,*

$$\partial U = \overline{U} \setminus U$$

Let C be a closed subset of $\mathbb{R}^n$ and let $\mathcal{K}$ denote the set of components of $\mathbb{R}^n \setminus C$. Then if K is one of these components, it is open and

$$\partial K \subseteq C$$

Proof: Let $\mathbf{x} \in \overline{U} \setminus U$. If $B(\mathbf{x}, r)$ contains no points of U, then $\mathbf{x} \notin \overline{U}$. If $B(\mathbf{x}, r)$ contains no points of U^C, then $\mathbf{x} \in U$ and so $\mathbf{x} \notin \overline{U} \setminus U$. Therefore, $\overline{U} \setminus U \subseteq \partial U$. Now let $\mathbf{x} \in \partial U$. If $\mathbf{x} \in U$, then since U is open there is a ball containing $\mathbf{x}$ which is contained in U contrary to $\mathbf{x} \in \partial U$. Therefore, $\mathbf{x} \notin U$. If $\mathbf{x}$ is not a limit point of U, then some ball containing $\mathbf{x}$ contains no points of U contrary to $\mathbf{x} \in \partial U$. Therefore, $\mathbf{x} \in \overline{U} \setminus U$ which shows the two sets are equal.

Why is K open for K a component of $\mathbb{R}^n \setminus C$? This is obvious because in $\mathbb{R}^n$ an open ball is connected. Thus if $k \in K$,letting $B(k, r) \subseteq C^C$, it follows $K \cup B(k, r)$ is connected and contained in C^C. Thus $K \cup B(k, r)$ is connected, contained in C^C, and therefore is contained in K because K is maximal with respect to being connected and contained in C^C.

Now for K a component of $\mathbb{R}^n \setminus C$, why is $\partial K \subseteq C$? Let $\mathbf{x} \in \partial K$. If $\mathbf{x} \notin C$, then $\mathbf{x} \in K_1$, some component of $\mathbb{R}^n \setminus C$. If $K_1 \neq K$ then $\mathbf{x}$ cannot be a limit point of K and so it cannot be in ∂K. Therefore, $K = K_1$ but this also is a contradiction because if $\mathbf{x} \in \partial K$ then $\mathbf{x} \notin K$. ■

I will give a shorter version of the proof and a longer version. First is the shorter version which leaves out a few details which may or may not be clear. Sometimes, it seems to me that when you put in too many details, one loses the forest by stumbling around hitting trees. It may still have too many details.

Theorem 10.5.7 *(Jordan separation theorem) Let $\mathbf{f}$ be a homeomorphism of C and $\mathbf{f}(C)$ where C is a compact set in $\mathbb{R}^n$. Then $\mathbb{R}^n \setminus C$ and $\mathbb{R}^n \setminus \mathbf{f}(C)$ have the same number of connected components.*

Proof: Denote by $\mathcal{K}$ the bounded components of $\mathbb{R}^n \setminus C$ and denote by $\mathcal{L}$, the bounded components of $\mathbb{R}^n \setminus \mathbf{f}(C)$. Also, using the Tietze extension theorem, there exists $\overline{\mathbf{f}}$ an extension of $\mathbf{f}$ to all of $\mathbb{R}^n$ which maps into a bounded set and let $\overline{\mathbf{f}^{-1}}$ be an extension of $\mathbf{f}^{-1}$ to all of $\mathbb{R}^n$ which also maps into a bounded set. Pick $K \in \mathcal{K}$ and take $\mathbf{y} \in K$. Then $\partial K \subseteq C$ and so

$$\mathbf{y} \notin \overline{\mathbf{f}^{-1}}\left(\overline{\mathbf{f}}(\partial K)\right)$$

Since $\overline{\mathbf{f}^{-1}} \circ \overline{\mathbf{f}}$ equals the identity I on ∂K, it follows from the properties of the degree that

$$1 = d(I, K, \mathbf{y}) = d\left(\overline{\mathbf{f}^{-1}} \circ \overline{\mathbf{f}}, K, \mathbf{y}\right).$$

Recall that if two functions agree on the boundary, then they have the same degree. Let $\mathcal{H}$ denote the set of bounded components of $\mathbb{R}^n \setminus \mathbf{f}(\partial K)$. These will be as large as those in $\mathcal{L}$ and if a set in $\mathcal{L}$ intersects one of these larger $H \in \mathcal{H}$ then H contains the component in $\mathcal{L}$. By the product formula,

$$1 = d\left(\overline{\mathbf{f}^{-1}} \circ \overline{\mathbf{f}}, K, \mathbf{y}\right) = \sum_{H \in \mathcal{H}} d\left(\overline{\mathbf{f}}, K, H\right) d\left(\overline{\mathbf{f}^{-1}}, H, \mathbf{y}\right), \tag{10.18}$$

the sum being a finite sum from the product formula. That is, there are finitely many H involved in the sum, the other terms being zero.

What about those sets of $\mathcal{H}$ which contain no set of $\mathcal{L}$? These sets also have empty intersection with all sets of $\mathcal{L}$. Therefore, for H one of these, $H \subseteq \mathbf{f}(C)$. Therefore,

$$d\left(\overline{\mathbf{f}^{-1}}, H, \mathbf{y}\right) = d\left(\mathbf{f}^{-1}, H, \mathbf{y}\right) = 0$$

because $\mathbf{y} \in K$ a component of $\mathbb{R}^n \setminus C$, but for $\mathbf{u} \in H \subseteq \mathbf{f}(C), \mathbf{f}^{-1}(\mathbf{u}) \in C$ so $\mathbf{f}^{-1}(\mathbf{u}) \neq \mathbf{y}$ implying that $d\left(\mathbf{f}^{-1}, H, \mathbf{y}\right) = 0$. Thus in 10.18, all such terms are zero. Then letting $\mathcal{H}_1$ be those sets of $\mathcal{H}$ which contain (intersect) some sets of $\mathcal{L}$, the above sum reduces to

$$\begin{aligned}\sum_{H\in\mathcal{H}_1} d\left(\overline{\mathbf{f}}, K, H\right) d\left(\overline{\mathbf{f}^{-1}}, H, \mathbf{y}\right) &= \sum_{H\in\mathcal{H}_1} d\left(\overline{\mathbf{f}}, K, H\right) \sum_{L\in\mathcal{L}_H} d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right) \\ &= \sum_{H\in\mathcal{H}_1} \sum_{L\in\mathcal{L}_H} d\left(\overline{\mathbf{f}}, K, H\right) d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right)\end{aligned}$$

where $\mathcal{L}_H$ are those sets of $\mathcal{L}$ contained in H. If $\mathcal{L}_H = \emptyset$, the above shows that the second sum is 0 with the convention that $\sum_\emptyset = 0$. Now $d\left(\overline{\mathbf{f}}, K, H\right) = d\left(\overline{\mathbf{f}}, K, L\right)$ where $L \in \mathcal{L}_H$. Therefore,

$$\sum_{H\in\mathcal{H}_1} \sum_{L\in\mathcal{L}_H} d\left(\overline{\mathbf{f}}, K, H\right) d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right) = \sum_{H\in\mathcal{H}_1} \sum_{L\in\mathcal{L}_H} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right)$$

As noted above, there are finitely many $H \in \mathcal{H}$ which are involved. $\mathbb{R}^n \setminus \mathbf{f}(C) \subseteq \mathbb{R}^n \setminus \mathbf{f}(\partial K)$ and so every L must be contained in some $H \in \mathcal{H}_1$. It follows that the above reduces to

$$\sum_{L\in\mathcal{L}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right)$$

Thus from 10.18,

$$1 = \sum_{L\in\mathcal{L}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right) = \sum_{L\in\mathcal{L}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, K\right) \tag{10.19}$$

Let $|\mathcal{K}|$ denote the number of components in $\mathcal{K}$ and similarly, $|\mathcal{L}|$ denotes the number of components in $\mathcal{L}$. Thus

$$|\mathcal{K}| = \sum_{K\in\mathcal{K}} 1 = \sum_{K\in\mathcal{K}} \sum_{L\in\mathcal{L}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, K\right)$$

Similarly,

$$|\mathcal{L}| = \sum_{L\in\mathcal{L}} 1 = \sum_{L\in\mathcal{L}} \sum_{K\in\mathcal{K}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, K\right)$$

If $|\mathcal{K}| < \infty$, then $\sum_{K\in\mathcal{K}} \overbrace{\sum_{L\in\mathcal{L}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, K\right)}^{1} < \infty$. The summation which equals 1 is a finite sum and so is the outside sum. Hence we can switch the order of summation and get

$$|\mathcal{K}| = \sum_{L\in\mathcal{L}} \sum_{K\in\mathcal{K}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, K\right) = |\mathcal{L}|$$

A similar argument applies if $|\mathcal{L}| < \infty$. Thus if one of these numbers is finite, so is the other and they are equal. It follows that $|\mathcal{L}| = |\mathcal{K}|$.■

Now is the same proof with more details included.

Theorem 10.5.8 *(Jordan separation theorem) Let* $\mathbf{f}$ *be a homeomorphism of* C *and* $\mathbf{f}(C)$ *where* C *is a compact set in* $\mathbb{R}^n$*. Then* $\mathbb{R}^n \setminus C$ *and* $\mathbb{R}^n \setminus \mathbf{f}(C)$ *have the same number of connected components.*

Proof: Denote by $\mathcal{K}$ the bounded components of $\mathbb{R}^n \setminus C$ and denote by $\mathcal{L}$, the bounded components of $\mathbb{R}^n \setminus \mathbf{f}(C)$. Also, using the Tietze extension theorem, there exists $\overline{\mathbf{f}}$ an extension of $\mathbf{f}$ to all of $\mathbb{R}^n$ which maps into a bounded set and let $\overline{\mathbf{f}^{-1}}$ be an extension of $\mathbf{f}^{-1}$ to all of $\mathbb{R}^n$ which also maps into a bounded set. Pick $K \in \mathcal{K}$ and take $\mathbf{y} \in K$. Then

$$\mathbf{y} \notin \overline{\mathbf{f}^{-1}}\left(\overline{\mathbf{f}}(\partial K)\right)$$

because by Lemma 10.5.6, $\partial K \subseteq C$ and on $C, \overline{\mathbf{f}} = \mathbf{f}$. Thus the right side is of the form

$$\overline{\mathbf{f}^{-1}}\left(\overbrace{\mathbf{f}(\partial K)}^{\subseteq \mathbf{f}(C)}\right) = \mathbf{f}^{-1}(\mathbf{f}(\partial K)) \subseteq C$$

and $\mathbf{y} \notin C$. Since $\overline{\mathbf{f}^{-1}} \circ \overline{\mathbf{f}}$ equals the identity I on ∂K, it follows from the properties of the degree that

$$1 = d(I, K, \mathbf{y}) = d\left(\overline{\mathbf{f}^{-1}} \circ \overline{\mathbf{f}}, K, \mathbf{y}\right).$$

Recall that if two functions agree on the boundary, then they have the same degree. Let $\mathcal{H}$ denote the set of bounded components of $\mathbb{R}^n \setminus \mathbf{f}(\partial K)$. (These will be as large as those in $\mathcal{L}$) By the product formula,

$$1 = d\left(\overline{\mathbf{f}^{-1}} \circ \overline{\mathbf{f}}, K, \mathbf{y}\right) = \sum_{H \in \mathcal{H}} d\left(\overline{\mathbf{f}}, K, H\right) d\left(\overline{\mathbf{f}^{-1}}, H, \mathbf{y}\right), \tag{10.20}$$

the sum being a finite sum from the product formula. It might help to consult the following diagram.

$$\begin{array}{lcl} \mathbb{R}^n \setminus C & \overset{\overline{\mathbf{f}}}{\underset{\overline{\mathbf{f}^{-1}}}{\rightleftarrows}} & \mathbb{R}^n \setminus \mathbf{f}(C) \\ \mathcal{K} & & \mathcal{L} \\ K & & \mathbb{R}^n \setminus \mathbf{f}(\partial K) \\ \mathbf{y} \in K & & \mathcal{H}, \mathcal{H}_1 \\ & & H \\ & & \mathcal{L}_H \end{array}$$

Now letting $\mathbf{x} \in L \in \mathcal{L}$, if S is a connected set containing $\mathbf{x}$ and contained in $\mathbb{R}^n \setminus \mathbf{f}(C)$, then it follows S is contained in $\mathbb{R}^n \setminus \mathbf{f}(\partial K)$ because $\partial K \subseteq C$. Therefore, every set of $\mathcal{L}$ is contained in some set of $\mathcal{H}$. Furthermore, if any $L \in \mathcal{L}$ has nonempty intersection with $H \in \mathcal{H}$ then it must be contained in H. This is because

$$L = (L \cap H) \cup (L \cap \partial H) \cup \left(L \cap \overline{H}^C\right).$$

Now by Lemma 10.5.6,

$$L \cap \partial H \subseteq L \cap \mathbf{f}(\partial K) \subseteq L \cap \mathbf{f}(C) = \emptyset.$$

Since L is connected, $L \cap \overline{H}^C = \emptyset$. Letting $\mathcal{L}_H$ denote those sets of $\mathcal{L}$ which are contained in H equivalently having nonempty intersection with H, if $\mathbf{p} \in H \setminus \cup\mathcal{L}_H = H \setminus \cup\mathcal{L}$, then $\mathbf{p} \in H \cap \mathbf{f}(C)$ and so

$$H = (\cup\mathcal{L}_H) \cup (H \cap \mathbf{f}(C)) \tag{10.21}$$

Claim 1:

$$\overline{H} \setminus \cup \mathcal{L}_H \subseteq \mathbf{f}(C).$$

Proof of the claim: Suppose $\mathbf{p} \in \overline{H} \setminus \cup \mathcal{L}_H$ but $\mathbf{p} \notin \mathbf{f}(C)$. Then $\mathbf{p} \in L \in \mathcal{L}$. It must be the case that L has nonempty intersection with H since otherwise $\mathbf{p}$ could not be in $\overline{H}$. However, as shown above, this requires $L \subseteq H$ and now by 10.21 and $\mathbf{p} \notin \cup \mathcal{L}_H$, it follows $\mathbf{p} \in \mathbf{f}(C)$ after all. This proves the claim.

Claim 2: $\mathbf{y} \notin \overline{\mathbf{f}^{-1}}\left(\overline{H} \setminus \cup \mathcal{L}_H\right)$. Recall $\mathbf{y} \in K \in \mathcal{K}$ the bounded components of $\mathbb{R}^n \setminus C$.

Proof of the claim: If not, then $\overline{\mathbf{f}^{-1}}(\mathbf{z}) = \mathbf{y}$ where $\mathbf{z} \in \overline{H} \setminus \cup \mathcal{L}_H \subseteq \mathbf{f}(C)$ and so $\mathbf{z} = \mathbf{f}(\mathbf{w})$ for some $\mathbf{w} \in C$ and so $\mathbf{y} = \mathbf{f}^{-1}(\mathbf{f}(\mathbf{w})) = \mathbf{w} \in C$ contrary to $\mathbf{y} \in K$, a component of $\mathbb{R}^n \setminus C$.

Now every set of $\mathcal{L}$ is contained in some set of $\mathcal{H}$. What about those sets of $\mathcal{H}$ which contain no set of $\mathcal{L}$ so that $\mathcal{L}_H = \emptyset$? From 10.21 it follows $H \subseteq \mathbf{f}(C)$. Therefore,

$$d\left(\overline{\mathbf{f}^{-1}}, H, \mathbf{y}\right) = d\left(\mathbf{f}^{-1}, H, \mathbf{y}\right) = 0$$

because $\mathbf{y} \in K$ a component of $\mathbb{R}^n \setminus C$. Therefore, letting $\mathcal{H}_1$ denote those sets of $\mathcal{H}$ which contain some set of $\mathcal{L}$, 10.20 is of the form

$$1 = \sum_{H \in \mathcal{H}_1} d\left(\overline{\mathbf{f}}, K, H\right) d\left(\overline{\mathbf{f}^{-1}}, H, \mathbf{y}\right).$$

and it is still a finite sum because the terms in the sum are 0 for all but finitely many $H \in \mathcal{H}_1$. I want to expand $d\left(\overline{\mathbf{f}^{-1}}, H, \mathbf{y}\right)$ as a sum of the form

$$\sum_{L \in \mathcal{L}_H} d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right)$$

using Lemma 10.5.5. Therefore, I must verify

$$\mathbf{y} \notin \overline{\mathbf{f}^{-1}}\left(\overline{H} \setminus \cup \mathcal{L}_H\right)$$

but this is just **Claim 2**. By Lemma 10.5.5, I can write the above sum in place of $d\left(\overline{\mathbf{f}^{-1}}, H, \mathbf{y}\right)$. Therefore,

$$1 = \sum_{H \in \mathcal{H}_1} d\left(\overline{\mathbf{f}}, K, H\right) d\left(\overline{\mathbf{f}^{-1}}, H, \mathbf{y}\right) = \sum_{H \in \mathcal{H}_1} d\left(\overline{\mathbf{f}}, K, H\right) \sum_{L \in \mathcal{L}_H} d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right)$$

where there are only finitely many H which give a nonzero term and for each of these, there are only finitely many L in $\mathcal{L}_H$ which yield $d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right) \neq 0$. Now the above equals

$$= \sum_{H \in \mathcal{H}_1} \sum_{L \in \mathcal{L}_H} d\left(\overline{\mathbf{f}}, K, H\right) d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right). \tag{10.22}$$

By definition,

$$d\left(\overline{\mathbf{f}}, K, H\right) = d\left(\overline{\mathbf{f}}, K, \mathbf{x}\right)$$

where $\mathbf{x}$ is any point of H. In particular $d\left(\overline{\mathbf{f}}, K, H\right) = d\left(\overline{\mathbf{f}}, K, L\right)$ for any $L \in \mathcal{L}_H$. Therefore, the above reduces to

$$= \sum_{L \in \mathcal{L}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right) \tag{10.23}$$

Here is why. There are finitely many $H \in \mathcal{H}_1$ for which the term in the double sum of 10.22 is not zero, say $H_1, \cdots, H_m$. Then the above sum in 10.23 equals

$$\sum_{k=1}^{m} \sum_{L \in \mathcal{L}_{H_k}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right) + \sum_{\mathcal{L} \setminus \cup_{k=1}^{m} \mathcal{L}_{H_k}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right)$$

The second sum equals 0 because those L are contained in some $H \in \mathcal{H}$ for which

$$\begin{aligned} 0 &= d\left(\overline{\mathbf{f}}, K, H\right) d\left(\overline{\mathbf{f}^{-1}}, H, \mathbf{y}\right) = d\left(\overline{\mathbf{f}}, K, H\right) \sum_{L \in \mathcal{L}_H} d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right) \\ &= \sum_{L \in \mathcal{L}_H} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right). \end{aligned}$$

Therefore, the sum in 10.23 reduces to

$$\sum_{k=1}^{m} \sum_{L \in \mathcal{L}_{H_k}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, \mathbf{y}\right)$$

which is the same as the sum in 10.22. Therefore, 10.23 does follow. Then the sum in 10.23 reduces to

$$= \sum_{L \in \mathcal{L}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, K\right)$$

and all but finitely many terms in the sum are 0.

By the same argument,

$$1 = \sum_{K \in \mathcal{K}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, K\right)$$

and all but finitely many terms in the sum are 0. Letting $|\mathcal{K}|$ denote the number of elements in $\mathcal{K}$, similar for $\mathcal{L}$,

$$\begin{aligned} |\mathcal{K}| &= \sum_{K \in \mathcal{K}} 1 = \sum_{K \in \mathcal{K}} \left(\sum_{L \in \mathcal{L}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, K\right) \right) \\ |\mathcal{L}| &= \sum_{L \in \mathcal{L}} 1 = \sum_{L \in \mathcal{L}} \left(\sum_{K \in \mathcal{K}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, K\right) \right) \end{aligned}$$

Suppose $|\mathcal{K}| < \infty$. Then you can switch the order of summation in the double sum for $|\mathcal{K}|$ and so

$$\begin{aligned} |\mathcal{K}| &= \sum_{K \in \mathcal{K}} \left(\sum_{L \in \mathcal{L}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, K\right) \right) \\ &= \sum_{L \in \mathcal{L}} \left(\sum_{K \in \mathcal{K}} d\left(\overline{\mathbf{f}}, K, L\right) d\left(\overline{\mathbf{f}^{-1}}, L, K\right) \right) = |\mathcal{L}| \end{aligned}$$

It follows that if either $|\mathcal{K}|$ or $|\mathcal{L}|$ is finite, then they are equal. Thus if one is infinite, so is the other. This proves the theorem because if $n > 1$ there is exactly one unbounded component to both $\mathbb{R}^n \setminus C$ and $\mathbb{R}^n \setminus \mathbf{f}(C)$ and if $n = 1$ there are exactly two unbounded components. ■

As an application, here is a very interesting little result. It has to do with $d\left(\mathbf{f}, \Omega, \mathbf{f}(\mathbf{x})\right)$ in the case where $\mathbf{f}$ is one to one and Ω is connected. You might imagine this should equal 1 or -1 based on one dimensional analogies. In fact this is the case and it is a nice application of the Jordan separation theorem and the product formula.

Proposition 10.5.9 *Let Ω be an open connected bounded set in $\mathbb{R}^n, n \geq 1$ such that $\mathbb{R}^n \setminus \partial\Omega$ consists of two, three if $n = 1$, connected components. Let $\mathbf{f} \in C\left(\overline{\Omega}; \mathbb{R}^n\right)$ be continuous and one to one. Then $\mathbf{f}(\Omega)$ is the bounded component of $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$ and for $\mathbf{y} \in \mathbf{f}(\Omega)$, $d(\mathbf{f}, \Omega, \mathbf{y})$ either equals 1 or -1.*

Proof: First suppose $n \geq 2$. By the Jordan separation theorem, $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$ consists of two components, a bounded component B and an unbounded component U. Using the Tietze extention theorem, there exists $\mathbf{g}$ defined on $\mathbb{R}^n$ such that $\mathbf{g} = \mathbf{f}^{-1}$ on $\mathbf{f}\left(\overline{\Omega}\right)$. Thus on $\partial\Omega, \mathbf{g} \circ \mathbf{f} = \mathrm{id}$. It follows from this and the product formula that

$$\begin{aligned} 1 &= d(\mathrm{id}, \Omega, \mathbf{g}(\mathbf{y})) = d(\mathbf{g} \circ \mathbf{f}, \Omega, \mathbf{g}(\mathbf{y})) \\ &= d(\mathbf{g}, B, \mathbf{g}(\mathbf{y}))\, d(\mathbf{f}, \Omega, B) + d(\mathbf{f}, \Omega, U)\, d(\mathbf{g}, U, \mathbf{g}(\mathbf{y})) \\ &= d(\mathbf{g}, B, \mathbf{g}(\mathbf{y}))\, d(\mathbf{f}, \Omega, B) \end{aligned}$$

The reduction happens because $d(\mathbf{f}, \Omega, U) = 0$ as explained above. Since U is unbounded, there are points in U which cannot be in the compact set $\mathbf{f}\left(\overline{\Omega}\right)$. For such, the degree is 0 but the degree is constant on U, one of the components of $\mathbf{f}(\partial\Omega)$. Therefore, $d(\mathbf{f}, \Omega, B) \neq 0$ and so for every $\mathbf{z} \in B$, it follows $\mathbf{z} \in \mathbf{f}(\Omega)$. Thus $B \subseteq \mathbf{f}(\Omega)$. On the other hand, $\mathbf{f}(\Omega)$ cannot have points in both U and B because it is a connected set. Therefore $\mathbf{f}(\Omega) \subseteq B$ and this shows $B = \mathbf{f}(\Omega)$. Thus $d(\mathbf{f}, \Omega, B) = d(\mathbf{f}, \Omega, \mathbf{y})$ for each $\mathbf{y} \in B$ and the above formula shows this equals either 1 or -1 because the degree is an integer. In the case where $n = 1$, the argument is similar but here you have 3 components in $\mathbb{R}^1 \setminus \mathbf{f}(\partial\Omega)$ so there are more terms in the above sum although two of them give 0. ■

10.6 Integration And The Degree

There is a very interesting application of the degree to integration [18]. Recall Lemma 10.2.5. I want to generalize this to the case where $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^n$ is only $C^1(\mathbb{R}^n; \mathbb{R}^n)$, vanishing outside a bounded set. In the following proposition, let ψ_m be a symmetric nonnegative mollifier,

$$\psi_m(\mathbf{x}) \equiv m^n \psi(m\mathbf{x}), \operatorname{spt}\psi \subseteq B(\mathbf{0}, 1)$$

and let ϕ_ε be a mollifier as $\varepsilon \to 0$

$$\phi_\varepsilon(\mathbf{x}) \equiv \left(\frac{1}{\varepsilon}\right)^n \phi\left(\frac{\mathbf{x}}{\varepsilon}\right), \operatorname{spt}\phi \subseteq B(\mathbf{0}, 1)$$

Ω will be a bounded open set.

Proposition 10.6.1 *Let $S \subseteq \mathbf{h}(\partial\Omega)^C$ such that*

$$\operatorname{dist}(S, \mathbf{h}(\partial\Omega)) > 0$$

where Ω is a bounded open set and also let $\mathbf{h}$ be $C^1(\mathbb{R}^n; \mathbb{R}^n)$, vanishing outside some bounded set. Then there exists $\varepsilon_0 > 0$ such that whenever $0 < \varepsilon < \varepsilon_0$

$$d(\mathbf{h}, \Omega, \mathbf{y}) = \int_\Omega \phi_\varepsilon(\mathbf{h}(\mathbf{x}) - \mathbf{y}) \det D\mathbf{h}(\mathbf{x})\, dx$$

for all $\mathbf{y} \in S$.

Proof: Let $\varepsilon_0 > 0$ be small enough that for all $\mathbf{y} \in S$,

$$B(\mathbf{y}, 5\varepsilon_0) \cap \mathbf{h}(\partial\Omega) = \emptyset.$$

Now let ψ_m be a mollifier as $m \to \infty$ with support in $B\left(\mathbf{0}, m^{-1}\right)$ and let

$$\mathbf{h}_m \equiv \mathbf{h} * \psi_m.$$

Thus $\mathbf{h}_m \in C^\infty\left(\overline{\Omega}; \mathbb{R}^n\right)$ and $\mathbf{h}_m$ converges uniformly to $\mathbf{h}$ while $D\mathbf{h}_m$ converges uniformly to $D\mathbf{h}$. Denote by $||\cdot||_\infty$ this norm defined by

$$||\mathbf{h}||_\infty \equiv \sup\{|\mathbf{h}(\mathbf{x})| : \mathbf{x} \in \mathbb{R}^n\}$$

$$||D\mathbf{h}||_\infty \equiv \sup\left\{\max_{i,j}\left|\frac{\partial h_i(\mathbf{x})}{\partial x_j}\right| : \mathbf{x} \in \mathbb{R}^n\right\}$$

It is finite because it is given that $\mathbf{h}$ vanishes off a bounded set. Choose M such that for $m \geq M$,

$$||\mathbf{h}_m - \mathbf{h}||_\infty < \varepsilon_0. \tag{10.24}$$

Thus $\mathbf{h}_m \in \mathcal{U}_\mathbf{y} \cap C^2\left(\overline{\Omega}; \mathbb{R}^n\right)$ for all $\mathbf{y} \in S$ where $\mathcal{U}_\mathbf{y}$ is defined on Page 267 and consists of those functions $\mathbf{f}$ in $C\left(\overline{\Omega}; \mathbb{R}^n\right)$ for which $\mathbf{y} \notin \mathbf{f}(\partial\Omega)$.

For $\mathbf{y} \in S$, let $\mathbf{z} \in B(\mathbf{y}, \varepsilon)$ where $\varepsilon < \varepsilon_0$ and suppose $\mathbf{x} \in \partial\Omega$, and $k, m \geq M$. Then for $t \in [0,1]$,

$$\begin{aligned} |(1-t)\mathbf{h}_m(\mathbf{x}) + \mathbf{h}_k(\mathbf{x})t - \mathbf{z}| &\geq |\mathbf{h}_m(\mathbf{x}) - \mathbf{z}| - t|\mathbf{h}_k(\mathbf{x}) - \mathbf{h}_m(\mathbf{x})| \\ &> 2\varepsilon_0 - t2\varepsilon_0 \geq 0 \end{aligned}$$

showing that for each $\mathbf{y} \in S$, $B(\mathbf{y}, \varepsilon) \cap ((1-t)\mathbf{h}_m + t\mathbf{h}_k)(\partial\Omega) = \emptyset$. By Lemma 10.2.5, for all $\mathbf{y} \in S$,

$$\int_\Omega \phi_\varepsilon(\mathbf{h}_m(\mathbf{x}) - \mathbf{y}) \det(D\mathbf{h}_m(\mathbf{x}))\, dx =$$

$$\int_\Omega \phi_\varepsilon(\mathbf{h}_k(\mathbf{x}) - \mathbf{y}) \det(D\mathbf{h}_k(\mathbf{x}))\, dx \tag{10.25}$$

for all $k, m \geq M$. By this lemma again, which says that for small enough ε the integral is constant and the definition of the degree in Definition 10.2.4,

$$d(\mathbf{y},\Omega,\mathbf{h}_m) = \int_\Omega \phi_\varepsilon(\mathbf{h}_m(\mathbf{x}) - \mathbf{y}) \det(D\mathbf{h}_m(\mathbf{x}))\, dx \tag{10.26}$$

for all ε small enough. For $\mathbf{x} \in \partial\Omega$, $\mathbf{y} \in S$, and $t \in [0,1]$,

$$\begin{aligned} |(1-t)\mathbf{h}(\mathbf{x}) + \mathbf{h}_m(\mathbf{x})t - \mathbf{y}| &\geq |\mathbf{h}(\mathbf{x}) - \mathbf{y}| - t|\mathbf{h}(\mathbf{x}) - \mathbf{h}_m(\mathbf{x})| \\ &> 3\varepsilon_0 - t2\varepsilon_0 > 0 \end{aligned}$$

and so by Theorem 10.2.8, the part about homotopy, for each $\mathbf{y} \in S$,

$$d(\mathbf{y},\Omega,\mathbf{h}) = d(\mathbf{y},\Omega,\mathbf{h}_m) =$$

$$\int_\Omega \phi_\varepsilon(\mathbf{h}_m(\mathbf{x}) - \mathbf{y}) \det(D\mathbf{h}_m(\mathbf{x}))\, dx$$

whenever ε is small enough. Fix such an $\varepsilon < \varepsilon_0$ and use 10.25 to conclude the right side of the above equation is independent of $m > M$. Now from the uniform convergence noted above,

$$\begin{aligned} d(\mathbf{y},\Omega,\mathbf{h}) &= \lim_{m\to\infty} \int_\Omega \phi_\varepsilon(\mathbf{h}_m(\mathbf{x}) - \mathbf{y}) \det(D\mathbf{h}_m(\mathbf{x}))\, dx \\ &= \int_\Omega \phi_\varepsilon(\mathbf{h}(\mathbf{x}) - \mathbf{y}) \det(D\mathbf{h}(\mathbf{x}))\, dx. \end{aligned}$$

This proves the proposition. ■

The next lemma is quite interesting. It says a C^1 function maps sets of measure zero to sets of measure zero. This was proved earlier in Lemma 9.8.1 but I am stating a special case here for convenience.

Lemma 10.6.2 *Let* $\mathbf{h} \in C^1(\mathbb{R}^n;\mathbb{R}^n)$ *and* $\mathbf{h}$ *vanishes off a bounded set. Let* $m_n(A) = 0$. *Then* $\mathbf{h}(A)$ *also has measure zero.*

Next is an interesting change of variables theorem. Let Ω be a bounded open set with the property that $\partial\Omega$ has measure zero and let $\mathbf{h}$ be C^1 and vanish off a bounded set. Then from Lemma 10.6.2, $\mathbf{h}(\partial\Omega)$ also has measure zero.

Now suppose $f \in C_c\left(\mathbf{h}(\partial\Omega)^C\right)$. By compactness, there are finitely many components of $\mathbf{h}(\partial\Omega)^C$ which have nonempty intersection with $\mathrm{spt}(f)$. From the Proposition above,

$$\int f(\mathbf{y})\, d(\mathbf{y},\Omega,\mathbf{h})\, dy = \int f(\mathbf{y}) \lim_{\varepsilon\to 0} \int_\Omega \phi_\varepsilon(\mathbf{h}(\mathbf{x}) - \mathbf{y}) \det D\mathbf{h}(\mathbf{x})\, dx dy$$

Actually, there exists an ε small enough that for all $\mathbf{y} \in \mathrm{spt}(f)$,

$$\begin{aligned} \lim_{\delta\to 0} \int_\Omega \phi_\delta(\mathbf{h}(\mathbf{x}) - \mathbf{y}) \det D\mathbf{h}(\mathbf{x})\, dx &= \int_\Omega \phi_\varepsilon(\mathbf{h}(\mathbf{x}) - \mathbf{y}) \det D\mathbf{h}(\mathbf{x})\, dx \\ &= d(\mathbf{y},\Omega,\mathbf{h}) \end{aligned}$$

This is because $\mathrm{spt}(f)$ is at a positive distance from the compact set $\mathbf{h}(\partial\Omega)^C$ so this follows from Proposition 10.6.1. Therefore, for all ε small enough,

$$\begin{aligned} \int f(\mathbf{y})\, d(\mathbf{y},\Omega,\mathbf{h})\, dy &= \int\int_\Omega f(\mathbf{y})\, \phi_\varepsilon(\mathbf{h}(\mathbf{x}) - \mathbf{y}) \det D\mathbf{h}(\mathbf{x})\, dx dy \\ &= \int_\Omega \det D\mathbf{h}(\mathbf{x}) \int f(\mathbf{y})\, \phi_\varepsilon(\mathbf{h}(\mathbf{x}) - \mathbf{y})\, dy dx \end{aligned}$$

$$= \int_\Omega \det D\mathbf{h}(\mathbf{x})\, f(\mathbf{h}(\mathbf{x}))\, dx + \int_\Omega \det D\mathbf{h}(\mathbf{x}) \int (f(\mathbf{y}) - f(\mathbf{h}(\mathbf{x})))\, \phi_\varepsilon(\mathbf{h}(\mathbf{x}) - \mathbf{y})\, dy dx$$

Using the uniform continuity of f, you can now pass to a limit and obtain

$$\int f(\mathbf{y})\, d(\mathbf{y},\Omega,\mathbf{h})\, dy = \int_\Omega f(\mathbf{h}(\mathbf{x})) \det D\mathbf{h}(\mathbf{x})\, dx$$

This has essentially proved the following interesting Theorem.

Theorem 10.6.3 *Let* $f \in C_c\left(\mathbf{h}(\partial\Omega)^C\right)$ *for* Ω *a bounded open set where* $\mathbf{h}$ *is in* $C^1\left(\overline{\Omega};\mathbb{R}^n\right)$. *Suppose* $\partial\Omega$ *has measure zero. Then*

$$\int f(\mathbf{y})\, d(\mathbf{y},\Omega,\mathbf{h})\, dy = \int_\Omega \det(D\mathbf{h}(\mathbf{x}))\, f(\mathbf{h}(\mathbf{x}))\, dx.$$

Proof: Both sides depend only on $\mathbf{h}$ restricted to $\overline{\Omega}$ and so the above results apply and give the formula of the theorem for any $\overline{\mathbf{h}} \in C^1(\mathbb{R}^n;\mathbb{R}^n)$ which vanishes off a bounded set which coincides with $\mathbf{h}$ on $\overline{\Omega}$. This proves the lemma. ■

Lemma 10.6.4 *If* $\mathbf{h} \in C^1\left(\overline{\Omega},\mathbb{R}^n\right)$ *for* Ω *a bounded connected open set with* $\partial\Omega$ *having measure zero and* $\mathbf{h}$ *is one to one on* $\overline{\Omega}$. *Then for any* E *a Borel set,*

$$\int_{\mathbf{h}(\Omega)} \mathcal{X}_E(\mathbf{y})\, d(\mathbf{y},\Omega,\mathbf{h})\, dy = \int_\Omega \det(D\mathbf{h}(\mathbf{x}))\, \mathcal{X}_E(\mathbf{h}(\mathbf{x}))\, dx$$

Furthermore, off a set of measure zero, $\det(D\mathbf{h}(\mathbf{x}))$ *has constant sign equal to the sign of* $d(\mathbf{y},\Omega,\mathbf{h})$.

Proof: First here is a simple observation about the existence of a sequence of nonnegative continuous functions which increase to $\mathcal{X}_O$ for O open.

Let O be an open set and let H_j denote those points $\mathbf{y}$ of O such that

$$\mathrm{dist}\left(\mathbf{y}, O^C\right) \geq 1/j, j = 1, 2, \cdots.$$

Then define $K_j \equiv \overline{B\left(\mathbf{0}, j\right)} \cap H_j$ and

$$W_j \equiv K_j + B\left(\mathbf{0}, \frac{1}{2j}\right).$$

where this means

$$\left\{\mathbf{k} + \mathbf{b} : \mathbf{k} \in K_j, \mathbf{b} \in B\left(\mathbf{0}, \frac{1}{2j}\right)\right\}.$$

Let

$$f_j\left(\mathbf{y}\right) \equiv \frac{\mathrm{dist}\left(\mathbf{y}, W_j^C\right)}{\mathrm{dist}\left(\mathbf{y}, K_j\right) + \mathrm{dist}\left(\mathbf{y}, W_j^C\right)}$$

Thus f_j is nonnegative, increasing in j, has compact support in $\overline{W_j}$, is continuous, and eventually $f_j\left(\mathbf{y}\right) = 1$ for all $\mathbf{y} \in O$ and $f_j\left(\mathbf{y}\right) = 0$ for all $\mathbf{y} \notin O$. Thus $\lim_{j\to\infty} f_j\left(\mathbf{y}\right) = \mathcal{X}_O\left(\mathbf{y}\right)$.

Now let $O \subseteq \mathbf{h}\left(\partial\Omega\right)^C$. Then from the above, let f_j be as described above for the open set $O \cap \mathbf{h}\left(\Omega\right)$. (By invariance of domain, $\mathbf{h}\left(\Omega\right)$ is open.)

$$\int_{\mathbf{h}(\Omega)} f_j\left(\mathbf{y}\right) d\left(\mathbf{y}, \Omega, \mathbf{h}\right) dy = \int_\Omega \det\left(D\mathbf{h}\left(\mathbf{x}\right)\right) f_j\left(\mathbf{h}\left(\mathbf{x}\right)\right) dx$$

From Proposition 10.5.9, $d\left(\mathbf{y}, \Omega, \mathbf{h}\right)$ either equals 1 or -1 for all $\mathbf{y} \in \mathbf{h}\left(\Omega\right)$. Then by the monotone convergence theorem on the left, using the fact that $d\left(\mathbf{y}, \Omega, \mathbf{h}\right)$ is either always 1 or always -1 and the dominated convergence theorem on the right, it follows

$$\begin{aligned}\int_{\mathbf{h}(\Omega)} \mathcal{X}_O\left(\mathbf{y}\right) d\left(\mathbf{y}, \Omega, \mathbf{h}\right) dy &= \int_\Omega \det\left(D\mathbf{h}\left(\mathbf{x}\right)\right) \mathcal{X}_{O\cap\mathbf{h}(\Omega)}\left(\mathbf{h}\left(\mathbf{x}\right)\right) dx \\ &= \int_\Omega \det\left(D\mathbf{h}\left(\mathbf{x}\right)\right) \mathcal{X}_O\left(\mathbf{h}\left(\mathbf{x}\right)\right) dx\end{aligned}$$

If $\mathbf{y} \in \mathbf{h}\left(\partial\Omega\right)$, then since $\mathbf{h}\left(\Omega\right) \cap \mathbf{h}\left(\partial\Omega\right)$ is assumed to be empty, it follows $\mathbf{y} \notin \mathbf{h}\left(\Omega\right)$. Therefore, the above formula holds for any O open.

Now let $\mathcal{G}$ denote those Borel sets of $\mathbb{R}^n$, E such that

$$\int_{\mathbf{h}(\Omega)} \mathcal{X}_E\left(\mathbf{y}\right) d\left(\mathbf{y}, \Omega, \mathbf{h}\right) dy = \int_\Omega \det\left(D\mathbf{h}\left(\mathbf{x}\right)\right) \mathcal{X}_E\left(\mathbf{h}\left(\mathbf{x}\right)\right) dx$$

Then as shown above, $\mathcal{G}$ contains the π system of open sets. Since $\left|\det\left(D\mathbf{h}\left(\mathbf{x}\right)\right)\right|$ is bounded uniformly, it follows easily that if $E \in \mathcal{G}$ then $E^C \in \mathcal{G}$. This is because, since $\mathbb{R}^n$ is an open set,

$$\int_{\mathbf{h}(\Omega)} \mathcal{X}_E\left(\mathbf{y}\right) d\left(\mathbf{y}, \Omega, \mathbf{h}\right) dy + \int_{\mathbf{h}(\Omega)} \mathcal{X}_{E^C}\left(\mathbf{y}\right) d\left(\mathbf{y}, \Omega, \mathbf{h}\right) dy$$

$$\begin{aligned}&= \int_{\mathbf{h}(\Omega)} d\left(\mathbf{y}, \Omega, \mathbf{h}\right) dy = \int_\Omega \det\left(D\mathbf{h}\left(\mathbf{x}\right)\right) dx \\ &= \int_\Omega \det\left(D\mathbf{h}\left(\mathbf{x}\right)\right) \mathcal{X}_{E^C}\left(\mathbf{h}\left(\mathbf{x}\right)\right) dx + \int_\Omega \det\left(D\mathbf{h}\left(\mathbf{x}\right)\right) \mathcal{X}_E\left(\mathbf{h}\left(\mathbf{x}\right)\right) dx\end{aligned}$$

Now cancelling

$$\int_{\Omega} \det\left(D\mathbf{h}\left(\mathbf{x}\right)\right) \mathcal{X}_{E}\left(\mathbf{h}\left(\mathbf{x}\right)\right)$$

from the right with

$$\int_{\mathbf{h}(\Omega)} \mathcal{X}_{E}\left(\mathbf{y}\right) d\left(\mathbf{y},\Omega,\mathbf{h}\right) dy$$

from the left, since $E \in \mathcal{G}$, yields the desired result.

In addition, if $\{E_i\}_{i=1}^{\infty}$ is a sequence of disjoint sets of $\mathcal{G}$, the monotone convergence and dominated convergence theorems imply the union of these disjoint sets is in $\mathcal{G}$. By the Lemma on π systems, Lemma 9.1.2 it follows $\mathcal{G}$ equals the Borel sets.

Now consider the last claim. Suppose $d\left(\mathbf{y},\Omega,\mathbf{h}\right) = -1$. Consider the compact set

$$E_{\varepsilon} \equiv \left\{\mathbf{x} \in \Omega : \det\left(D\mathbf{h}\left(\mathbf{x}\right)\right) \geq \varepsilon\right\}$$

and $f\left(\mathbf{y}\right) \equiv \mathcal{X}_{\mathbf{h}(E_{\varepsilon})}\left(\mathbf{y}\right)$. Then from the first part,

$$0 \geq \int_{\mathbf{h}(\Omega)} \mathcal{X}_{\mathbf{h}(E_{\varepsilon})}\left(\mathbf{y}\right)(-1)\, dy = \int_{\Omega} \det D\mathbf{h}\left(\mathbf{x}\right) \mathcal{X}_{E_{\varepsilon}}\left(\mathbf{x}\right) dx$$

$$\geq \varepsilon m_n\left(E_{\varepsilon}\right)$$

and so $m_n\left(E_{\varepsilon}\right) = 0$. Therefore, if E is the set of $\mathbf{x} \in \Omega$ where $\det\left(D\mathbf{h}\left(\mathbf{x}\right)\right) > 0$, it equals $\cup_{k=1}^{\infty} E_{1/k}$, a set of measure zero. Thus off this set $\det D\mathbf{h}\left(\mathbf{x}\right) \leq 0$. Similarly, if $d\left(\mathbf{y},\Omega,\mathbf{h}\right) = 1$, $\det D\mathbf{h}\left(\mathbf{x}\right) \geq 0$ off a set of measure zero. This proves the lemma. ■

Theorem 10.6.5 *Let $f \geq 0$ and measurable for Ω a bounded open connected set where $\mathbf{h}$ is in $C^1\left(\overline{\Omega};\mathbb{R}^n\right)$ and is one to one on $\overline{\Omega}$. Then*

$$\int_{\mathbf{h}(\Omega)} f\left(\mathbf{y}\right) d\left(\mathbf{y},\Omega,\mathbf{h}\right) dy = \int_{\Omega} \det\left(D\mathbf{h}\left(\mathbf{x}\right)\right) f\left(\mathbf{h}\left(\mathbf{x}\right)\right) dx.$$

which amounts to the same thing as

$$\int_{\mathbf{h}(\Omega)} f\left(\mathbf{y}\right) dy = \int_{\Omega} \left|\det\left(D\mathbf{h}\left(\mathbf{x}\right)\right)\right| f\left(\mathbf{h}\left(\mathbf{x}\right)\right) dx.$$

Proof: First suppose $\partial\Omega$ has measure zero. From Proposition 10.5.9, $d\left(\mathbf{y},\Omega,\mathbf{h}\right)$ either equals 1 or -1 for all $\mathbf{y} \in \mathbf{h}\left(\Omega\right)$ and $\det\left(D\mathbf{h}\left(\mathbf{x}\right)\right)$ has the same sign as the sign of the degree a.e. Suppose $d\left(\mathbf{y},\Omega,\mathbf{h}\right) = -1$. By Theorem 9.9.10, if $f \geq 0$ and is Lebesgue measurable,

$$\int_{\mathbf{h}(\Omega)} f\left(\mathbf{y}\right) dy = \int_{\Omega} \left|\det\left(D\mathbf{h}\left(\mathbf{x}\right)\right)\right| f\left(\mathbf{h}\left(\mathbf{x}\right)\right) dx = \int_{\Omega} (-1) \det\left(D\mathbf{h}\left(\mathbf{x}\right)\right) f\left(\mathbf{h}\left(\mathbf{x}\right)\right) dx$$

and so multiplying both sides by -1,

$$\int_{\Omega} \det\left(D\mathbf{h}\left(\mathbf{x}\right)\right) f\left(\mathbf{h}\left(\mathbf{x}\right)\right) dx = \int_{\mathbf{h}(\Omega)} f\left(\mathbf{y}\right)(-1)\, dy = \int_{\mathbf{h}(\Omega)} f\left(\mathbf{y}\right) d\left(\mathbf{y},\Omega,\mathbf{h}\right) dy$$

The case where $d\left(\mathbf{y},\Omega,\mathbf{h}\right) = 1$ is similar. This proves the theorem when $\partial\Omega$ has measure zero.

Next I show it is not necessary to assume $\partial\Omega$ has measure zero. By Corollary 9.7.6 there exists a sequence of disjoint balls $\{B_i\}$ such that

$$m_n\left(\Omega \setminus \cup_{i=1}^{\infty} B_i\right) = 0.$$

Since ∂B_i has measure zero, the above holds for each B_i and so

$$\int_{\mathbf{h}(B_i)} f(\mathbf{y})\, d(\mathbf{y}, B_i, \mathbf{h})\, dy = \int_{B_i} \det(D\mathbf{h}(\mathbf{x}))\, f(\mathbf{h}(\mathbf{x}))\, dx$$

Since $\mathbf{h}$ is one to one, if $\mathbf{y} \in \mathbf{h}(B_i)$, then $\mathbf{y} \notin \mathbf{h}(\Omega \setminus B_i)$. Therefore, it follows from Theorem 10.2.8 that

$$d(\mathbf{y}, B_i, \mathbf{h}) = d(\mathbf{y}, \Omega, \mathbf{h})$$

Thus

$$\int_{\mathbf{h}(B_i)} f(\mathbf{y})\, d(\mathbf{y}, \Omega, \mathbf{h})\, dy = \int_{B_i} \det(D\mathbf{h}(\mathbf{x}))\, f(\mathbf{h}(\mathbf{x}))\, dx$$

From Lemma 10.6.4 $\det(D\mathbf{h}(\mathbf{x}))$ has the same sign off a set of measure zero as the constant sign of $d(\mathbf{y}, \Omega, \mathbf{h})$. Therefore, using the monotone convergence theorem and that $\mathbf{h}$ is one to one,

$$\begin{aligned}
&\int_{\Omega} \det(D\mathbf{h}(\mathbf{x}))\, f(\mathbf{h}(\mathbf{x}))\, dx = \int_{\cup_{i=1}^{\infty} B_i} \det(D\mathbf{h}(\mathbf{x}))\, f(\mathbf{h}(\mathbf{x}))\, dx \\
&= \sum_{i=1}^{\infty} \int_{B_i} \det(D\mathbf{h}(\mathbf{x}))\, f(\mathbf{h}(\mathbf{x}))\, dx = \sum_{i=1}^{\infty} \int_{\mathbf{h}(B_i)} f(\mathbf{y})\, d(\mathbf{y}, \Omega, \mathbf{h})\, dy \\
&= \int_{\cup_{i=1}^{\infty} \mathbf{h}(B_i)} f(\mathbf{y})\, d(\mathbf{y}, \Omega, \mathbf{h})\, dy = \int_{\mathbf{h}\left(\cup_{i=1}^{\infty} B_i\right)} f(\mathbf{y})\, d(\mathbf{y}, \Omega, \mathbf{h})\, dy \\
&= \int_{\mathbf{h}(\Omega)} f(\mathbf{y})\, d(\mathbf{y}, \Omega, \mathbf{h})\, dy,
\end{aligned}$$

the last line following from Lemma 10.6.2. This proves the theorem. ■

10.7 Exercises

1. Show the Brouwer fixed point theorem is equivalent to the nonexistence of a continuous retraction onto the boundary of $B(\mathbf{0}, r)$.

2. Using the Jordan separation theorem, prove the invariance of domain theorem. **Hint:** You might consider $B(\mathbf{x}, r)$ and show $\mathbf{f}$ maps the inside to one of two components of $\mathbb{R}^n \setminus \mathbf{f}(\partial B(\mathbf{x}, r))$. Thus an open ball goes to some open set.

3. Give a version of Proposition 10.5.9 which is valid for the case where $n = 1$.

4. It was shown that if $\mathbf{f}$ is locally one to one and continuous, $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, and

$$\lim_{|\mathbf{x}| \to \infty} |\mathbf{f}(\mathbf{x})| = \infty,$$

   then $\mathbf{f}$ maps $\mathbb{R}^n$ onto $\mathbb{R}^n$. Suppose you have $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^n$ where $\mathbf{f}$ is one to one and $\lim_{|\mathbf{x}| \to \infty} |\mathbf{f}(\mathbf{x})| = \infty$. Show that $\mathbf{f}$ cannot be onto.

5. Can there exist a one to one onto continuous map, $\mathbf{f}$ which takes the unit interval to the unit disk? **Hint:** Think in terms of invariance of domain and use the hint to Problem **??**.

6. Let $m < n$ and let $B_m(\mathbf{0}, r)$ be the ball in $\mathbb{R}^m$ and $B_n(\mathbf{0}, r)$ be the ball in $\mathbb{R}^n$. Show that there is no one to one continuous map from $\overline{B_m(\mathbf{0}, r)}$ to $\overline{B_n(\mathbf{0}, r)}$. **Hint:** It is like the above problem.

7. Consider the unit disk,

$$\left\{(x,y) : x^2+y^2 \leq 1\right\} \equiv D$$

and the annulus

$$\left\{(x,y) : \frac{1}{2} \leq x^2+y^2 \leq 1\right\} \equiv A$$

Is it possible there exists a one to one onto continuous map $\mathbf{f}$ such that $\mathbf{f}(D) = A$? Thus D has no holes and A is really like D but with one hole punched out. Can you generalize to different numbers of holes? **Hint:** Consider the invariance of domain theorem. The interior of D would need to be mapped to the interior of A. Where do the points of the boundary of A come from? Consider Theorem 5.3.5.

8. Suppose C is a compact set in $\mathbb{R}^n$ which has empty interior and $\mathbf{f} : C \to \Gamma \subseteq \mathbb{R}^n$ is one to one onto and continuous with continuous inverse. Could Γ have nonempty interior? Show also that if $\mathbf{f}$ is one to one and onto Γ then if it is continuous, so is $\mathbf{f}^{-1}$.

9. Let K be a nonempty closed and convex subset of $\mathbb{R}^n$. Recall K is convex means that if $\mathbf{x}, \mathbf{y} \in K$, then for all $t \in [0,1]$, $t\mathbf{x} + (1-t)\mathbf{y} \in K$. Show that if $\mathbf{x} \in \mathbb{R}^n$ there exists a unique $\mathbf{z} \in K$ such that

$$|\mathbf{x}-\mathbf{z}| = \min\{|\mathbf{x}-\mathbf{y}| : \mathbf{y} \in K\}.$$

This $\mathbf{z}$ will be denoted as $P\mathbf{x}$. **Hint:** First note you do not know K is compact. Establish the parallelogram identity if you have not already done so,

$$|\mathbf{u}-\mathbf{v}|^2 + |\mathbf{u}+\mathbf{v}|^2 = 2|\mathbf{u}|^2 + 2|\mathbf{v}|^2.$$

Then let $\{\mathbf{z}_k\}$ be a minimizing sequence,

$$\lim_{k\to\infty} |\mathbf{z}_k - \mathbf{x}|^2 = \inf\{|\mathbf{x}-\mathbf{y}| : \mathbf{y} \in K\} \equiv \lambda.$$

Now using convexity, explain why

$$\left|\frac{\mathbf{z}_k - \mathbf{z}_m}{2}\right|^2 + \left|\mathbf{x} - \frac{\mathbf{z}_k + \mathbf{z}_m}{2}\right|^2 = 2\left|\frac{\mathbf{x}-\mathbf{z}_k}{2}\right|^2 + 2\left|\frac{\mathbf{x}-\mathbf{z}_m}{2}\right|^2$$

and then use this to argue $\{\mathbf{z}_k\}$ is a Cauchy sequence. Then if $\mathbf{z}_i$ works for $i = 1, 2$, consider $(\mathbf{z}_1 + \mathbf{z}_2)/2$ to get a contradiction.

10. In Problem 9 show that $P\mathbf{x}$ satisfies the following variational inequality.

$$(\mathbf{x} - P\mathbf{x}) \cdot (\mathbf{y} - P\mathbf{x}) \leq 0$$

for all $\mathbf{y} \in K$. Then show that $|P\mathbf{x}_1 - P\mathbf{x}_2| \leq |\mathbf{x}_1 - \mathbf{x}_2|$. **Hint:** For the first part note that if $\mathbf{y} \in K$, the function $t \to |\mathbf{x} - (P\mathbf{x} + t(\mathbf{y} - P\mathbf{x}))|^2$ achieves its minimum on $[0,1]$ at $t = 0$. For the second part,

$$(\mathbf{x}_1 - P\mathbf{x}_1) \cdot (P\mathbf{x}_2 - P\mathbf{x}_1) \leq 0,\ (\mathbf{x}_2 - P\mathbf{x}_2) \cdot (P\mathbf{x}_1 - P\mathbf{x}_2) \leq 0.$$

Explain why

$$(\mathbf{x}_2 - P\mathbf{x}_2 - (\mathbf{x}_1 - P\mathbf{x}_1)) \cdot (P\mathbf{x}_2 - P\mathbf{x}_1) \geq 0$$

and then use a some manipulations and the Cauchy Schwarz inequality to get the desired inequality.

11. Establish the Brouwer fixed point theorem for any convex compact set in $\mathbb{R}^n$. **Hint:** If K is a compact and convex set, let R be large enough that the closed ball, $D(\mathbf{0}, R) \supseteq K$. Let P be the projection onto K as in Problem 10 above. If $\mathbf{f}$ is a continuous map from K to K, consider $\mathbf{f} \circ P$. You want to show $\mathbf{f}$ has a fixed point in K.

12. Suppose D is a set which is homeomorphic to $\overline{B(\mathbf{0}, 1)}$. This means there exists a continuous one to one map, $\mathbf{h}$ such that $\mathbf{h}\left(\overline{B(\mathbf{0}, 1)}\right) = D$ such that $\mathbf{h}^{-1}$ is also one to one. Show that if $\mathbf{f}$ is a continuous function which maps D to D then $\mathbf{f}$ has a fixed point. Now show that it suffices to say that $\mathbf{h}$ is one to one and continuous. In this case the continuity of $\mathbf{h}^{-1}$ is automatic. Sets which have the property that continuous functions taking the set to itself have at least one fixed point are said to have the fixed point property. Work Problem 7 using this notion of fixed point property. What about a solid ball and a donut? Could these be homeomorphic?

13. There are many different proofs of the Brouwer fixed point theorem. Let l be a line segment. Label one end with A and the other end B. Now partition the segment into n little pieces and label each of these partition points with either A or B. Show there is an odd number of little segments with one end labeled with A and the other labeled with B. If $\mathbf{f} : l \to l$ is continuous, use the fact it is uniformly continuous and this little labeling result to give a proof for the Brouwer fixed point theorem for a one dimensional segment. Next consider a triangle. Label the vertices with A, B, C and subdivide this triangle into little triangles, $T_1, \cdots, T_m$ in such a way that any pair of these little triangles intersects either along an entire edge or a vertex. Now label the unlabeled vertices of these little triangles with either A, B, or C in any way. Show there is an odd number of little triangles having their vertices labeled as A, B, C. Use this to show the Brouwer fixed point theorem for any triangle. This approach generalizes to higher dimensions and you will see how this would take place if you are successful in going this far. This is an outline of the Sperner's lemma approach to the Brouwer fixed point theorem. Are there other sets besides compact convex sets which have the fixed point property?

14. Using the definition of the derivative and the Vitali covering theorem, show that if $\mathbf{f} \in C^1\left(\overline{U}, \mathbb{R}^n\right)$ and ∂U has n dimensional measure zero then $\mathbf{f}(\partial U)$ also has measure zero. (This problem has little to do with this chapter. It is a review.)

15. Suppose Ω is any open bounded subset of $\mathbb{R}^n$ which contains $\mathbf{0}$ and that $\mathbf{f} : \overline{\Omega} \to \mathbb{R}^n$ is continuous with the property that

$$\mathbf{f}(\mathbf{x}) \cdot \mathbf{x} \geq 0$$

for all $\mathbf{x} \in \partial\Omega$. Show that then there exists $\mathbf{x} \in \Omega$ such that $\mathbf{f}(\mathbf{x}) = \mathbf{0}$. Give a similar result in the case where the above inequality is replaced with $\leq$. **Hint:** You might consider the function

$$\mathbf{h}(t, \mathbf{x}) \equiv t\mathbf{f}(\mathbf{x}) + (1 - t)\mathbf{x}.$$

16. Suppose Ω is an open set in $\mathbb{R}^n$ containing $\mathbf{0}$ and suppose that $\mathbf{f} : \overline{\Omega} \to \mathbb{R}^n$ is continuous and $|\mathbf{f}(\mathbf{x})| \leq |\mathbf{x}|$ for all $\mathbf{x} \in \partial\Omega$. Show $\mathbf{f}$ has a fixed point in $\overline{\Omega}$. **Hint:** Consider $\mathbf{h}(t, \mathbf{x}) \equiv t(\mathbf{x} - \mathbf{f}(\mathbf{x})) + (1 - t)\mathbf{x}$ for $t \in [0, 1]$. If $t = 1$ and some $\mathbf{x} \in \partial\Omega$ is sent to $\mathbf{0}$, then you are done. Suppose therefore, that no fixed point exists on $\partial\Omega$. Consider $t < 1$ and use the given inequality.

17. Let Ω be an open bounded subset of $\mathbb{R}^n$ and let $\mathbf{f}, \mathbf{g} : \overline{\Omega} \to \mathbb{R}^n$ both be continuous such that

$$|\mathbf{f}(\mathbf{x})| - |\mathbf{g}(\mathbf{x})| > 0$$

for all $\mathbf{x} \in \partial\Omega$. Show that then

$$d(\mathbf{f}-\mathbf{g},\Omega,\mathbf{0}) = d(\mathbf{f},\Omega,\mathbf{0})$$

Show that if there exists $\mathbf{x} \in \mathbf{f}^{-1}(\mathbf{0})$, then there exists $\mathbf{x} \in (\mathbf{f}-\mathbf{g})^{-1}(\mathbf{0})$. **Hint:** You might consider $\mathbf{h}(t,\mathbf{x}) \equiv (1-t)\mathbf{f}(\mathbf{x}) + t(\mathbf{f}(\mathbf{x}) - \mathbf{g}(\mathbf{x}))$ and argue $\mathbf{0} \notin \mathbf{h}(t,\partial\Omega)$ for $t \in [0,1]$.

18. Let $f : \mathbb{C} \to \mathbb{C}$ where $\mathbb{C}$ is the field of complex numbers. Thus f has a real and imaginary part. Letting $z = x + iy$,

$$f(z) = u(x,y) + iv(x,y)$$

Recall that the norm in $\mathbb{C}$ is given by $|x+iy| = \sqrt{x^2+y^2}$ and this is the usual norm in $\mathbb{R}^2$ for the ordered pair (x,y). Thus complex valued functions defined on $\mathbb{C}$ can be considered as $\mathbb{R}^2$ valued functions defined on some subset of $\mathbb{R}^2$. Such a complex function is said to be analytic if the usual definition holds. That is

$$f'(z) = \lim_{h\to 0} \frac{f(z+h) - f(z)}{h}.$$

In other words,

$$f(z+h) = f(z) + f'(z)h + o(h) \tag{10.27}$$

at a point z where the derivative exists. Let $f(z) = z^n$ where n is a positive integer. Thus $z^n = p(x,y) + iq(x,y)$ for p, q suitable polynomials in x and y. Show this function is analytic. Next show that for an analytic function and u and v the real and imaginary parts, the Cauchy Riemann equations hold.

$$u_x = v_y,\ u_y = -v_x.$$

In terms of mappings show 10.27 has the form

$$\begin{pmatrix} u(x+h_1, y+h_2) \\ v(x+h_1, y+h_2) \end{pmatrix}$$

$$= \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix} + \begin{pmatrix} u_x(x,y) & u_y(x,y) \\ v_x(x,y) & v_y(x,y) \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + \mathbf{o}(\mathbf{h})$$

$$= \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix} + \begin{pmatrix} u_x(x,y) & -v_x(x,y) \\ v_x(x,y) & u_x(x,y) \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + \mathbf{o}(\mathbf{h})$$

where $\mathbf{h} = (h_1, h_2)^T$ and h is given by $h_1 + ih_2$. Thus the determinant of the above matrix is always nonnegative. Letting B_r denote the ball $B(0,r) = B((0,0),r)$ show

$$d(f, B_r, \mathbf{0}) = n.$$

where $f(z) = z^n$. In terms of mappings on $\mathbb{R}^2$,

$$\mathbf{f}(x,y) = \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix}.$$

Thus show

$$d(\mathbf{f}, B_r, \mathbf{0}) = n.$$

Hint: You might consider

$$g(z) \equiv \prod_{j=1}^{n} (z - a_j)$$

where the a_j are small real distinct numbers and argue that both this function and f are analytic but that $\mathbf{0}$ is a regular value for $\mathbf{g}$ although it is not so for $\mathbf{f}$. However, for each a_j small but distinct $d(\mathbf{f}, B_r, \mathbf{0}) = d(\mathbf{g}, B_r, \mathbf{0})$.

19. Using Problem 18, prove the fundamental theorem of algebra as follows. Let $p(z)$ be a nonconstant polynomial of degree n,

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots$$

Show that for large enough $r, |p(z)| > |p(z) - a_n z^n|$ for all $z \in \partial B(0,r)$. Now from Problem 17 you can conclude $d(p, B_r, 0) = d(f, B_r, 0) = n$ where $f(z) = a_n z^n$.

20. Generalize Theorem 10.6.5 to the situation where Ω is not necessarily a connected open set. You may need to make some adjustments on the hypotheses.

21. Suppose $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ satisfies

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| \geq \alpha |\mathbf{x} - \mathbf{y}|, \ \alpha > 0,$$

Show that $\mathbf{f}$ must map $\mathbb{R}^n$ onto $\mathbb{R}^n$. **Hint:** First show $\mathbf{f}$ is one to one. Then use invariance of domain. Next show, using the inequality, that the points not in $\mathbf{f}(\mathbb{R}^n)$ must form an open set because if $\mathbf{y}$ is such a point, then there can be no sequence $\{\mathbf{f}(\mathbf{x}_n)\}$ converging to it. Finally recall that $\mathbb{R}^n$ is connected.

Chapter 11

Integration Of Differential Forms

11.1 Manifolds

Manifolds are sets which resemble $\mathbb{R}^n$ locally. A manifold with boundary resembles half of $\mathbb{R}^n$ locally. To make this concept of a manifold more precise, here is a definition.

Definition 11.1.1 *Let $\Omega \subseteq \mathbb{R}^m$. A set, U, is open in Ω if it is the intersection of an open set from $\mathbb{R}^m$ with Ω. Equivalently, a set, U is open in Ω if for every point, $\mathbf{x} \in U$, there exists $\delta > 0$ such that if $|\mathbf{x} - \mathbf{y}| < \delta$ and $\mathbf{y} \in \Omega$, then $\mathbf{y} \in U$. A set, H, is closed in Ω if it is the intersection of a closed set from $\mathbb{R}^m$ with Ω. Equivalently, a set, H, is closed in Ω if whenever, $\mathbf{y}$ is a limit point of H and $\mathbf{y} \in \Omega$, it follows $\mathbf{y} \in H$.*

Recall the following definition.

Definition 11.1.2 *Let $V \subseteq \mathbb{R}^n$. $C^k\left(\overline{V}; \mathbb{R}^m\right)$ is the set of functions which are restrictions to V of some function defined on $\mathbb{R}^n$ which has k continuous derivatives and compact support which has values in $\mathbb{R}^m$. When $k = 0$, it means the restriction to V of continuous functions with compact support.*

Definition 11.1.3 *A closed and bounded subset of $\mathbb{R}^m$ Ω, will be called an n dimensional manifold with boundary, $n \geq 1$, if there are finitely many sets U_i, open in Ω and continuous one to one on $\overline{U_i}$ functions, $\mathbf{R}_i \in C^0\left(\overline{U_i}, \mathbb{R}^n\right)$ such that $\mathbf{R}_i U_i$ is relatively open in $\mathbb{R}^n_{\leq} \equiv \{\mathbf{u} \in \mathbb{R}^n : u_1 \leq 0\}$, $\mathbf{R}_i^{-1}$ is continuous and one to one on $\mathbf{R}_i\left(\overline{U_i}\right)$. These mappings, $\mathbf{R}_i$, together with the relatively open sets U_i, are called charts and the totality of all the charts, $(U_i, \mathbf{R}_i)$ just described is called an atlas for the manifold. Define*

$$\text{int}\,(\Omega) \equiv \{\mathbf{x} \in \Omega : \text{ for some } i, \mathbf{R}_i \mathbf{x} \in \mathbb{R}^n_{<}\}$$

where $\mathbb{R}^n_{<} \equiv \{\mathbf{u} \in \mathbb{R}^n : u_1 < 0\}$. Also define

$$\partial\Omega \equiv \{\mathbf{x} \in \Omega : \text{ for some } i, \mathbf{R}_i \mathbf{x} \in \mathbb{R}^n_0\}$$

where

$$\mathbb{R}^n_0 \equiv \{\mathbf{u} \in \mathbb{R}^n : u_1 = 0\}$$

and $\partial\Omega$ is called the boundary of Ω. Note that if $n = 1$, $\mathbb{R}^n_0$ is just the single point 0. By convention, we will consider the boundary of such a 0 dimensional manifold to be empty.

Theorem 11.1.4 *Let $\partial\Omega$ and $\mathrm{int}(\Omega)$ be as defined above. Then $\mathrm{int}(\Omega)$ is open in Ω and $\partial\Omega$ is closed in Ω. Furthermore, $\partial\Omega \cap \mathrm{int}(\Omega) = \emptyset$, $\Omega = \partial\Omega \cup \mathrm{int}(\Omega)$, and for $n \geq 1$, $\partial\Omega$ is an $n-1$ dimensional manifold for which $\partial(\partial\Omega) = \emptyset$. The property of being in $\mathrm{int}(\Omega)$ or $\partial\Omega$ does not depend on the choice of atlas.*

Proof: It is clear that $\Omega = \partial\Omega \cup \mathrm{int}(\Omega)$. First consider the claim that $\partial\Omega \cap \mathrm{int}(\Omega) = \emptyset$. Suppose this does not happen. Then there would exist $\mathbf{x} \in \partial\Omega \cap \mathrm{int}(\Omega)$. Therefore, there would exist two mappings $\mathbf{R}_i$ and $\mathbf{R}_j$ such that $\mathbf{R}_j\mathbf{x} \in \mathbb{R}^n_0$ and $\mathbf{R}_i\mathbf{x} \in \mathbb{R}^n_<$ with $\mathbf{x} \in U_i \cap U_j$. Now consider the map, $\mathbf{R}_j \circ \mathbf{R}_i^{-1}$, a continuous one to one map from $\mathbb{R}^n_\leq$ to $\mathbb{R}^n_\leq$ having a continuous inverse. By continuity, there exists $r > 0$ small enough that,

$$\mathbf{R}_i^{-1} B(\mathbf{R}_i\mathbf{x}, r) \subseteq U_i \cap U_j.$$

Therefore, $\mathbf{R}_j \circ \mathbf{R}_i^{-1}(B(\mathbf{R}_i\mathbf{x}, r)) \subseteq \mathbb{R}^n_\leq$ and contains a point on $\mathbb{R}^n_0, \mathbf{R}_j\mathbf{x}$. However, this cannot occur because it contradicts the theorem on invariance of domain, Theorem 10.4.3, which requires that $\mathbf{R}_j \circ \mathbf{R}_i^{-1}(B(\mathbf{R}_i\mathbf{x}, r))$ must be an open subset of $\mathbb{R}^n$ and this one isn't because of the point on $\mathbb{R}^n_0$. Therefore, $\partial\Omega \cap \mathrm{int}(\Omega) = \emptyset$ as claimed. This same argument shows that the property of being in $\mathrm{int}(\Omega)$ or $\partial\Omega$ does not depend on the choice of the atlas.

To verify that $\partial(\partial\Omega) = \emptyset$, let $\mathbf{S}_i$ be the restriction of $\mathbf{R}_i$ to $\partial\Omega \cap U_i$. Thus

$$\mathbf{S}_i(\mathbf{x}) = (0, (\mathbf{R}_i\mathbf{x})_2, \cdots, (\mathbf{R}_i\mathbf{x})_n)$$

and the collection of such points for $\mathbf{x} \in \partial\Omega \cap U_i$ is an open bounded subset of

$$\{\mathbf{u} \in \mathbb{R}^n : u_1 = 0\},$$

identified with $\mathbb{R}^{n-1}$. $\mathbf{S}_i(\partial\Omega \cap U_i)$ is bounded because $\mathbf{S}_i$ is the restriction of a continuous function defined on $\mathbb{R}^m$ and $\partial\Omega \cap U_i \equiv V_i$ is contained in the compact set Ω. Thus if $\mathbf{S}_i$ is modified slightly, to be of the form

$$\mathbf{S}_i'(\mathbf{x}) = ((\mathbf{R}_i\mathbf{x})_2 - k_i, \cdots, (\mathbf{R}_i\mathbf{x})_n)$$

where k_i is chosen sufficiently large that $(\mathbf{R}_i(V_i))_2 - k_i < 0$, it follows that $\{(V_i, \mathbf{S}_i')\}$ is an atlas for $\partial\Omega$ as an $n-1$ dimensional manifold such that every point of $\partial\Omega$ is sent to to $\mathbb{R}^{n-1}_<$ and none gets sent to $\mathbb{R}^{n-1}_0$. It follows $\partial\Omega$ is an $n-1$ dimensional manifold with empty boundary. In case $n = 1$, the result follows by definition of the boundary of a 0 dimensional manifold.

Next consider the claim that $\mathrm{int}(\Omega)$ is open in Ω. If $\mathbf{x} \in \mathrm{int}(\Omega)$, are all points of Ω which are sufficiently close to $\mathbf{x}$ also in $\mathrm{int}(\Omega)$? If this were not true, there would exist $\{\mathbf{x}_n\}$ such that $\mathbf{x}_n \in \partial\Omega$ and $\mathbf{x}_n \to \mathbf{x}$. Since there are only finitely many charts of interest, this would imply the existence of a subsequence, still denoted by $\mathbf{x}_n$ and a single map, $\mathbf{R}_i$ such that $\mathbf{R}_i(\mathbf{x}_n) \in \mathbb{R}^n_0$. But then $\mathbf{R}_i(\mathbf{x}_n) \to \mathbf{R}_i(\mathbf{x})$ and so $\mathbf{R}_i(\mathbf{x}) \in \mathbb{R}^n_0$ showing $\mathbf{x} \in \partial\Omega$, a contradiction to $\mathrm{int}(\Omega) \cap \partial\Omega = \emptyset$. Now it follows that $\partial\Omega$ is closed in Ω because $\partial\Omega = \Omega \setminus \mathrm{int}(\Omega)$. This proves the theorem. ■

Definition 11.1.5 *An n dimensional manifold with boundary, Ω is a C^k manifold with boundary for some $k \geq 0$ if*

$$\mathbf{R}_j \circ \mathbf{R}_i^{-1} \in C^k\left(\overline{\mathbf{R}_i(U_i \cap U_j)}; \mathbb{R}^n\right)$$

and $\mathbf{R}_i^{-1} \in C^k\left(\overline{\mathbf{R}_i U_i}; \mathbb{R}^m\right)$. It is called a continuous manifold with boundary if the mappings, $\mathbf{R}_j \circ \mathbf{R}_i^{-1}$, $\mathbf{R}_i^{-1}, \mathbf{R}_i$ are continuous. In the case where Ω is a $C^k, k \geq 1$

manifold, it is called orientable if in addition to this there exists an atlas, $(U_r, \mathbf{R}_r)$*, such that whenever* $U_i \cap U_j \neq \emptyset$*,*

$$\det\left(D\left(\mathbf{R}_j \circ \mathbf{R}_i^{-1}\right)\right)(\mathbf{u}) > 0 \text{ for all } \mathbf{u} \in \mathbf{R}_i\left(U_i \cap U_j\right) \tag{11.1}$$

The mappings, $\mathbf{R}_i \circ \mathbf{R}_j^{-1}$ *are called the overlap maps. In the case where* $k = 0$*, the* $\mathbf{R}_i$ *are only assumed continuous so there is no differentiability available and in this case, the manifold is oriented if whenever* A *is an open connected subset of* $\operatorname{int}\left(\mathbf{R}_i\left(U_i \cap U_j\right)\right)$ *whose boundary has measure zero and separates* $\mathbb{R}^n$ *into two components,*

$$d\left(\mathbf{y}, A, \mathbf{R}_j \circ \mathbf{R}_i^{-1}\right) \in \{1, 0\} \tag{11.2}$$

depending on whether $\mathbf{y} \in \mathbf{R}_j \circ \mathbf{R}_i^{-1}(A)$*. An atlas satisfying 11.1 or more generally 11.2 is called an oriented atlas. By Lemma 10.6.4 and Proposition 10.5.9, this definition in terms of the degree when applied to the situation of a* C^k *manifold, gives the same thing as the earlier definition in terms of the determinant of the derivative.*

The advantage of using the degree in the above definition to define orientation is that it does not depend on any kind of differentiability and since I am trying to relax smoothness of the boundary, this is a good idea.

In calculus, you probably looked at piecewise smooth curves. The following is an attempt to generalize this to the present situation.

Definition 11.1.6 *In the above context, I will call* Ω *a* PC^1 *manifold if it is a* C^0 *manifold with charts* $(\mathbf{R}_i, U_i)$ *and there exists a closed set* $L \subseteq \Omega$ *such that* $\mathbf{R}_i\left(L \cap U_i\right)$ *is closed in* $\mathbf{R}_i\left(U_i\right)$ *and has* m_n *measure zero,* $\mathbf{R}_i\left(L \cap U_i \cap \partial\Omega\right)$ *is closed in* $\mathbb{R}^{n-1}$ *and has* m_{n-1} *measure zero, and the following conditions hold.*

$$\mathbf{R}_j \circ \mathbf{R}_i^{-1} \in C^1\left(\mathbf{R}_i\left(\left(U_i \cap U_j\right) \setminus L\right); \mathbb{R}^n\right) \tag{11.3}$$

$$\sup\left\{\left|\left|D\mathbf{R}_i^{-1}(\mathbf{u})\right|\right|_F : \mathbf{u} \in \mathbf{R}_i\left(U_i \setminus L\right)\right\} < \infty \tag{11.4}$$

Also, to deal with technical questions, assume that

$$\mathbf{R}_i, \mathbf{R}_i^{-1} \text{ are Lipschitz continuous.} \tag{11.5}$$

This assumption is made so that $\mathbf{R}_i \circ \mathbf{R}_j^{-1}$ *will map a set of measure zero to a set of measure zero. You can take the norm in the above to be the Frobenius norm*

$$||M||_F \equiv \left(\sum_{i,j} |M_{ij}|^2\right)^{1/2}$$

or the operator norm, whichever is more convenient. Note that 11.4 follows from 11.5. This is seen from taking a difference quotient and a limit.

The study of manifolds is really a generalization of something with which everyone who has taken a normal calculus course is familiar. We think of a point in three dimensional space in two ways. There is a geometric point and there are coordinates associated with this point. There are many different coordinate systems which describe a point. There are spherical coordinates, cylindrical coordinates and rectangular coordinates to name the three most popular coordinate systems. These coordinates are like the vector $\mathbf{u}$. The point, $\mathbf{x}$ is like the geometric point although it is always assumed here $\mathbf{x}$ has rectangular coordinates in $\mathbb{R}^m$ for some m. Under fairly general conditions, it can be shown there is no loss of generality in making such an assumption.

Now it will be convenient to use the following equivalent definition of orientable in the case of a PC^1 manifold.

Proposition 11.1.7 *Let Ω be a PC^1 n dimensional manifold with boundary. Then it is an orientable manifold if and only if there exists an atlas $\{(\mathbf{R}_i, U_i)\}$ such that for each i, j*

$$\det D\left(\mathbf{R}_i \circ \mathbf{R}_j^{-1}\right)(\mathbf{u}) \geq 0 \text{ a.e. } \mathbf{u} \in \operatorname{int}\left(\mathbf{R}_j\left(U_j \cap U_i\right)\right) \tag{11.6}$$

If $\mathbf{v} = \mathbf{R}_i \circ \mathbf{R}_j^{-1}(\mathbf{u})$, I will often write

$$\frac{\partial(v_1 \cdots v_n)}{\partial(u_1 \cdots u_n)} \equiv \det D\mathbf{R}_i \circ \mathbf{R}_j^{-1}(\mathbf{u})$$

Thus in this situation $\frac{\partial(v_1 \cdots v_n)}{\partial(u_1 \cdots u_n)} \geq 0$.

Proof: Suppose first the chart is an oriented chart so

$$d\left(\mathbf{v}, A, \mathbf{R}_i \circ \mathbf{R}_j^{-1}\right) = 1$$

whenever $\mathbf{v} \in \operatorname{int}\mathbf{R}_j(A)$ where A is an open ball contained in $\mathbf{R}_i \circ \mathbf{R}_j^{-1}\left(U_i \cap U_j \setminus L\right)$. Then by Theorem 10.6.5, if $E \subseteq A$ is any Borel measurable set,

$$0 \leq \int_{\mathbf{R}_i \circ \mathbf{R}_j^{-1}(A)} \mathcal{X}_{\mathbf{R}_i \circ \mathbf{R}_j^{-1}(E)}(\mathbf{v})\, 1 dv = \int_A \det\left(D\left(\mathbf{R}_i \circ \mathbf{R}_j^{-1}\right)(\mathbf{u})\right) \mathcal{X}_E(\mathbf{u})\, du$$

Since this is true for arbitrary $E \subseteq A$, it follows $\det\left(D\left(\mathbf{R}_i \circ \mathbf{R}_j^{-1}\right)(\mathbf{u})\right) \geq 0$ *a.e.* $\mathbf{u} \in A$ because if not so, then you could take $E_\delta \equiv \left\{\mathbf{u} : \det\left(D\left(\mathbf{R}_i \circ \mathbf{R}_j^{-1}\right)(\mathbf{u})\right) < -\delta\right\}$ and for some $\delta > 0$ this would have positive measure. Then the right side of the above is negative while the left is nonnegative. By the Vitali covering theorem Corollary 9.7.6, and the assumptions of PC^1, there exists a sequence of disjoint open balls contained in $\mathbf{R}_i \circ \mathbf{R}_j^{-1}\left(U_i \cap U_j \setminus L\right), \{A_k\}$ such that

$$\operatorname{int}\left(\mathbf{R}_i \circ \mathbf{R}_j^{-1}\left(U_j \cap U_i\right)\right) = L \cup \cup_{k=1}^{\infty} A_k$$

and from the above, there exist sets of measure zero $N_k \subseteq A_k$ such that

$$\det D\left(\mathbf{R}_i \circ \mathbf{R}_j^{-1}\right)(\mathbf{u}) \geq 0$$

for all $\mathbf{u} \in A_k \setminus N_k$. Then $\det D\left(\mathbf{R}_i \circ \mathbf{R}_j^{-1}\right)(\mathbf{u}) \geq 0$ on $\operatorname{int}\left(\mathbf{R}_i \circ \mathbf{R}_j^{-1}\left(U_j \cap U_i\right)\right) \setminus \left(L \cup \cup_{k=1}^{\infty} N_k\right)$. This proves one direction. Now consider the other direction.

Suppose the condition $\det D\left(\mathbf{R}_i \circ \mathbf{R}_j^{-1}\right)(\mathbf{u}) \geq 0$ a.e. Then by Theorem 10.6.5

$$\int_{\mathbf{R}_i \circ \mathbf{R}_j^{-1}(A)} d\left(\mathbf{v}, A, \mathbf{R}_i \circ \mathbf{R}_j^{-1}\right) dv = \int_A \det\left(D\left(\mathbf{R}_i \circ \mathbf{R}_j^{-1}\right)(\mathbf{u})\right) du \geq 0$$

The degree is constant on the connected open set $\mathbf{R}_i \circ \mathbf{R}_j^{-1}(A)$. By Proposition 10.5.9, the degree equals either -1 or 1. The above inequality shows it can't equal -1 and so it must equal 1. This proves the proposition. ■

This shows it would be fine to simply use 11.6 as the definition of orientable in the case of a PC^1 manifold and not bother with the definiton in terms of the degree. This is exactly what will be done in what follows. The version defined in terms of the degree is more general because it does not depend on any differentiability.

11.2 Some Important Measure Theory

11.2.1 Eggoroff's Theorem

Eggoroff's theorem says that if a sequence converges pointwise, then it almost converges uniformly in a certain sense.

Theorem 11.2.1 *(Egoroff) Let $(\Omega, \mathcal{F}, \mu)$ be a finite measure space,*

$$(\mu(\Omega) < \infty)$$

and let f_n, f be complex valued functions such that $\operatorname{Re} f_n, \operatorname{Im} f_n$ are all measurable and

$$\lim_{n\to\infty} f_n(\omega) = f(\omega)$$

for all $\omega \notin E$ where $\mu(E) = 0$. Then for every $\varepsilon > 0$, there exists a set,

$$F \supseteq E,\ \mu(F) < \varepsilon,$$

such that f_n converges uniformly to f on F^C.

Proof: First suppose $E = \emptyset$ so that convergence is pointwise everywhere. It follows then that $\operatorname{Re} f$ and $\operatorname{Im} f$ are pointwise limits of measurable functions and are therefore measurable. Let $E_{km} = \{\omega \in \Omega : |f_n(\omega) - f(\omega)| \geq 1/m \text{ for some } n > k\}$. Note that

$$|f_n(\omega) - f(\omega)| = \sqrt{(\operatorname{Re} f_n(\omega) - \operatorname{Re} f(\omega))^2 + (\operatorname{Im} f_n(\omega) - \operatorname{Im} f(\omega))^2}$$

and so,

$$\left[|f_n - f| \geq \frac{1}{m}\right]$$

is measurable. Hence E_{km} is measurable because

$$E_{km} = \cup_{n=k+1}^{\infty} \left[|f_n - f| \geq \frac{1}{m}\right].$$

For fixed $m, \cap_{k=1}^{\infty} E_{km} = \emptyset$ because f_n converges to f. Therefore, if $\omega \in \Omega$ there exists k such that if $n > k$, $|f_n(\omega) - f(\omega)| < \frac{1}{m}$ which means $\omega \notin E_{km}$. Note also that

$$E_{km} \supseteq E_{(k+1)m}.$$

Since $\mu(E_{1m}) < \infty$, Theorem 7.3.2 on Page 163 implies

$$0 = \mu(\cap_{k=1}^{\infty} E_{km}) = \lim_{k\to\infty} \mu(E_{km}).$$

Let $k(m)$ be chosen such that $\mu(E_{k(m)m}) < \varepsilon 2^{-m}$ and let

$$F = \bigcup_{m=1}^{\infty} E_{k(m)m}.$$

Then $\mu(F) < \varepsilon$ because

$$\mu(F) \leq \sum_{m=1}^{\infty} \mu\left(E_{k(m)m}\right) < \sum_{m=1}^{\infty} \varepsilon 2^{-m} = \varepsilon$$

Now let $\eta > 0$ be given and pick m_0 such that $m_0^{-1} < \eta$. If $\omega \in F^C$, then

$$\omega \in \bigcap_{m=1}^{\infty} E_{k(m)m}^C.$$

Hence $\omega \in E_{k(m_0)m_0}^C$ so

$$|f_n(\omega) - f(\omega)| < 1/m_0 < \eta$$

for all $n > k(m_0)$. This holds for all $\omega \in F^C$ and so f_n converges uniformly to f on F^C.

Now if $E \neq \emptyset$, consider $\{\mathcal{X}_{E^C} f_n\}_{n=1}^{\infty}$. Each $\mathcal{X}_{E^C} f_n$ has real and imaginary parts measurable and the sequence converges pointwise to $\mathcal{X}_E f$ everywhere. Therefore, from the first part, there exists a set of measure less than ε, F such that on $F^C, \{\mathcal{X}_{E^C} f_n\}$ converges uniformly to $\mathcal{X}_{E^C} f$. Therefore, on $(E \cup F)^C$, $\{f_n\}$ converges uniformly to f. This proves the theorem. ■

11.2.2 The Vitali Convergence Theorem

The Vitali convergence theorem is a convergence theorem which in the case of a finite measure space is superior to the dominated convergence theorem.

Definition 11.2.2 *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $\mathcal{S} \subseteq L^1(\Omega)$. $\mathcal{S}$ is uniformly integrable if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all $f \in \mathcal{S}$*

$$|\int_E f d\mu| < \varepsilon \text{ whenever } \mu(E) < \delta.$$

Lemma 11.2.3 *If $\mathcal{S}$ is uniformly integrable, then $|\mathcal{S}| \equiv \{|f| : f \in \mathcal{S}\}$ is uniformly integrable. Also $\mathcal{S}$ is uniformly integrable if $\mathcal{S}$ is finite.*

Proof: Let $\varepsilon > 0$ be given and suppose $\mathcal{S}$ is uniformly integrable. First suppose the functions are real valued. Let δ be such that if $\mu(E) < \delta$, then

$$\left|\int_E f d\mu\right| < \frac{\varepsilon}{2}$$

for all $f \in \mathcal{S}$. Let $\mu(E) < \delta$. Then if $f \in \mathcal{S}$,

$$\begin{aligned} \int_E |f| \, d\mu &\leq \int_{E\cap[f\leq 0]} (-f) \, d\mu + \int_{E\cap[f>0]} f d\mu \\ &= \left|\int_{E\cap[f\leq 0]} f d\mu\right| + \left|\int_{E\cap[f>0]} f d\mu\right| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

In general, if $\mathcal{S}$ is a uniformly integrable set of complex valued functions, the inequalities,

$$\left|\int_E \operatorname{Re} f d\mu\right| \leq \left|\int_E f d\mu\right|, \left|\int_E \operatorname{Im} f d\mu\right| \leq \left|\int_E f d\mu\right|,$$

imply $\operatorname{Re}\mathcal{S} \equiv \{\operatorname{Re} f : f \in \mathcal{S}\}$ and $\operatorname{Im}\mathcal{S} \equiv \{\operatorname{Im} f : f \in \mathcal{S}\}$ are also uniformly integrable. Therefore, applying the above result for real valued functions to these sets of functions, it follows $|\mathcal{S}|$ is uniformly integrable also.

For the last part, is suffices to verify a single function in $L^1(\Omega)$ is uniformly integrable. To do so, note that from the dominated convergence theorem,

$$\lim_{R\to\infty} \int_{[|f|>R]} |f| \, d\mu = 0.$$

Let $\varepsilon > 0$ be given and choose R large enough that $\int_{[|f|>R]} |f| \, d\mu < \frac{\varepsilon}{2}$. Now let $\mu(E) < \frac{\varepsilon}{2R}$. Then

$$\begin{aligned} \int_E |f| \, d\mu &= \int_{E\cap[|f|\leq R]} |f| \, d\mu + \int_{E\cap[|f|>R]} |f| \, d\mu \\ &< R\mu(E) + \frac{\varepsilon}{2} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

This proves the lemma. ■

The following gives a nice way to identify a uniformly integrable set of functions.

Lemma 11.2.4 *Let $\mathfrak{S}$ be a subset of $L^1(\Omega,\mu)$ where $\mu(\Omega)<\infty$. Let $t\to h(t)$ be a continuous function which satisfies*

$$\lim_{t\to\infty}\frac{h(t)}{t}=\infty$$

Then $\mathfrak{S}$ is uniformly integrable and bounded in $L^1(\Omega)$ if

$$\sup\left\{\int_\Omega h(|f|)\,d\mu : f\in\mathfrak{S}\right\}=N<\infty.$$

Proof: First I show $\mathfrak{S}$ is bounded in $L^1(\Omega;\mu)$ which means there exists a constant M such that for all $f\in\mathfrak{S}$,

$$\int_\Omega |f|\,d\mu\le M.$$

From the properties of h, there exists R_n such that if $t\ge R_n$, then $h(t)\ge nt$. Therefore,

$$\int_\Omega |f|\,d\mu=\int_{[|f|\ge R_n]}|f|\,d\mu+\int_{[|f|<R_n]}|f|\,d\mu$$

Letting $n=1$,

$$\begin{aligned}\int_\Omega |f|\,d\mu &\le \int_{[|f|\ge R_1]}h(|f|)\,d\mu+R_1\mu([|f|<R_1])\\ &\le N+R_1\mu(\Omega)\equiv M.\end{aligned}$$

Next let E be a measurable set. Then for every $f\in\mathfrak{S}$,

$$\int_E |f|\,d\mu=\int_{[|f|\ge R_n]\cap E}|f|\,d\mu+\int_{[|f|<R_n]\cap E}|f|\,d\mu$$

$$\le\frac{1}{n}\int_\Omega |f|\,d\mu+R_n\mu(E)\le\frac{N}{n}+R_n\mu(E)$$

and letting n be large enough, this is less than

$$\varepsilon/2+R_n\mu(E)$$

Now if $\mu(E)<\varepsilon/2R_n$, it follows that for all $f\in\mathfrak{S}$,

$$\int_E |f|\,d\mu<\varepsilon$$

This proves the lemma. ■

Letting $h(t)=t^2$, it follows that if all the functions in $\mathfrak{S}$ are bounded, then the collection of functions is uniformly integrable.

The following theorem is Vitali's convergence theorem.

Theorem 11.2.5 *Let $\{f_n\}$ be a uniformly integrable set of complex valued functions, $\mu(\Omega)<\infty$, and $f_n(x)\to f(x)$ a.e. where f is a measurable complex valued function. Then $f\in L^1(\Omega)$ and*

$$\lim_{n\to\infty}\int_\Omega |f_n-f|d\mu=0. \tag{11.7}$$

Proof: First it will be shown that $f \in L^1(\Omega)$. By uniform integrability, there exists $\delta > 0$ such that if $\mu(E) < \delta$, then

$$\int_E |f_n| \, d\mu < 1$$

for all n. By Egoroff's theorem, there exists a set, E of measure less than δ such that on E^C, $\{f_n\}$ converges uniformly. Therefore, for p large enough, and $n > p$,

$$\int_{E^C} |f_p - f_n| \, d\mu < 1$$

which implies

$$\int_{E^C} |f_n| \, d\mu < 1 + \int_\Omega |f_p| \, d\mu.$$

Then since there are only finitely many functions, f_n with $n \leq p$, there exists a constant, M_1 such that for all n,

$$\int_{E^C} |f_n| \, d\mu < M_1.$$

But also,

$$\begin{aligned} \int_\Omega |f_m| \, d\mu &= \int_{E^C} |f_m| \, d\mu + \int_E |f_m| \\ &\leq M_1 + 1 \equiv M. \end{aligned}$$

Therefore, by Fatou's lemma,

$$\int_\Omega |f| \, d\mu \leq \lim \inf_{n\to\infty} \int |f_n| \, d\mu \leq M,$$

showing that $f \in L^1$ as hoped.

Now $\mathfrak{S} \cup \{f\}$ is uniformly integrable so there exists $\delta_1 > 0$ such that if $\mu(E) < \delta_1$, then $\int_E |g| \, d\mu < \varepsilon/3$ for all $g \in \mathfrak{S} \cup \{f\}$.

By Egoroff's theorem, there exists a set, F with $\mu(F) < \delta_1$ such that f_n converges uniformly to f on F^C. Therefore, there exists N such that if $n > N$, then

$$\int_{F^C} |f - f_n| \, d\mu < \frac{\varepsilon}{3}.$$

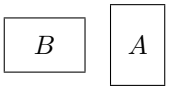
It follows that for $n > N$,

$$\begin{aligned} \int_\Omega |f - f_n| \, d\mu &\leq \int_{F^C} |f - f_n| \, d\mu + \int_F |f| \, d\mu + \int_F |f_n| \, d\mu \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon, \end{aligned}$$

which verifies 11.7. ■

11.3 The Binet Cauchy Formula

The Binet Cauchy formula is a generalization of the theorem which says the determinant of a product is the product of the determinants. The situation is illustrated in the following picture where A, B are matrices.



Theorem 11.3.1 *Let A be an $n \times m$ matrix with $n \geq m$ and let B be a $m \times n$ matrix. Also let A_i*

$$i = 1, \cdots, C(n,m)$$

be the $m \times m$ submatrices of A which are obtained by deleting $n-m$ rows and let B_i be the $m \times m$ submatrices of B which are obtained by deleting corresponding $n-m$ columns. Then

$$\det(BA) = \sum_{k=1}^{C(n,m)} \det(B_k) \det(A_k)$$

Proof: This follows from a computation. By Corollary 3.5.6 on Page 42, $\det(BA) =$

$$\frac{1}{m!} \sum_{(i_1 \cdots i_m)} \sum_{(j_1 \cdots j_m)} \operatorname{sgn}(i_1 \cdots i_m) \operatorname{sgn}(j_1 \cdots j_m) (BA)_{i_1 j_1} (BA)_{i_2 j_2} \cdots (BA)_{i_m j_m}$$

$$\frac{1}{m!} \sum_{(i_1 \cdots i_m)} \sum_{(j_1 \cdots j_m)} \operatorname{sgn}(i_1 \cdots i_m) \operatorname{sgn}(j_1 \cdots j_m) \cdot \sum_{r_1=1}^{n} B_{i_1 r_1} A_{r_1 j_1} \sum_{r_2=1}^{n} B_{i_2 r_2} A_{r_2 j_2} \cdots \sum_{r_m=1}^{n} B_{i_m r_m} A_{r_m j_m}$$

Now denote by I_k one subsets of $\{1, \cdots, n\}$ having m elements. Thus there are $C(n,m)$ of these. Then the above equals

$$= \sum_{k=1}^{C(n,m)} \sum_{\{r_1, \cdots, r_m\} = I_k} \frac{1}{m!} \sum_{(i_1 \cdots i_m)} \sum_{(j_1 \cdots j_m)} \operatorname{sgn}(i_1 \cdots i_m) \operatorname{sgn}(j_1 \cdots j_m) \cdot B_{i_1 r_1} A_{r_1 j_1} B_{i_2 r_2} A_{r_2 j_2} \cdots B_{i_m r_m} A_{r_m j_m}$$

$$= \sum_{k=1}^{C(n,m)} \sum_{\{r_1, \cdots, r_m\} = I_k} \frac{1}{m!} \sum_{(i_1 \cdots i_m)} \operatorname{sgn}(i_1 \cdots i_m) B_{i_1 r_1} B_{i_2 r_2} \cdots B_{i_m r_m} \cdot \sum_{(j_1 \cdots j_m)} \operatorname{sgn}(j_1 \cdots j_m) A_{r_1 j_1} A_{r_2 j_2} \cdots A_{r_m j_m}$$

$$= \sum_{k=1}^{C(n,m)} \sum_{\{r_1, \cdots, r_m\} = I_k} \frac{1}{m!} \operatorname{sgn}(r_1 \cdots r_m)^2 \det(B_k) \det(A_k)$$

$$= \sum_{k=1}^{C(n,m)} \det(B_k) \det(A_k)$$

since there are $m!$ ways of arranging the indices $\{r_1, \cdots, r_m\}$. ■

11.4 The Area Measure On A Manifold

It is convenient to specify a "surface measure" on a manifold. This concept is a little easier because you don't have to worry about orientation. It will involve the following definition.

Definition 11.4.1 *Let $(U_i, \mathbf{R}_i)$ be an atlas for an n dimensional PC^1 manifold Ω. Also let $\{\psi_i\}_{i=1}^p$ be a C^∞ partition of unity, $\operatorname{spt}\psi_i \subseteq U_i$. Then for E a Borel subset of Ω, define*

$$\sigma_n(E) \equiv \sum_{i=1}^p \int_{\mathbf{R}_i U_i} \psi_i\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \mathcal{X}_E\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) J_i(\mathbf{u})\, du$$

where

$$J_i(\mathbf{u}) \equiv \left(\det\left(D\mathbf{R}_i^{-1}(\mathbf{u})^* D\mathbf{R}_i^{-1}(\mathbf{u})\right)\right)^{1/2}$$

By the Binet Cauchy theorem, this equals

$$\left(\sum_{i_1,\cdots,i_n}\left(\frac{\partial(x_{i_1}\cdots x_{i_n})}{\partial(u_1\cdots u_n)}(\mathbf{u})\right)^2\right)^{1/2}$$

where the sum is taken over all increasing strings of n indices $(i_1,\cdots,i_n)$ and

$$\frac{\partial(x_{i_1}\cdots x_{i_n})}{\partial(u_1\cdots u_n)}(\mathbf{u}) \equiv \det\begin{pmatrix} x_{i_1,u_1} & x_{i_1,u_2} & \cdots & x_{i_1,u_n} \\ x_{i_2,u_1} & x_{i_2,u_2} & \cdots & x_{i_2,u_2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i_n,u_1} & x_{i_n,u_2} & \cdots & x_{i_n,u_n} \end{pmatrix}(\mathbf{u}) \tag{11.8}$$

Suppose $(V_j, \mathbf{S}_j)$ is another atlas and you have

$$\mathbf{v} = \mathbf{S}_j \circ \mathbf{R}_i^{-1}(\mathbf{u}) \tag{11.9}$$

for $\mathbf{u} \in \mathbf{R}_i\left((V_j \cap U_i) \setminus L\right)$. Then $\mathbf{R}_i^{-1} = \mathbf{S}_j^{-1} \circ \left(\mathbf{S}_j \circ \mathbf{R}_i^{-1}\right)$ and so by the chain rule,

$$D\mathbf{S}_j^{-1}(\mathbf{v})\, D\left(\mathbf{S}_j \circ \mathbf{R}_i^{-1}\right)(\mathbf{u}) = D\mathbf{R}_i^{-1}(\mathbf{u})$$

Therefore,

$$J_i(\mathbf{u}) = \left(\det\left(D\mathbf{R}_i^{-1}(\mathbf{u})^* D\mathbf{R}_i^{-1}(\mathbf{u})\right)\right)^{1/2}$$

$$= \left(\det\left(\overbrace{D\left(\mathbf{S}_j \circ \mathbf{R}_i^{-1}\right)(\mathbf{u})^*}^{n\times n} \overbrace{D\mathbf{S}_j^{-1}(\mathbf{v})^*}^{n\times n} \overbrace{D\mathbf{S}_j^{-1}(\mathbf{v}) D\left(\mathbf{S}_j \circ \mathbf{R}_i^{-1}\right)(\mathbf{u})}^{n\times n}\right)\right)^{1/2}$$

$$= \left|\det\left(D\left(\mathbf{S}_j \circ \mathbf{R}_i^{-1}\right)(\mathbf{u})\right)\right| J_j(\mathbf{v}) \tag{11.10}$$

Similarly

$$J_j(\mathbf{v}) = \left|\det\left(D\left(\mathbf{R}_i \circ \mathbf{S}_j^{-1}\right)(\mathbf{v})\right)\right| J_i(\mathbf{u}). \tag{11.11}$$

In the situation of 11.9, it is convenient to use the notation

$$\frac{\partial(v_1\cdots v_n)}{\partial(u_1\cdots u_n)} \equiv \det\left(D\left(\mathbf{S}_j \circ \mathbf{R}_i^{-1}\right)(\mathbf{u})\right)$$

and this will be used occasionally below.

It is necessary to show the above definition is well defined.

Theorem 11.4.2 *The above definition of surface measure is well defined. That is, suppose Ω is an n dimensional orientable PC^1 manifold with boundary and let $\{(U_i, \mathbf{R}_i)\}_{i=1}^p$ and $\{(V_j, \mathbf{S}_j)\}_{j=1}^q$ be two atlass of Ω. Then for E a Borel set, the computation of $\sigma_n(E)$ using the two different atlass gives the same thing. This defines a Borel measure on Ω. Furthermore, if $E \subseteq U_i$, $\sigma_n(E)$ reduces to*

$$\int_{\mathbf{R}_i U_i} \mathcal{X}_E\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) J_i(\mathbf{u})\, du$$

Also $\sigma_n(L) = 0$.

Proof: Let $\{\psi_i\}$ be a partition of unity as described in Lemma 11.5.3 which is associated with the atlas $(U_i, \mathbf{R}_i)$ and let $\{\eta_i\}$ be a partition of unity associated in the same manner with the atlas $(V_i, \mathbf{S}_i)$. First note the following.

$$\int_{\mathbf{R}_i U_i} \psi_i\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \mathcal{X}_E\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) J_i(\mathbf{u})\, du$$

$$= \sum_{j=1}^{q} \int_{\mathbf{R}_i(U_i \cap V_j)} \eta_j\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \psi_i\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \mathcal{X}_E\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) J_i(\mathbf{u})\, du$$

$$= \sum_{j=1}^{q} \int_{\operatorname{int} \mathbf{R}_i(U_i \cap V_j \setminus L)} \eta_j\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \psi_i\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \mathcal{X}_E\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) J_i(\mathbf{u})\, du$$

The reason this can be done is that points not on the interior of $\mathbf{R}_i(U_i \cap V_j)$ are on the plane $u_1 = 0$ which is a set of measure zero and by assumption $\mathbf{R}_i(L \cap U_i \cap V_j)$ has measure zero. Of course the above determinants in the definition of $J_i(\mathbf{u})$ in the integrand are not even defined on the set of measure zero $\mathbf{R}_i(L \cap U_i)$. It follows the definition of $\sigma_n(E)$ in terms of the atlas $\{(U_i, \mathbf{R}_i)\}_{i=1}^{p}$ and specified partition of unity is

$$\sum_{i=1}^{p} \sum_{j=1}^{q} \int_{\operatorname{int} \mathbf{R}_i(U_i \cap V_j \setminus L)} \eta_j\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \psi_i\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \mathcal{X}_E\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) J_i(\mathbf{u})\, du$$

By the change of variables formula, Theorem 9.9.4, this equals

$$\sum_{i=1}^{p} \sum_{j=1}^{q} \int_{\left(\mathbf{S}_j \circ \mathbf{R}_i^{-1}\right)(\operatorname{int} \mathbf{R}_i(U_i \cap V_j \setminus L))} \eta_j\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) \psi_i\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) \cdot$$
$$\mathcal{X}_E\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) J_i(\mathbf{u}) \left|\det\left(\mathbf{R}_i \circ \mathbf{S}_j^{-1}(\mathbf{v})\right)\right| dv$$

The integral is unchanged if it is taken over $\mathbf{S}_j(U_i \cap V_j)$. This is because the map $\mathbf{S}_j \circ \mathbf{R}_i^{-1}$ is Lipschitz and so it takes a set of measure zero to one of measure zero by Corollary 9.8.2. By 11.11, this equals

$$\sum_{i=1}^{p} \sum_{j=1}^{q} \int_{\mathbf{S}_j(U_i \cap V_j)} \eta_j\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) \psi_i\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) \mathcal{X}_E\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) J_j(\mathbf{v})\, dv$$
$$= \sum_{j=1}^{q} \int_{\mathbf{S}_j(U_i \cap V_j)} \eta_j\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) \mathcal{X}_E\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) J_j(\mathbf{v})\, dv$$

which equals the definition of $\sigma_n(E)$ taken with respect to the other atlas and partition of unity. Thus the definition is well defined. This also has shown that the partition of unity can be picked at will.

It remains to verify the claim. First suppose $E = K$ a compact subset of U_i. Then using Lemma 11.5.3 there exists a partition of unity such that $\psi_k = 1$ on K. Consider the sum used to define $\sigma_n(K)$,

$$\sum_{i=1}^{p} \int_{\mathbf{R}_i U_i} \psi_i\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \mathcal{X}_K\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) J_i(\mathbf{u})\, du$$

Unless $\mathbf{R}_i^{-1}(\mathbf{u}) \in K$, the integrand equals 0. Assume then that $\mathbf{R}_i^{-1}(\mathbf{u}) \in K$. If $i \neq k, \psi_i\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) = 0$ because $\psi_k\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) = 1$ and these functions sum to 1. Therefore, the above sum reduces to

$$\int_{\mathbf{R}_k U_k} \mathcal{X}_K\left(\mathbf{R}_k^{-1}(\mathbf{u})\right) J_k(\mathbf{u})\, du.$$

Next consider the general case. By Theorem 7.4.6 the Borel measure σ_n is regular. Also Lebesgue measure is regular. Therefore, there exists an increasing sequence of compact subsets of E, $\{K_r\}_{k=1}^{\infty}$ such that

$$\lim_{r\to\infty} \sigma_n(K_r) = \sigma_n(E)$$

Then letting $F = \cup_{r=1}^{\infty} K_r$, the monotone convergence theorem implies

$$\begin{aligned}\sigma_n(E) &= \lim_{r\to\infty} \sigma_n(K_r) = \int_{\mathbf{R}_k U_k} \mathcal{X}_F\left(\mathbf{R}_k^{-1}(\mathbf{u})\right) J_k(\mathbf{u})\, du \\ &\leq \int_{\mathbf{R}_k U_k} \mathcal{X}_E\left(\mathbf{R}_k^{-1}(\mathbf{u})\right) J_k(\mathbf{u})\, du\end{aligned}$$

Next take an increasing sequence of compact sets contained in $\mathbf{R}_k(E)$ such that

$$\lim_{r\to\infty} m_n(K_r) = m_n(\mathbf{R}_k(E)).$$

Thus $\left\{\mathbf{R}_k^{-1}(K_r)\right\}_{r=1}^{\infty}$ is an increasing sequence of compact subsets of E. Therefore,

$$\begin{aligned}\sigma_n(E) &\geq \lim_{r\to\infty} \sigma_n\left(\mathbf{R}_k^{-1}(K_r)\right) = \lim_{r\to\infty} \int_{\mathbf{R}_k U_k} \mathcal{X}_{K_r}(\mathbf{u})\, J_k(\mathbf{u})\, du \\ &= \int_{\mathbf{R}_k U_k} \mathcal{X}_{\mathbf{R}_k(E)}(\mathbf{u})\, J_k(\mathbf{u})\, du \\ &= \int_{\mathbf{R}_k U_k} \mathcal{X}_E\left(\mathbf{R}_k^{-1}(\mathbf{u})\right) J_k(\mathbf{u})\, du\end{aligned}$$

Thus

$$\sigma_n(E) = \int_{\mathbf{R}_k U_k} \mathcal{X}_E\left(\mathbf{R}_k^{-1}(\mathbf{u})\right) J_k(\mathbf{u})\, du$$

as claimed.

So what is the measure of L? By definition it equals

$$\begin{aligned}\sigma_n(L) &\equiv \sum_{i=1}^{p} \int_{\mathbf{R}_i U_i} \psi_i\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \mathcal{X}_L\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) J_i(\mathbf{u})\, du \\ &= \sum_{i=1}^{p} \int_{\mathbf{R}_i U_i} \psi_i\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \mathcal{X}_{\mathbf{R}_i(L\cap U_i)}(\mathbf{u})\, J_i(\mathbf{u})\, du = 0\end{aligned}$$

because by assumption, the measure of each $\mathbf{R}_i(L \cap U_i)$ is zero. This proves the theorem. ■

All of this ends up working if you only assume the overlap maps $\mathbf{R}_i \circ \mathbf{R}_j^{-1}$ are Lipschitz. However, this involves Rademacher's theorem which says that Lipschitz maps are differentiable almost everywhere and this is not something which has been discussed.

11.5 Integration Of Differential Forms On Manifolds

This section presents the integration of differential forms on manifolds. This topic is a higher dimensional version of what is done in calculus in finding the work done by a force field on an object which moves over some path. There you evaluated line integrals. Differential forms are just a higher dimensional version of this idea and it turns out they are what it makes sense to integrate on manifolds. The following lemma, on Page 253 used in establishing the definition of the degree and in giving a proof of the Brouwer fixed point theorem is also a fundamental result in discussing the integration of differential forms.

Lemma 11.5.1 *Let $\mathbf{g} : U \to V$ be C^2 where U and V are open subsets of $\mathbb{R}^n$. Then*

$$\sum_{j=1}^{n} (cof(D\mathbf{g}))_{ij,j} = 0,$$

where here $(D\mathbf{g})_{ij} \equiv g_{i,j} \equiv \frac{\partial g_i}{\partial x_j}$.

Recall Proposition 10.5.9.

Proposition 11.5.2 *Let Ω be an open connected bounded set in $\mathbb{R}^n$ such that $\mathbb{R}^n \setminus \partial\Omega$ consists of two, three if $n = 1$, connected components. Let $\mathbf{f} \in C\left(\overline{\Omega}; \mathbb{R}^n\right)$ be continuous and one to one. Then $\mathbf{f}(\Omega)$ is the bounded component of $\mathbb{R}^n \setminus \mathbf{f}(\partial\Omega)$ and for $\mathbf{y} \in \mathbf{f}(\Omega)$, $d(\mathbf{f}, \Omega, \mathbf{y})$ either equals 1 or -1.*

Also recall the following fundamental lemma on partitions of unity in Lemma 9.5.15 and Corollary 9.5.14.

Lemma 11.5.3 *Let K be a compact set in $\mathbb{R}^n$ and let $\{U_i\}_{i=1}^m$ be an open cover of K. Then there exist functions, $\psi_k \in C_c^\infty(U_i)$ such that $\psi_i \prec U_i$ and for all $\mathbf{x} \in K$,*

$$\sum_{i=1}^{m} \psi_i(\mathbf{x}) = 1.$$

If K is a compact subset of U_1 (U_i)there exist such functions such that also $\psi_1(\mathbf{x}) = 1$ ($\psi_i(\mathbf{x}) = 1$) for all $\mathbf{x} \in K$.

With the above, what follows is the definition of what a differential form is and how to integrate one.

Definition 11.5.4 *Let I denote an ordered list of n indices taken from the set, $\{1, \cdots, m\}$. Thus $I = (i_1, \cdots, i_n)$. It is an ordered list because the order matters. A differential form of order n in $\mathbb{R}^m$ is a formal expression,*

$$\omega = \sum_I a_I(\mathbf{x})\, d\mathbf{x}^I$$

where a_I is at least Borel measurable $d\mathbf{x}^I$ is short for the expression

$$dx_{i_1} \wedge \cdots \wedge dx_{i_n},$$

and the sum is taken over all ordered lists of indices taken from the set, $\{1, \cdots, m\}$. For Ω an orientable n dimensional manifold with boundary, let $\{(U_i, \mathbf{R}_i)\}$ be an oriented atlas for Ω. Each U_i is the intersection of an open set in $\mathbb{R}^m$, O_i, with Ω and so there exists a C^∞ partition of unity subordinate to the open cover, $\{O_i\}$ which sums to 1 on Ω. Thus $\psi_i \in C_c^\infty(O_i)$, has values in $[0,1]$ and satisfies $\sum_i \psi_i(\mathbf{x}) = 1$ for all $\mathbf{x} \in \Omega$. Define

$$\int_\Omega \omega \equiv \sum_{i=1}^{p} \sum_I \int_{\mathbf{R}_i U_i} \psi_i\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) a_I\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \frac{\partial(x_{i_1} \cdots x_{i_n})}{\partial(u_1 \cdots u_n)} du \tag{11.12}$$

Note

$$\frac{\partial(x_{i_1} \cdots x_{i_n})}{\partial(u_1 \cdots u_n)},$$

given by 11.8 is not defined on $\mathbf{R}_i(U_i \cap L)$ but this is assumed a set of measure zero so it is not important in the integral.

Of course there are all sorts of questions related to whether this definition is well defined. What if you had a different atlas and a different partition of unity? Would $\int_\Omega \omega$ change? In general, the answer is yes. However, there is a sense in which the integral of a differential form is well defined. This involves the concept of orientation.

Definition 11.5.5 *Suppose Ω is an n dimensional orientable manifold with boundary and let $(U_i, \mathbf{R}_i)$ and $(V_i, \mathbf{S}_i)$ be two oriented atlass of Ω. They have the same orientation if for all open connected sets $A \subseteq \mathbf{S}_j(V_j \cap U_i)$ with ∂A having measure zero and separating $\mathbb{R}^n$ into two components,*

$$d\left(\mathbf{u}, \mathbf{R}_i \circ \mathbf{S}_j^{-1}, A\right) \in \{0,1\}$$

depending on whether $\mathbf{u} \in \mathbf{R}_i \circ \mathbf{S}_j^{-1}(A)$. In terms of the derivative in the case where the manifold is PC^1, this is equivalent to having

$$\det\left(D\left(\mathbf{R}_i \circ \mathbf{S}_j^{-1}\right)\right) > 0 \text{ on } \mathbf{S}_j(V_j \cap U_i \setminus L)$$

The above definition of $\int_\Omega \omega$ is well defined in the sense that any two atlass which have the same orientation deliver the same value for this symbol.

Theorem 11.5.6 *Suppose Ω is an n dimensional orientable PC^1 manifold with boundary and let $(U_i, \mathbf{R}_i)$ and $(V_i, \mathbf{S}_i)$ be two oriented atlass of Ω. Suppose the two atlass have the same orientation. Then if $\int_\Omega \omega$ is computed with respect to the two atlass the same number is obtained. Assume each a_I is Borel measurable and bounded.*[1]

Proof: Let $\{\psi_i\}$ be a partition of unity as described in Lemma 11.5.3 which is associated with the atlas $(U_i, \mathbf{R}_i)$ and let $\{\eta_i\}$ be a partition of unity associated in the same manner with the atlas $(V_i, \mathbf{S}_i)$. Then the definition using the atlas $\{(U_i, \mathbf{R}_i)\}$ is

$$\sum_{i=1}^{p} \sum_I \int_{\mathbf{R}_i U_i} \psi_i\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) a_I\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \frac{\partial(x_{i_1} \cdots x_{i_n})}{\partial(u_1 \cdots u_n)} du \tag{11.13}$$

$$\begin{aligned}
&= \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_I \int_{\mathbf{R}_i(U_i \cap V_j)} \eta_j\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \psi_i\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) a_I\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \frac{\partial(x_{i_1} \cdots x_{i_n})}{\partial(u_1 \cdots u_n)} du \\
&= \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_I \int_{\operatorname{int} \mathbf{R}_i(U_i \cap V_j \setminus L)} \eta_j\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \psi_i\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) a_I\left(\mathbf{R}_i^{-1}(\mathbf{u})\right) \frac{\partial(x_{i_1} \cdots x_{i_n})}{\partial(u_1 \cdots u_n)} du
\end{aligned}$$

The reason this can be done is that points not on the interior of $\mathbf{R}_i(U_i \cap V_j)$ are on the plane $u_1 = 0$ which is a set of measure zero and by assumption $\mathbf{R}_i(L \cap U_i \cap V_j)$ has measure zero. Of course the above determinant in the integrand is not even defined on $\mathbf{R}_i(L \cap U_i \cap V_j)$. By the change of variables formula Theorem 9.9.10 and Proposition 11.1.7, this equals

$$\begin{aligned}
&\sum_{i=1}^{p} \sum_{j=1}^{q} \sum_I \int_{\operatorname{int} \mathbf{S}_j(U_i \cap V_j \setminus L)} \eta_j\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) \psi_i\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) a_I\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) \frac{\partial(x_{i_1} \cdots x_{i_n})}{\partial(v_1 \cdots v_n)} dv \\
&= \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_I \int_{\mathbf{S}_j(U_i \cap V_j)} \eta_j\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) \psi_i\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) a_I\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) \frac{\partial(x_{i_1} \cdots x_{i_n})}{\partial(v_1 \cdots v_n)} dv \\
&= \sum_{j=1}^{q} \sum_I \int_{\mathbf{S}_j(V_j)} \eta_j\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) a_I\left(\mathbf{S}_j^{-1}(\mathbf{v})\right) \frac{\partial(x_{i_1} \cdots x_{i_n})}{\partial(v_1 \cdots v_n)} dv
\end{aligned}$$

which is the definition $\int_\Omega \omega$ using the other atlas $\{(V_j, \mathbf{S}_j)\}$ and partition of unity. This proves the theorem. ■

[1] This is so issues of existence for the various integrals will not arrise. This is leading to Stoke's theorem in which even more will be assumed on a_I.

11.5.1 The Derivative Of A Differential Form

The derivative of a differential form is defined next.

Definition 11.5.7 *Let $\omega = \sum_I a_I(\mathbf{x})\, dx_{i_1} \wedge \cdots \wedge dx_{i_{n-1}}$ be a differential form of order $n-1$ where a_I is C^1. Then define $d\omega$, a differential form of order n by replacing $a_I(\mathbf{x})$ with*

$$da_I(\mathbf{x}) \equiv \sum_{k=1}^{m} \frac{\partial a_I(\mathbf{x})}{\partial x_k} dx_k \tag{11.14}$$

and putting a wedge after the dx_k. Therefore,

$$d\omega \equiv \sum_I \sum_{k=1}^{m} \frac{\partial a_I(\mathbf{x})}{\partial x_k} dx_k \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_{n-1}}. \tag{11.15}$$

11.6 Stoke's Theorem And The Orientation Of ∂Ω

Here Ω will be an n dimensional orientable PC^1 manifold with boundary in $\mathbb{R}^m$. Let an oriented atlas for it be $\{U_i, \mathbf{R}_i\}_{i=1}^p$ and let a C^∞ partition of unity be $\{\psi_i\}_{i=1}^p$. Also let

$$\omega = \sum_I a_I(\mathbf{x})\, dx_{i_1} \wedge \cdots \wedge dx_{i_{n-1}}$$

be a differential form such that a_I is $C^1\left(\overline{\Omega}\right)$. Since $\sum \psi_i(\mathbf{x}) = 1$ on Ω,

$$d\omega \equiv \sum_I \sum_{k=1}^{m} \frac{\partial a_I(\mathbf{x})}{\partial x_k} dx_k \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_{n-1}}$$

$$= \sum_I \sum_{k=1}^{m} \sum_{j=1}^{p} \frac{\partial\left(\psi_j a_I\right)}{\partial x_k}(\mathbf{x})\, dx_k \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_{n-1}}$$

because the right side equals

$$\sum_I \sum_{k=1}^{m} \sum_{j=1}^{p} \frac{\partial \psi_j}{\partial x_k} a_I(\mathbf{x})\, dx_k \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_{n-1}}$$

$$+ \sum_I \sum_{k=1}^{m} \sum_{j=1}^{p} \frac{\partial a_I}{\partial x_k} \psi_j(\mathbf{x})\, dx_k \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_{n-1}}$$

$$= \sum_I a_I(\mathbf{x}) \sum_{k=1}^{m} \frac{\partial}{\partial x_k} \left(\overbrace{\sum_{j=1}^{p} \psi_j}^{=1} \right) dx_k \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_{n-1}}$$

$$+ \sum_I \sum_{k=1}^{m} \frac{\partial a_I}{\partial x_k} \sum_{j=1}^{p} \psi_j(\mathbf{x})\, dx_k \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_{n-1}}$$

$$= \sum_I \sum_{k=1}^{m} \frac{\partial a_I}{\partial x_k}(\mathbf{x})\, dx_k \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_{n-1}} \equiv d\omega$$

It follows

$$\int d\omega = \sum_I \sum_{k=1}^{m} \sum_{j=1}^{p} \int_{\mathbf{R}_j(U_j)} \frac{\partial\left(\psi_j a_I\right)}{\partial x_k}\left(\mathbf{R}_j^{-1}(\mathbf{u})\right) \frac{\partial\left(x_k, x_{i_1} \cdots x_{i_{n-1}}\right)}{\partial(u_1, \cdots, u_n)} d\mathbf{u}$$

$$= \sum_I \sum_{k=1}^m \sum_{j=1}^p \int_{\mathbf{R}_j(U_j)} \frac{\partial\left(\psi_j a_I\right)}{\partial x_k}\left(\mathbf{R}_{j\varepsilon}^{-1}(\mathbf{u})\right) \frac{\partial\left(x_{k\varepsilon}, x_{i_1\varepsilon}\cdots x_{i_{n-1}\varepsilon}\right)}{\partial\left(u_1,\cdots,u_n\right)} d\mathbf{u} +$$

$$\sum_I \sum_{k=1}^m \sum_{j=1}^p \int_{\mathbf{R}_j(U_j)} \frac{\partial\left(\psi_j a_I\right)}{\partial x_k}\left(\mathbf{R}_j^{-1}(\mathbf{u})\right) \frac{\partial\left(x_k, x_{i_1}\cdots x_{i_{n-1}}\right)}{\partial\left(u_1,\cdots,u_n\right)} d\mathbf{u}$$

$$- \sum_I \sum_{k=1}^m \sum_{j=1}^p \int_{\mathbf{R}_j(U_j)} \frac{\partial\left(\psi_j a_I\right)}{\partial x_k}\left(\mathbf{R}_{j\varepsilon}^{-1}(\mathbf{u})\right) \frac{\partial\left(x_{k\varepsilon}, x_{i_1\varepsilon}\cdots x_{i_{n-1}\varepsilon}\right)}{\partial\left(u_1,\cdots,u_n\right)} d\mathbf{u} \tag{11.16}$$

Here

$$\mathbf{R}_{j\varepsilon}^{-1}(\mathbf{u}) \equiv \mathbf{R}_j^{-1} * \phi_\varepsilon(\mathbf{u})$$

for ϕ_ε a mollifier and $x_{i\varepsilon}$ is the i^{th} component mollified. Thus by Lemma 9.5.7, this function with the subscript ε is infinitely differentiable. The last two expressions in 11.16 sum to $e(\varepsilon)$ which converges to 0 as $\varepsilon \to 0$. Here is why.

$$\frac{\partial\left(\psi_j a_I\right)}{\partial x_k}\left(\mathbf{R}_{j\varepsilon}^{-1}(\mathbf{u})\right) \to \frac{\partial\left(\psi_j a_I\right)}{\partial x_k}\left(\mathbf{R}_j^{-1}(\mathbf{u})\right) \ a.e.$$

because of the pointwise convergence of $\mathbf{R}_{j\varepsilon}^{-1}$ to $\mathbf{R}_j^{-1}$ which follows from Lemma 9.5.7. In addition to this,

$$\frac{\partial\left(x_{k\varepsilon}, x_{i_1\varepsilon}\cdots x_{i_{n-1}\varepsilon}\right)}{\partial\left(u_1,\cdots,u_n\right)} \to \frac{\partial\left(x_k, x_{i_1}\cdots x_{i_{n-1}}\right)}{\partial\left(u_1,\cdots,u_n\right)} \ a.e.$$

because of this lemma used again on each of the component functions. This convergence happens on $\mathbf{R}_j(U_j \setminus L)$ for each j. Thus $e(\varepsilon)$ is a finite sum of integrals of integrands which converges to 0 $a.e.$ By assumption 11.4, these integrands are uniformly integrable and so it follows from the Vitali convergence theorem, Theorem 11.2.5, the integrals converge to 0.

Then 11.16 equals

$$= \sum_I \sum_{k=1}^m \sum_{j=1}^p \int_{\mathbf{R}_j(U_j)} \frac{\partial\left(\psi_j a_I\right)}{\partial x_k}\left(\mathbf{R}_{j\varepsilon}^{-1}(\mathbf{u})\right) \sum_{l=1}^m \frac{\partial x_{k\varepsilon}}{\partial u_l} A_{1l} d\mathbf{u} + e(\varepsilon)$$

where A_{1l} is the $1l^{th}$ cofactor for the determinant

$$\frac{\partial\left(x_{k\varepsilon}, x_{i_1\varepsilon}\cdots x_{i_{n-1}\varepsilon}\right)}{\partial\left(u_1,\cdots,u_n\right)}$$

which is determined by a particular I. I am suppressing the ε and I for the sake of notation. Then the above reduces to

$$= \sum_I \sum_{j=1}^p \int_{\mathbf{R}_j(U_j)} \sum_{l=1}^n A_{1l} \sum_{k=1}^m \frac{\partial\left(\psi_j a_I\right)}{\partial x_k}\left(\mathbf{R}_{j\varepsilon}^{-1}(\mathbf{u})\right) \frac{\partial x_{k\varepsilon}}{\partial u_l} d\mathbf{u} + e(\varepsilon)$$

$$= \sum_I \sum_{j=1}^p \sum_{l=1}^n \int_{\mathbf{R}_j(U_j)} A_{1l} \frac{\partial}{\partial u_l}\left(\psi_j a_I \circ \mathbf{R}_{j\varepsilon}^{-1}\right)(\mathbf{u})\, d\mathbf{u} + e(\varepsilon) \tag{11.17}$$

(Note l goes up to n not m.) Recall $\mathbf{R}_j(U_j)$ is relatively open in $\mathbb{R}^n_{\leq}$. Consider the integral where $l > 1$. Integrate first with respect to u_l. In this case the boundary term vanishes because of ψ_j and you get

$$-\int_{\mathbf{R}_j(U_j)} A_{1l,l}\left(\psi_j a_I \circ \mathbf{R}_{j\varepsilon}^{-1}\right)(\mathbf{u})\, d\mathbf{u} \tag{11.18}$$

Next consider the case where $l = 1$. Integrating first with respect to u_1, the term reduces to

$$\int_{\mathbf{R}_j V_j} \psi_j a_I \circ \mathbf{R}_{j\varepsilon}^{-1}(0, u_2, \cdots, u_n)\, A_{11} d\mathbf{u}_1 - \int_{\mathbf{R}_j(U_j)} A_{11,1}\left(\psi_j a_I \circ \mathbf{R}_{j\varepsilon}^{-1}\right)(\mathbf{u})\, d\mathbf{u} \quad (11.19)$$

where $\mathbf{R}_j V_j$ is an open set in $\mathbb{R}^{n-1}$ consisting of

$$\left\{(u_2, \cdots, u_n) \in \mathbb{R}^{n-1} : (0, u_2, \cdots, u_n) \in \mathbf{R}_j(U_j)\right\}$$

and $d\mathbf{u}_1$ represents $du_2 du_3 \cdots du_n$ on $\mathbf{R}_j V_j$ for short. Thus V_j is just the part of $\partial\Omega$ which is in U_j and the mappings $\mathbf{S}_j^{-1}$ given on $\mathbf{R}_j V_j = \mathbf{R}_j(U_j \cap \partial\Omega)$ by

$$\mathbf{S}_j^{-1}(u_2, \cdots, u_n) \equiv \mathbf{R}_j^{-1}(0, u_2, \cdots, u_n)$$

are such that $\{(\mathbf{S}_j, V_j)\}$ is an atlas for $\partial\Omega$. Then if 11.18 and 11.19 are placed in 11.17, it follows from Lemma 11.5.1 that 11.17 reduces to

$$\sum_I \sum_{j=1}^p \int_{\mathbf{R}_j V_j} \psi_j a_I \circ \mathbf{R}_{j\varepsilon}^{-1}(0, u_2, \cdots, u_n)\, A_{11} d\mathbf{u}_1 + e(\varepsilon)$$

Now as before, each $\partial x_{s\varepsilon}/\partial u_r$ converges pointwise *a.e.* to $\partial x_s/\partial u_r$, off $\mathbf{R}_j(V_j \cap L)$ assumed to be a set of measure zero, and the integrands are bounded. Using the Vitali convergence theorem again, pass to a limit as $\varepsilon \to 0$ to obtain

$$\begin{aligned} &\sum_I \sum_{j=1}^p \int_{\mathbf{R}_j V_j} \psi_j a_I \circ \mathbf{R}_j^{-1}(0, u_2, \cdots, u_n)\, A_{11} d\mathbf{u}_1 \\ = &\sum_I \sum_{j=1}^p \int_{\mathbf{S}_j V_j} \psi_j a_I \circ \mathbf{S}_j^{-1}(u_2, \cdots, u_n)\, A_{11} d\mathbf{u}_1 \end{aligned}$$

$$= \sum_I \sum_{j=1}^p \int_{\mathbf{S}_j V_j} \psi_j a_I \circ \mathbf{S}_j^{-1}(u_2, \cdots, u_n)\, \frac{\partial\left(x_{i_1} \cdots x_{i_{n-1}}\right)}{\partial(u_2, \cdots, u_n)}(0, u_2, \cdots, u_n)\, d\mathbf{u}_1 \quad (11.20)$$

This of course is the definition of $\int_{\partial\Omega} \omega$ provided $\{\mathbf{S}_j, V_j\}$ is an oriented atlas. Note the integral is well defined because of the assumption that $\mathbf{R}_i(L \cap U_i \cap \partial\Omega)$ has m_{n-1} measure zero. That $\partial\Omega$ is orientable and that this atlas is an oriented atlas is shown next. I will write $\mathbf{u}_1 \equiv (u_2, \cdots, u_n)$.

What if $\operatorname{spt} a_I \subseteq K \subseteq U_i \cap U_j$ for each I? Then using Lemma 11.5.3 it follows that $\int d\omega =$

$$\sum_I \int_{\mathbf{S}_j(V_j \cap V_j)} a_I \circ \mathbf{S}_j^{-1}(u_2, \cdots, u_n)\, \frac{\partial\left(x_{i_1} \cdots x_{i_{n-1}}\right)}{\partial(u_2, \cdots, u_n)}(0, u_2, \cdots, u_n)\, d\mathbf{u}_1$$

This is done by using a partition of unity which has the property that ψ_j equals 1 on K which forces all the other ψ_k to equal zero there. Using the same trick involving a judicious choice of the partition of unity, $\int d\omega$ is also equal to

$$\sum_I \int_{\mathbf{S}_i(V_j \cap V_j)} a_I \circ \mathbf{S}_i^{-1}(v_2, \cdots, v_n)\, \frac{\partial\left(x_{i_1} \cdots x_{i_{n-1}}\right)}{\partial(v_2, \cdots, v_n)}(0, v_2, \cdots, v_n)\, d\mathbf{v}_1$$

Since $\mathbf{S}_i(L \cap U_i), \mathbf{S}_j(L \cap U_j)$ have measure zero, the above integrals may be taken over

$$\mathbf{S}_j(V_j \cap V_j \setminus L), \mathbf{S}_i(V_j \cap V_j \setminus L)$$

respectively. Also these are equal, both being $\int d\omega$. To simplify the notation, let π_I denote the projection onto the components corresponding to I. Thus if $I = (i_1, \cdots, i_n)$,

$$\pi_I \mathbf{x} \equiv (x_{i_1}, \cdots, x_{i_n}).$$

Then writing in this simpler notation, the above would say

$$\sum_I \int_{\mathbf{S}_j(V_j \cap V_j \setminus L)} a_I \circ \mathbf{S}_j^{-1}(\mathbf{u}_1) \det D\pi_I \mathbf{S}_j^{-1}(\mathbf{u}_1)\, d\mathbf{u}_1$$

$$= \sum_I \int_{\mathbf{S}_i(V_j \cap V_j \setminus L)} a_I \circ \mathbf{S}_i^{-1}(\mathbf{v}_1) \det D\pi_I \mathbf{S}_i^{-1}(\mathbf{v}_1)\, d\mathbf{v}_1$$

and both equal to $\int d\omega$. Thus using the change of variables formula, Theorem 9.9.10, it follows the second of these equals

$$\sum_I \int_{\mathbf{S}_j(V_j \cap V_j \setminus L)} a_I \circ \mathbf{S}_j^{-1}(\mathbf{u}_1) \det D\pi_I \mathbf{S}_i^{-1}\left(\mathbf{S}_i \circ \mathbf{S}_j^{-1}(\mathbf{u}_1)\right) \left|\det D\left(\mathbf{S}_i \circ \mathbf{S}_j^{-1}\right)(\mathbf{u}_1)\right| d\mathbf{u}_1 \tag{11.21}$$

I want to argue $\det D\left(\mathbf{S}_i \circ \mathbf{S}_j^{-1}\right)(\mathbf{u}_1) \geq 0$. Let A be the open subset of $\mathbf{S}_j(V_j \cap V_j \setminus L)$ on which for $\delta > 0$,

$$\det D\left(\mathbf{S}_i \circ \mathbf{S}_j^{-1}\right)(\mathbf{u}_1) < -\delta \tag{11.22}$$

I want to show $A = \emptyset$ so assume A is nonempty. If this is the case, we could consider an open ball contained in A. To simplify notation, assume A is an open ball. Letting f_I be a smooth function which vanishes off a compact subset of $\mathbf{S}_j^{-1}(A)$ the above argument and the chain rule imply

$$\begin{aligned} & \sum_I \int_{\mathbf{S}_j(V_j \cap V_j \setminus L)} f_I \circ \mathbf{S}_j^{-1}(\mathbf{u}_1) \det D\pi_I \mathbf{S}_j^{-1}(\mathbf{u}_1)\, d\mathbf{u}_1 \\ = & \sum_I \int_A f_I \circ \mathbf{S}_j^{-1}(\mathbf{u}_1) \det D\pi_I \mathbf{S}_j^{-1}(\mathbf{u}_1)\, d\mathbf{u}_1 \end{aligned}$$

$$\sum_I \int_A f_I \circ \mathbf{S}_j^{-1}(\mathbf{u}_1) \det D\pi_I \mathbf{S}_i^{-1}\left(\mathbf{S}_i \circ \mathbf{S}_j^{-1}(\mathbf{u}_1)\right) \det D\left(\mathbf{S}_i \circ \mathbf{S}_j^{-1}\right)(\mathbf{u}_1)\, d\mathbf{u}_1$$

Now from 11.21, this equals

$$= -\sum_I \int_A f_I \circ \mathbf{S}_j^{-1}(\mathbf{u}_1) \det D\pi_I \mathbf{S}_i^{-1}\left(\mathbf{S}_i \circ \mathbf{S}_j^{-1}(\mathbf{u}_1)\right) \det D\left(\mathbf{S}_i \circ \mathbf{S}_j^{-1}\right)(\mathbf{u}_1)\, d\mathbf{u}_1$$

and consequently

$$0 = 2\sum_I \int_A f_I \circ \mathbf{S}_j^{-1}(\mathbf{u}_1) \det D\pi_I \mathbf{S}_i^{-1}\left(\mathbf{S}_i \circ \mathbf{S}_j^{-1}(\mathbf{u}_1)\right) \det D\left(\mathbf{S}_i \circ \mathbf{S}_j^{-1}\right)(\mathbf{u}_1)\, d\mathbf{u}_1$$

Now for each I, let $\left\{f_I^k \circ \mathbf{S}_j^{-1}\right\}_{k=1}^{\infty}$ be a sequence of bounded functions having compact support in A which converge pointwise to $\det D\pi_I \mathbf{S}_i^{-1}\left(\mathbf{S}_i \circ \mathbf{S}_j^{-1}(\mathbf{u}_1)\right)$. Then it follows from the Vitali convergence theorem, one can pass to the limit and obtain

$$\begin{aligned} 0 &= 2\int_A \sum_I \left(\det D\pi_I \mathbf{S}_i^{-1}\left(\mathbf{S}_i \circ \mathbf{S}_j^{-1}(\mathbf{u}_1)\right)\right)^2 \det D\left(\mathbf{S}_i \circ \mathbf{S}_j^{-1}\right)(\mathbf{u}_1)\, d\mathbf{u}_1 \\ &\leq -2\delta \int_A \sum_I \left(\det D\pi_I \mathbf{S}_i^{-1}\left(\mathbf{S}_i \circ \mathbf{S}_j^{-1}(\mathbf{u}_1)\right)\right)^2 d\mathbf{u}_1 \end{aligned}$$

Since the integrand is continuous, this would require

$$\det D\pi_I \mathbf{S}_i^{-1}(\mathbf{v}_1) \equiv 0 \tag{11.23}$$

for each I and for each $\mathbf{v}_1 \in \mathbf{S}_i \circ \mathbf{S}_j^{-1}(A)$, an open set which must have positive measure. But since it has positive measure, it follows from the change of variables theorem and the chain rule,

$$D\left(\mathbf{S}_j \circ \mathbf{S}_i^{-1}\right)(\mathbf{v}_1) = \overbrace{D\mathbf{S}_j\left(\mathbf{S}_i^{-1}(\mathbf{v}_1)\right)}^{n\times m}\overbrace{D\mathbf{S}_i^{-1}(\mathbf{v}_1)}^{m\times n}$$

cannot be identically 0. By the Binet Cauchy theorem, at least some

$$D\pi_I \mathbf{S}_i^{-1}(\mathbf{v}_1) \neq 0$$

contradicting 11.23. Thus $A = \emptyset$ and since $\delta > 0$ was arbitrary, this shows

$$\det D\left(\mathbf{S}_i \circ \mathbf{S}_j^{-1}\right)(\mathbf{u}_1) \geq 0.$$

Hence this is an oriented atlas as claimed. This proves the theorem. ■

Theorem 11.6.1 *Let Ω be an oriented PC^1 manifold and let*

$$\omega = \sum_I a_I(\mathbf{x})\, dx_{i_1} \wedge \cdots \wedge dx_{i_{n-1}}.$$

where each a_I is $C^1\left(\overline{\Omega}\right)$. For $\{U_j, \mathbf{R}_j\}_{j=0}^p$ an oriented atlas for Ω where $\mathbf{R}_j(U_j)$ is a relatively open set in

$$\{\mathbf{u} \in \mathbb{R}^n : u_1 \leq 0\},$$

define an atlas for $\partial\Omega, \{V_j, \mathbf{S}_j\}$ where $V_j \equiv \partial\Omega \cap U_j$ and $\mathbf{S}_j$ is just the restriction of $\mathbf{R}_j$ to V_j. Then this is an oriented atlas for $\partial\Omega$ and

$$\int_{\partial\Omega} \omega = \int_\Omega d\omega$$

where the two integrals are taken with respect to the given oriented atlass.

11.7 Green's Theorem, An Example

Green's theorem is a well known result in calculus and it pertains to a region in the plane. I am going to generalize to an open set in $\mathbb{R}^n$with sufficiently smooth boundary using the methods of differential forms described above.

11.7.1 An Oriented Manifold

A bounded open subset Ω, of $\mathbb{R}^n, n \geq 2$ has PC^1 boundary and lies locally on one side of its boundary if it satisfies the following conditions.

For each $p \in \partial\Omega \equiv \overline{\Omega} \setminus \Omega$, there exists an open set, Q, containing p, an open interval (a, b), a bounded open set $B \subseteq \mathbb{R}^{n-1}$, and an orthogonal transformation R such that $\det R = 1$,

$$(a, b) \times B = RQ,$$

and letting $W = Q \cap \Omega$,

$$RW = \{\mathbf{u} \in \mathbb{R}^n : a < u_1 < g(u_2, \cdots, u_n), (u_2, \cdots, u_n) \in B\}$$

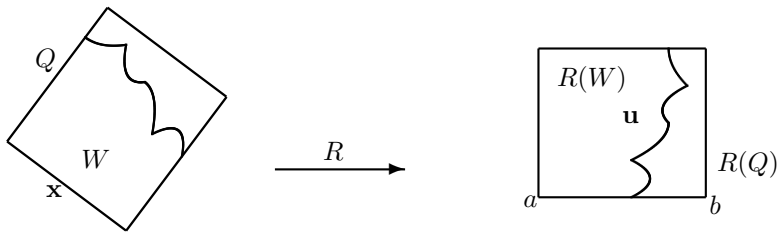
$g(u_2,\cdots,u_n) < b$ for $(u_2,\cdots,u_n) \in B$. Also g vanishes outside some compact set in $\mathbb{R}^{n-1}$ and g is continuous.

$$R(\partial\Omega \cap Q) = \{\mathbf{u} \in \mathbb{R}^n : u_1 = g(u_2,\cdots,u_n), (u_2,\cdots,u_n) \in B\}.$$

Note that finitely many of these sets Q cover $\partial\Omega$ because $\partial\Omega$ is compact. Assume there exists a closed subset of $\partial\Omega, L$ such that the closed set S_Q defined by

$$\{(u_2,\cdots,u_n) \in B : (g(u_2,\cdots,u_n), u_2,\cdots,u_n) \in R(L \cap Q)\} \tag{11.24}$$

has m_{n-1} measure zero. $g \in C^1(B \setminus S_Q)$ and all the partial derivatives of g are uniformly bounded on $B \setminus S_Q$. The following picture describes the situation. The pointy places symbolize the set L.



Define $\mathbf{P}_1 : \mathbb{R}^n \to \mathbb{R}^{n-1}$ by

$$\mathbf{P}_1\mathbf{u} \equiv (u_2,\cdots,u_n)$$

and $\Sigma : \mathbb{R}^n \to \mathbb{R}^n$ given by

$$\begin{aligned}\Sigma\mathbf{u} &\equiv \mathbf{u} - g(\mathbf{P}_1\mathbf{u})\mathbf{e}_1\\ &\equiv \mathbf{u} - g(u_2,\cdots,u_n)\mathbf{e}_1\\ &\equiv (u_1 - g(u_2,\cdots,u_n), u_2,\cdots,u_n)\end{aligned}$$

Thus Σ is invertible and

$$\begin{aligned}\Sigma^{-1}\mathbf{u} &= \mathbf{u} + g(\mathbf{P}_1\mathbf{u})\mathbf{e}_1\\ &\equiv (u_1 + g(u_2,\cdots,u_n), u_2,\cdots,u_n)\end{aligned}$$

For $\mathbf{x} \in \partial\Omega \cap Q$, it follows the first component of $R\mathbf{x}$ is $g(\mathbf{P}_1(R\mathbf{x}))$. Now define $\mathbf{R} : W \to \mathbb{R}^n_{\leq}$ as

$$\mathbf{u} \equiv \mathbf{R}\mathbf{x} \equiv R\mathbf{x} - g(\mathbf{P}_1(R\mathbf{x}))\mathbf{e}_1 \equiv \Sigma R\mathbf{x}$$

and so it follows

$$\mathbf{R}^{-1} = R^*\Sigma^{-1}.$$

These mappings $\mathbf{R}$ involve first a rotation followed by a variable sheer in the direction of the u_1 axis. From the above description, $\mathbf{R}(L \cap Q) = 0 \times S_Q$, a set of m_{n-1} measure zero. This is because

$$(u_2,\cdots,u_n) \in S_Q$$

if and only if

$$(g(u_2,\cdots,u_n), u_2,\cdots,u_n) \in R(L \cap Q)$$

if and only if

$$\begin{aligned}(0, u_2,\cdots,u_n) &\equiv \Sigma(g(u_2,\cdots,u_n), u_2,\cdots,u_n)\\ &\in \Sigma R(L \cap Q) \equiv \mathbf{R}(L \cap Q).\end{aligned}$$

Since $\partial\Omega$ is compact, there are finitely many of these open sets $Q_1,\cdots,Q_p$ which cover $\partial\Omega$. Let the orthogonal transformations and other quantities described above

also be indexed by k for $k = 1, \cdots, p$. Also let Q_0 be an open set with $\overline{Q_0} \subseteq \Omega$ and $\overline{\Omega}$ is covered by $Q_0, Q_1, \cdots, Q_p$. Let $\mathbf{u} \equiv \mathbf{R}_0\mathbf{x} \equiv \mathbf{x} - k\mathbf{e}_1$ where k is large enough that $\mathbf{R}_0 Q_0 \subseteq \mathbb{R}^n_<$. Thus in this case, the orthogonal transformation R_0 equals I and $\Sigma_0\mathbf{x} \equiv \mathbf{x} - k\mathbf{e}_1$. I claim Ω is an oriented manifold with boundary and the charts are $(W_i, \mathbf{R}_i)$.

To see this is an oriented atlas for the manifold, note that for a.e. points of $\mathbf{R}_i(W_i \cap W_j)$ the function g is differentiable. Then using the above notation, at these points $\mathbf{R}_j \circ \mathbf{R}_i^{-1}$ is of the form

$$\Sigma_j R_j R_i^* \Sigma_i^{-1}$$

and it is a one to one mapping. What is the determinant of its derivative? By the chain rule,

$$D\left(\Sigma_j R_j R_i^* \Sigma_i^{-1}\right) = D\Sigma_j\left(R_j R_i^* \Sigma_i^{-1}\right) DR_j\left(R_i^* \Sigma_i^{-1}\right) DR_i^*\left(\Sigma_i^{-1}\right) D\Sigma_i^{-1}$$

However,

$$\det(D\Sigma_j) = 1 = \det\left(D\Sigma_j^{-1}\right)$$

and $\det(R_i) = \det(R_i^*) = 1$ by assumption. Therefore, for a.e. $\mathbf{u} \in \left(\mathbf{R}_j \circ \mathbf{R}_i^{-1}\right)(A)$,

$$\det\left(D\left(\mathbf{R}_j \circ \mathbf{R}_i^{-1}\right)(\mathbf{u})\right) > 0.$$

By Proposition 11.1.7 Ω is indeed an oriented manifold with the given atlas.

11.7.2 Green's Theorem

The general Green's theorem is the following. It follows from Stoke's theorem above. Theorem 11.6.1.

First note that since $\Omega \subseteq \mathbb{R}^n$, there is no loss of generality in writing

$$\omega = \sum_{k=1}^n a_k(\mathbf{x})\, dx_1 \wedge \cdots \wedge \widehat{dx_k} \wedge \cdots \wedge dx_n$$

where the hat indicates the dx_k is omitted. Therefore,

$$d\omega = \sum_{k=1}^n \sum_{j=1}^n \frac{\partial a_k(\mathbf{x})}{\partial x_j} dx_j \wedge dx_1 \wedge \cdots \wedge \widehat{dx_k} \wedge \cdots \wedge dx_n$$

$$= \sum_{k=1}^n \frac{\partial a_k(\mathbf{x})}{\partial x_k} (-1)^{k-1}\, dx_1 \wedge \cdots \wedge dx_n$$

This follows from the definition of integration of differential forms. If there is a repeat in the dx_j then this will lead to a determinant of a matrix which has two equal columns in the definition of the integral of the differential form. Also, from the definition, and again from the properties of the determinant, when you switch two dx_k it changes the sign because it is equivalent to switching two columns in a determinant. Then with these observations, Green's theorem follows.

Theorem 11.7.1 *Let Ω be a bounded open set in $\mathbb{R}^n, n \geq 2$ and let it have PC^1 boundary and lie locally on one side of its boundary as described above. Also let*

$$\omega = \sum_{k=1}^n a_k(\mathbf{x})\, dx_1 \wedge \cdots \wedge \widehat{dx_k} \wedge \cdots \wedge dx_n$$

be a differential form where a_I is assumed to be in $C^1\left(\overline{\Omega}\right)$. Then

$$\int_{\partial\Omega} \omega = \int_\Omega \sum_{k=1}^n \frac{\partial a_k(\mathbf{x})}{\partial x_k} (-1)^{k-1}\, dm_n$$

Proof: From the definition and using the usual technique of ignoring the exceptional set of measure zero,

$$\int_\Omega d\omega \equiv \sum_{i=1}^{p}\sum_{k=1}^{n}\int_{\mathbf{R}_i W_i} \frac{\partial a_k\left(\mathbf{R}_i^{-1}(\mathbf{u})\right)}{\partial x_k}(-1)^{k-1}\psi_i\left(\mathbf{R}_i^{-1}(\mathbf{u})\right)\frac{\partial(x_1\cdots x_n)}{\partial(u_1\cdots u_n)}du$$

Now from the above description of $\mathbf{R}_i^{-1}$, the determinant in the above integrand equals 1. Therefore, the change of variables theorem applies and the above reduces to

$$\begin{aligned}\sum_{i=1}^{p}\sum_{k=1}^{n}\int_{W_i}\frac{\partial a_k(\mathbf{x})}{\partial x_k}(-1)^{k-1}\psi_i(\mathbf{x})\,dx &= \int_\Omega \sum_{i=1}^{p}\sum_{k=1}^{n}\frac{\partial a_k(\mathbf{x})}{\partial x_k}(-1)^{k-1}\psi_i(\mathbf{x})\,dm_n \\ &= \int_\Omega \sum_{k=1}^{n}\frac{\partial a_k(\mathbf{x})}{\partial x_k}(-1)^{k-1}\,dm_n\end{aligned}$$

This proves the theorem. ■

This Green's theorem may appear very general because it is an n dimensional theorem. However, the best versions of this theorem in the plane are considerably more general in terms of smoothness of the boundary. Later what is probably the best Green's theorem is discussed. The following is a specialization to the familiar calculus theorem.

Example 11.7.2 *The usual Green's theorem follows from the above specialized to the case of* $n = 2$.

$$\int_{\partial\Omega} P(x,y)\,dx + Q(x,y)\,dy = \int_\Omega (Q_x - P_y)\,dxdy$$

This follows because the differential form on the left is of the form

$$Pdx\wedge\widehat{dy} + Q\widehat{dx}\wedge dy$$

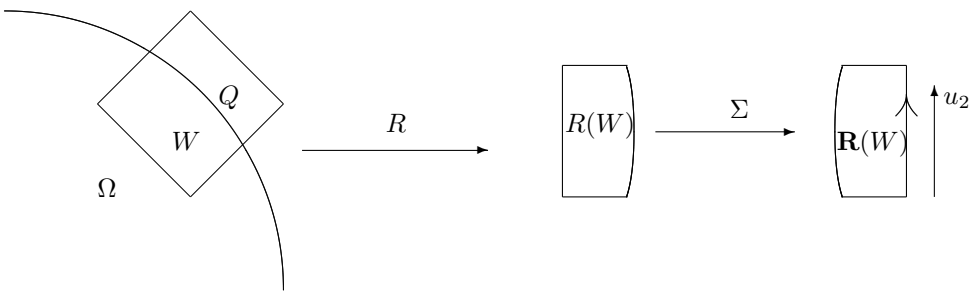
and so, as above, the derivative of this is

$$P_y dy\wedge dx + Q_x dx\wedge dy = (Q_x - P_y)\,dx\wedge dy$$

It is understood in all the above that the oriented atlas for Ω is the one described there where

$$\frac{\partial(x,y)}{\partial(u_1,u_2)} = 1.$$

Saying Ω is PC^1 reduces in this case to saying $\partial\Omega$ is a piecewise C^1 closed curve which is often the case stated in calculus courses. From calculus, the orientation of $\partial\Omega$ was defined not in the abstract manner described above but by saying that motion around the curve takes place in the counter clockwise direction. This is really a little vague although later it will be made very precise. However, it is not hard to see that this is what is taking place in the above formulation. To do this, consider the following pictures representing first the rotation and then the shear which were used to specify an atlas for Ω.



The vertical arrow at the end indicates the direction of increasing u_2. The vertical side of $\mathbf{R}(W)$ shown there corresponds to the curved side in $R(W)$ which corresponds to the part of $\partial\Omega$ which is selected by Q as shown in the picture. Here R is an orthogonal transformation which has determinant equal to 1. Now the shear which goes from the diagram on the right to the one on its left preserves the direction of motion relative to the surface the curve is bounding. This is geometrically clear. Similarly, the orthogonal transformation R^* which goes from the curved part of the boundary of $R(W)$ to the corresponding part of $\partial\Omega$ preserves the direction of motion relative to the surface. This is because orthogonal transformations in $\mathbb{R}^2$ whose determinants are 1 correspond to rotations. Thus increasing u_2 corresponds to counter clockwise motion around $\mathbf{R}(W)$ along the vertical side of $\mathbf{R}(W)$ which corresponds to counter clockwise motion around $R(W)$ along the curved side of $R(W)$ which corresponds to counter clockwise motion around Ω in the sense that the direction of motion along the curve is always such that if you were walking in this direction, your left hand would be over the surface. In other words this agrees with the usual calculus conventions.

11.8 The Divergence Theorem

From Green's theorem, one can quickly obtain a general Divergence theorem for Ω as described above in Section 11.7.1. First note that from the above description of the $\mathbf{R}_j$,

$$\frac{\partial\left(x_k, x_{i_1}, \cdots x_{i_{n-1}}\right)}{\partial\left(u_1, \cdots, u_n\right)} = \operatorname{sgn}\left(k, i_1 \cdots, i_{n-1}\right).$$

Let $\mathbf{F}(\mathbf{x})$ be a $C^1\left(\overline{\Omega}; \mathbb{R}^n\right)$ vector field. Say $\mathbf{F} = (F_1, \cdots, F_n)$. Consider the differential form

$$\omega(\mathbf{x}) \equiv \sum_{k=1}^{n} F_k(\mathbf{x}) (-1)^{k-1} dx_1 \wedge \cdots \wedge \widehat{dx_k} \wedge \cdots \wedge dx_n$$

where the hat means dx_k is being left out. Then

$$\begin{aligned} d\omega(\mathbf{x}) &= \sum_{k=1}^{n}\sum_{j=1}^{n} \frac{\partial F_k}{\partial x_j} (-1)^{k-1} dx_j \wedge dx_1 \wedge \cdots \wedge \widehat{dx_k} \wedge \cdots \wedge dx_n \\ &= \sum_{k=1}^{n} \frac{\partial F_k}{\partial x_k} dx_1 \wedge \cdots \wedge dx_k \wedge \cdots \wedge dx_n \\ &\equiv \operatorname{div}(\mathbf{F})\, dx_1 \wedge \cdots \wedge dx_k \wedge \cdots \wedge dx_n \end{aligned}$$

The assertion between the first and second lines follows right away from properties of determinants and the definition of the integral of the above wedge products in terms of determinants. From Green's theorem and the change of variables formula applied to the individual terms in the description of $\int_\Omega d\omega$

$$\int_\Omega \operatorname{div}(\mathbf{F})\, dx =$$

$$\sum_{j=1}^{p} \int_{B_j} \sum_{k=1}^{n} (-1)^{k-1} \frac{\partial\left(x_1, \cdots \widehat{x_k} \cdots, x_n\right)}{\partial\left(u_2, \cdots, u_n\right)} \left(\psi_j F_k\right) \circ \mathbf{R}_j^{-1}\left(0, u_2, \cdots, u_n\right) d\mathbf{u}_1,$$

$d\mathbf{u}_1$ short for $du_2 du_3 \cdots du_n$.

I want to write this in a more attractive manner which will give more insight. The above involves a particular partition of unity, the functions being the ψ_i. Replace $\mathbf{F}$ in

the above with $\psi_s \mathbf{F}$. Next let $\{\eta_j\}$ be a partition of unity $\eta_j \prec Q_j$ such that $\eta_s = 1$ on $\operatorname{spt}\psi_s$. This partition of unity exists by Lemma 11.5.3. Then

$$\int_\Omega \operatorname{div}(\psi_s \mathbf{F})\, dx =$$

$$\sum_{j=1}^{p} \int_{B_j} \sum_{k=1}^{n} (-1)^{k-1} \frac{\partial(x_1, \cdots \widehat{x_k} \cdots, x_n)}{\partial(u_2, \cdots, u_n)} \left(\eta_j \psi_s F_k\right) \circ \mathbf{R}_j^{-1}(0, u_2, \cdots, u_n)\, d\mathbf{u}_1$$

$$= \int_{B_s} \sum_{k=1}^{n} (-1)^{k-1} \frac{\partial(x_1, \cdots \widehat{x_k} \cdots, x_n)}{\partial(u_2, \cdots, u_n)} (\psi_s F_k) \circ \mathbf{R}_s^{-1}(0, u_2, \cdots, u_n)\, d\mathbf{u}_1 \tag{11.25}$$

because since $\eta_s = 1$ on $\operatorname{spt}\psi_s$, it follows all the other η_j equal zero there.

Consider the vector $\mathbf{N}$ defined for $\mathbf{u}_1 \in \mathbf{R}_s(W_s \setminus L) \cap \mathbb{R}_0^n$ whose k^{th} component is

$$N_k = (-1)^{k-1} \frac{\partial(x_1, \cdots \widehat{x_k} \cdots, x_n)}{\partial(u_2, \cdots, u_n)} = (-1)^{k+1} \frac{\partial(x_1, \cdots \widehat{x_k} \cdots, x_n)}{\partial(u_2, \cdots, u_n)} \tag{11.26}$$

Suppose you dot this vector with a tangent vector $\partial \mathbf{R}_s^{-1}/\partial u_i$. This yields

$$\sum_k (-1)^{k+1} \frac{\partial(x_1, \cdots \widehat{x_k} \cdots, x_n)}{\partial(u_2, \cdots, u_n)} \frac{\partial x_k}{\partial u_i} = 0$$

because it is the expansion of

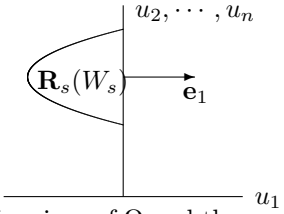
$$\begin{vmatrix} x_{1,i} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,i} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,i} & x_{n,2} & \cdots & x_{n,n} \end{vmatrix},$$

a determinant with two equal columns provided $i \geq 2$. Thus this vector is at least in some sense normal to $\partial\Omega$. If $i = 1$, then the above dot product is just

$$\frac{\partial(x_1 \cdots x_n)}{\partial(u_1 \cdots u_n)} = 1$$

This vector is called an exterior normal.

The important thing is the existence of the vector, but does it deserve to be called an **exterior** normal? Consider the following picture of $\mathbf{R}_s(W_s)$



We got this by first doing a rotation of a piece of Ω and then a shear in the direction of $\mathbf{e}_1$. Also it was shown above

$$\mathbf{R}_s^{-1}(\mathbf{u}) = R_s^*(u_1 + g(u_2, \cdots, u_n), u_2, \cdots, u_n)^T$$

where R_s^* is a rotation, an orthogonal transformation whose determinant is 1. Letting $\mathbf{x} = \mathbf{R}_s^{-1}(\mathbf{u})$, the above discussion shows $\partial\mathbf{x}/\partial u_1 \cdot \mathbf{N} = 1 > 0$. Thus it is also the case that for h small and positive,

$$\overbrace{\frac{\mathbf{x}(\mathbf{u} - h\mathbf{e}_1) - \mathbf{x}(\mathbf{u})}{h}}^{\approx -\partial\mathbf{x}/\partial u_1} \cdot \mathbf{N} < \mathbf{0}$$

Hence if θ is the angle between $\mathbf{N}$ and $\frac{\mathbf{x}(\mathbf{u}-h\mathbf{e}_1)-\mathbf{x}(\mathbf{u})}{h}$, it must be the case that $\theta > \pi/2$. However,

$$\frac{\mathbf{x}(\mathbf{u}-h\mathbf{e}_1)-\mathbf{x}(\mathbf{u})}{h}$$

points in to Ω for small $h > 0$ because $\mathbf{x}(\mathbf{u}-h\mathbf{e}_1) \in W_s$ while $\mathbf{x}(\mathbf{u})$ is on the boundary of W_s. Therefore, $\mathbf{N}$ should be pointing away from Ω at least at the points where $\mathbf{u} \to \mathbf{x}(\mathbf{u})$ is differentiable. Thus it is geometrically reasonable to use the word exterior on this vector $\mathbf{N}$.

One could normalize $\mathbf{N}$ given in 11.26 by dividing by its magnitude. Then it would be the **unit** exterior normal $\mathbf{n}$. The norm of this vector is

$$\left(\sum_{k=1}^{n}\left(\frac{\partial(x_1,\cdots\widehat{x_k}\cdots,x_n)}{\partial(u_2,\cdots,u_n)}\right)^2\right)^{1/2}$$

and by the Binet Cauchy theorem this equals

$$\det\left(D\mathbf{R}_s^{-1}(\mathbf{u}_1)^* D\mathbf{R}_s^{-1}(\mathbf{u}_1)\right)^{1/2} \equiv J(\mathbf{u}_1)$$

where as usual $\mathbf{u}_1 = (u_2,\cdots,u_n)$. Thus the expression in 11.25 reduces to

$$\int_{B_s}\left(\psi_s\mathbf{F}\circ\mathbf{R}_s^{-1}(\mathbf{u}_1)\right)\cdot\mathbf{n}\left(\mathbf{R}_s^{-1}(\mathbf{u}_1)\right)J(\mathbf{u}_1)\,d\mathbf{u}_1.$$

The integrand

$$\mathbf{u}_1 \to \left(\psi_s\mathbf{F}\circ\mathbf{R}_s^{-1}(\mathbf{u}_1)\right)\cdot\mathbf{n}\left(\mathbf{R}_s^{-1}(\mathbf{u}_1)\right)$$

is Borel measurable and bounded. Writing as a sum of positive and negative parts and using Theorem 7.7.12, there exists a sequence of bounded simple functions $\{s_k\}$ which converges pointwise a.e. to this function. Also the resulting integrands are uniformly integrable. Then by the Vitali convergence theorem, and Theorem 11.4.2 applied to these approximations,

$$\begin{aligned}
&\int_{B_s}\left(\psi_s\mathbf{F}\circ\mathbf{R}_s^{-1}(\mathbf{u}_1)\right)\cdot\mathbf{n}\left(\mathbf{R}_s^{-1}(\mathbf{u}_1)\right)J(\mathbf{u}_1)\,d\mathbf{u}_1\\
=\ &\lim_{k\to\infty}\int_{B_s}s_k\left(\mathbf{R}_s\left(\mathbf{R}_s^{-1}(\mathbf{u}_1)\right)\right)J(\mathbf{u}_1)\,d\mathbf{u}_1\\
=\ &\lim_{k\to\infty}\int_{W_s\cap\partial\Omega}s_k(\mathbf{R}_s(\mathbf{x}))\,d\sigma_{n-1}\\
=\ &\int_{W_s\cap\partial\Omega}\psi_s(\mathbf{x})\mathbf{F}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x})\,d\sigma_{n-1}\\
=\ &\int_{\partial\Omega}\psi_s(\mathbf{x})\mathbf{F}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x})\,d\sigma_{n-1}
\end{aligned}$$

Recall the exceptional set on $\partial\Omega$ has σ_{n-1} measure zero. Upon summing over all s using that the ψ_s add to 1,

$$\sum_s\int_{\partial\Omega}\psi_s(\mathbf{x})\mathbf{F}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x})\,d\sigma_{n-1} = \int_{\partial\Omega}\mathbf{F}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x})\,d\sigma_{n-1}$$

On the other hand, from 11.25, the left side of the above equals

$$\begin{aligned}\sum_s \int_\Omega \operatorname{div}(\psi_s \mathbf{F})\,dx &= \sum_s \int_\Omega \sum_{i=1}^n \psi_{s,i}F_i + \psi_s F_{i,i}dx \\ &= \int_\Omega \operatorname{div}(\mathbf{F})\,dx + \sum_{i=1}^n \left(\sum_s \psi_s\right)_{,i} F_i \\ &= \int_\Omega \operatorname{div}(\mathbf{F})\,dx\end{aligned}$$

■

This proves the following general divergence theorem.

Theorem 11.8.1 *Let Ω be a bounded open set having PC^1 boundary as described above. Also let $\mathbf{F}$ be a vector field with the property that for F_k a component function of $\mathbf{F}$, $F_k \in C^1\left(\overline{\Omega};\mathbb{R}^n\right)$. Then there exists an exterior normal vector $\mathbf{n}$ which is defined σ_{n-1} a.e. (off the exceptional set L) on $\partial\Omega$ such that*

$$\int_{\partial\Omega} \mathbf{F}\cdot\mathbf{n}d\sigma_{n-1} = \int_\Omega \operatorname{div}(\mathbf{F})\,dx$$

It is worth noting that it appears that everything above will work if you relax the requirement in PC^1 which requires the partial derivatives be bounded off an exceptional set. Instead, it would suffice to say that for some $p > n$ all integrals of the form

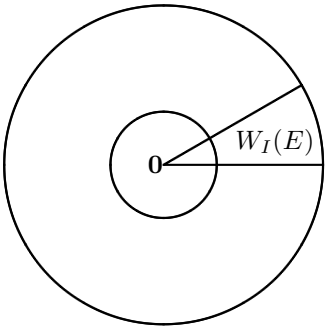
$$\int_{\mathbf{R}_i(U_i)} \left|\frac{\partial x_k}{\partial u_j}\right|^p du$$

are bounded. Here x_k is the k^{th} component of $\mathbf{R}_i^{-1}$. This is because this condition will suffice to use the Vitali convergence theorem. This would have required more work to show however so I have not included it. This is also a reason for featuring the Vitali convergence theorem rather than the dominated convergence theorem which could have been used in many of the steps in the above presentation. One difficulty is that you need to be sure that $\mathbf{R}_i \circ \mathbf{S}_j^{-1}(\mathbf{S}_j(L \cap U_i \cap V_j))$ has measure zero when $\mathbf{S}_j(L \cap U_i \cap V_j)$ has measure zero. In the above, I just assumed that the various functions were Lipschitz continuous and this made the issue an easy one. However, it seems clear that this can be generalized. I am not sure how this all works out because I have not been through it and do not have time to look at it.

All of the above can be done more elegantly and in greater generality if you have Rademacher's theorem which gives the almost everywhere differentiability of Lipschitz functions. In fact, some of the details become a little easier. However, this approach requires more real analysis than I want to include in this book, but the main ideas are all the same. You convolve with a mollifier and then do the hard computations with the mollified function exploiting equality of mixed partial derivatives and then pass to the limit.

11.9 Spherical Coordinates

Consider the following picture.



Definition 11.9.1 *The symbol $W_I(E)$ represents the piece of a wedge between the two concentric spheres such that the points $\mathbf{x} \in W_I(E)$ have the property that $\mathbf{x}/|\mathbf{x}| \in E$, a subset of the unit sphere in $\mathbb{R}^n$, S^{n-1} and $|\mathbf{x}| \in I$, an interval on the real line which does not contain 0.*

Now here are some technical results which are interesting for their own sake. The first gives the existence of a countable basis for $\mathbb{R}^n$. This is a countable set of open sets which has the property that every open set is the union of these special open sets.

Lemma 11.9.2 *Let $\mathcal{B}$ denote the countable set of all balls in $\mathbb{R}^n$ which have centers $\mathbf{x} \in \mathbb{Q}^n$ and rational radii. Then every open set is the union of sets of $\mathcal{B}$.*

Proof: Let U be an open set and let $\mathbf{y} \in U$. Then $B(\mathbf{y}, R) \subseteq U$ for some $R > 0$. Now by density of $\mathbb{Q}^n$ in $\mathbb{R}^n$, there exists $\mathbf{x} \in B(\mathbf{y}, R/10) \cap \mathbb{Q}^n$. Now let $r \in \mathbb{Q}$ and satisfy $R/10 < r < R/3$. Then $\mathbf{y} \in B(\mathbf{x}, r) \subseteq B(\mathbf{y}, R) \subseteq U$. This proves the lemma.

With the above countable basis, the following theorem is very easy to obtain. It is called the Lindelöf property.

Theorem 11.9.3 *Let $\mathcal{C}$ be any collection of open sets and let $U = \cup\mathcal{C}$. Then there exist countably many sets of $\mathcal{C}$ whose union is also equal to U.*

Proof: Let $\mathcal{B}'$ denote those sets of $\mathcal{B}$ in Lemma 11.9.2 which are contained in some set of $\mathcal{C}$. By this lemma, it follows $\cup\mathcal{B}' = U$. Now use axiom of choice to select for each $\mathcal{B}'$ a single set of $\mathcal{C}$ containing it. Denote the resulting countable collection $\mathcal{C}'$. Then

$$U = \cup\mathcal{B}' \subseteq \cup\mathcal{C}' \subseteq U$$

This proves the theorem.

Now consider all the open subsets of $\mathbb{R}^n \setminus \{\mathbf{0}\}$. If U is any such open set, it is clear that if $\mathbf{y} \in U$, then there exists a set open in S^{n-1}, E and an open interval I such that $\mathbf{y} \in W_I(E) \subseteq U$. It follows from Theorem 11.9.3 that every open set which does not contain $\mathbf{0}$ is the countable union of the sets of the form $W_I(E)$ for E open in S^{n-1}.

The divergence theorem and Green's theorem hold for sets $W_I(E)$ whenever E is the intersection of S^{n-1} with a finite intersection of balls. This is because the resulting set has PC^1 boundary. Therefore, from the divergence theorem and letting $I = (0,1)$

$$\int_{W_I(E)} \operatorname{div}(\mathbf{x})\, dx = \int_E \mathbf{x} \cdot \frac{\mathbf{x}}{|\mathbf{x}|} d\sigma + \int_{\text{straight part}} \mathbf{x} \cdot \mathbf{n} d\sigma$$

where I am going to denote by σ the measure on S^{n-1} which corresponds to the divergence theorem and other theorems given above. On the straight parts of the boundary of $W_I(E)$, the vector field $\mathbf{x}$ is parallel to the surface while $\mathbf{n}$ is perpendicular to it, all

this off a set of measure zero of course. Therefore, the integrand vanishes and the above reduces to

$$nm_n\left(W_I\left(E\right)\right) = \sigma\left(E\right)$$

Now let $\mathcal{G}$ denote those Borel sets of S^{n-1} such that the above holds for $I = (0,1)$, both sides making sense because both E and $W_I(E)$ are Borel sets in S^{n-1} and $\mathbb{R}^n \setminus \{\mathbf{0}\}$ respectively. Then $\mathcal{G}$ contains the π system of sets which are the finite intersection of balls with S^{n-1}. Also if $\{E_i\}$ are disjoint sets in $\mathcal{G}$, then $W_I\left(\cup_{i=1}^{\infty} E_i\right) = \cup_{i=1}^{\infty} W_I\left(E_i\right)$ and so

$$\begin{aligned} nm_n\left(W_I\left(\cup_{i=1}^{\infty} E_i\right)\right) &= nm_n\left(\cup_{i=1}^{\infty} W_I\left(E_i\right)\right) \\ &= n\sum_{i=1}^{\infty} m_n\left(W_I\left(E_i\right)\right) \\ &= \sum_{i=1}^{\infty} \sigma\left(E_i\right) = \sigma\left(\cup_{i=1}^{\infty} E_i\right) \end{aligned}$$

and so $\mathcal{G}$ is closed with respect to countable disjoint unions. Next let $E \in \mathcal{G}$. Then

$$\begin{aligned} nm_n\left(W_I\left(E^C\right)\right) + nm_n\left(W_I\left(E\right)\right) &= nm_n\left(W_I\left(S^{n-1}\right)\right) \\ &= \sigma\left(S^{n-1}\right) = \sigma\left(E\right) + \sigma\left(E^C\right) \end{aligned}$$

Now subtracting the equal quantities $nm_n\left(W_I\left(E\right)\right)$ and $\sigma\left(E\right)$ from both sides yields $E^C \in \mathcal{G}$ also. Therefore, by the Lemma on π systems Lemma 9.1.2, it follows $\mathcal{G}$ contains the σ algebra generated by these special sets E the intersection of finitely many open balls with S^{n-1}. Therefore, since any open set is the countable union of balls, it follows the sets open in S^{n-1} are contained in this σ algebra. Hence $\mathcal{G}$ equals the Borel sets. This has proved the following important theorem.

Theorem 11.9.4 *Let σ be the Borel measure on S^{n-1} which goes with the divergence theorems and other theorems like Green's and Stoke's theorem. Then for all E Borel,*

$$\sigma\left(E\right) = nm_n\left(W_I\left(E\right)\right)$$

where $I = (0,1)$. Furthermore, $W_I(E)$ is Borel for any interval. Also

$$m_n\left(W_{[a,b]}\left(E\right)\right) = m_n\left(W_{(a,b)}\left(E\right)\right) = \left(b^n - a^n\right) m_n\left(W_{(0,1)}\left(E\right)\right)$$

Proof: To show $W_I(E)$ is Borel for any I first suppose I is open of the form $(0,r)$. Then

$$W_I\left(E\right) = rW_{(0,1)}\left(E\right)$$

and this mapping $\mathbf{x} \to r\mathbf{x}$ is continuous with continuous inverse so it maps Borel sets to Borel sets. If $I = (0,r]$,

$$W_I\left(E\right) = \cap_{n=1}^{\infty} W_{\left(0,\frac{1}{n}+r\right)}\left(E\right)$$

and so it is Borel.

$$W_{[a,b]}\left(E\right) = W_{(0,b]}\left(E\right) \setminus W_{(0,a)}\left(E\right)$$

so this is also Borel. Similarly $W_{(a,b]}(E)$ is Borel. The last assertion is obvious and follows from the change of variables formula. This proves the theorem.

Now with this preparation, it is possible to discuss polar coordinates (spherical coordinates) a different way than before.

Note that if $\rho = |\mathbf{x}|$ and $\boldsymbol{\omega} \equiv \mathbf{x}/|\mathbf{x}|$, then $\mathbf{x} = \rho\boldsymbol{\omega}$. Also the map which takes $(0,\infty) \times S^{n-1}$ to $\mathbb{R}^n \setminus \{\mathbf{0}\}$ given by $(\rho, \boldsymbol{\omega}) \to \rho\boldsymbol{\omega} = \mathbf{x}$ is one to one and onto and

continuous. In addition to this, it follows right away from the definition that if I is any interval and $E \subseteq S^{n-1}$,

$$\mathcal{X}_{W_I(E)}(\rho\boldsymbol{\omega}) = \mathcal{X}_E(\boldsymbol{\omega})\,\mathcal{X}_I(\rho)$$

Lemma 11.9.5 *For any Borel set $F \subseteq \mathbb{R}^n$,*

$$m_n(F) = \int_0^\infty \int_{S^{n-1}} \rho^{n-1}\mathcal{X}_F(\rho\boldsymbol{\omega})\,d\sigma d\rho$$

and the iterated integral on the right makes sense.

Proof: First suppose $F = W_I(E)$ where E is Borel in S^{n-1} and I is an interval having endpoints $a \leq b$. Then

$$\begin{aligned}\int_0^\infty \int_{S^{n-1}} \rho^{n-1}\mathcal{X}_{W_I(E)}(\rho\boldsymbol{\omega})\,d\sigma d\rho &= \int_0^\infty \int_{S^{n-1}} \rho^{n-1}\mathcal{X}_E(\boldsymbol{\omega})\,\mathcal{X}_I(\rho)\,d\sigma d\rho \\ &= \int_a^b \rho^{n-1}\sigma(E)\,d\rho = \int_a^b \rho^{n-1} n m_n\left(W_{(0,1)}(E)\right)d\rho \\ &= (b^n - a^n)\,m_n\left(W_{(0,1)}(E)\right)\end{aligned}$$

and by Theorem 11.9.4, this equals $m_n(W_I(E))$. If I is an interval which contains 0, the above conclusion still holds because both sides are unchanged if $\mathbf{0}$ is included on the left and $\rho = 0$ is included on the right. In particular, the conclusion holds for $B(\mathbf{0}, r)$ in place of F.

Now let $\mathcal{G}$ be those Borel sets F such that the desired conclusion holds for $F \cap B(0, M)$.This contains the π system of sets of the form $W_I(E)$ and is closed with respect to countable unions of disjoint sets and complements. Therefore, it equals the Borel sets. Thus

$$m_n(F \cap B(\mathbf{0}, M)) = \int_0^\infty \int_{S^{n-1}} \rho^{n-1}\mathcal{X}_{F\cap B(\mathbf{0},M)}(\rho\boldsymbol{\omega})\,d\sigma d\rho$$

Now let $M \to \infty$ and use the monotone convergence theorem. This proves the lemma.

The lemma implies right away that for s a simple function

$$\int_{\mathbb{R}^n} s\,dm_n = \int_0^\infty \int_{S^{n-1}} \rho^{n-1}s(\rho\boldsymbol{\omega})\,d\sigma d\rho$$

Now the following polar coordinates theorem follows.

Theorem 11.9.6 *Let $f \geq 0$ be Borel measurable. Then*

$$\int_{\mathbb{R}^n} f\,dm_n = \int_0^\infty \int_{S^{n-1}} \rho^{n-1}f(\rho\boldsymbol{\omega})\,d\sigma d\rho$$

and the iterated integral on the right makes sense.

Proof: By Theorem 7.7.12 there exists a sequence of nonnegative simple functions $\{s_k\}$ which increases to f. Therefore, from the monotone convergence theorem and what was shown above,

$$\begin{aligned}\int_{\mathbb{R}^n} f\,dm_n &= \lim_{k\to\infty}\int_{\mathbb{R}^n} s_k\,dm_n \\ &= \lim_{k\to\infty}\int_0^\infty \int_{S^{n-1}} \rho^{n-1}s_k(\rho\boldsymbol{\omega})\,d\sigma d\rho \\ &= \int_0^\infty \int_{S^{n-1}} \rho^{n-1}f(\rho\boldsymbol{\omega})\,d\sigma d\rho\end{aligned}$$

This proves the theorem.

11.10 Exercises

1. Let
$$\omega(\mathbf{x}) \equiv \sum_I a_I(\mathbf{x})\, d\mathbf{x}^I$$
be a differential form where $\mathbf{x} \in \mathbb{R}^m$ and the I are increasing lists of n indices taken from $1, \cdots, m$. Also assume each $a_I(\mathbf{x})$ has the property that all mixed partial derivatives are equal. For example, from Corollary 6.10.2 this happens if the function is C^2. Show that under this condition, $d(d(\omega)) = 0$. To show this, first explain why
$$dx_i \wedge dx_j \wedge d\mathbf{x}^I = -dx_j \wedge dx_i \wedge d\mathbf{x}^I$$
When you integrate one you get -1 times the integral of the other. This is the sense in which the above formula holds. When you have a differential form ω with the property that $d\omega = 0$ this is called a closed form. If $\omega = d\alpha$, then ω is called exact. Thus every closed form is exact provided you have sufficient smoothness on the coefficients of the differential form.

2. Recall that in the definition of area measure, you use
$$J(\mathbf{u}) = \det\left(D\mathbf{R}^{-1}(\mathbf{u})^* D\mathbf{R}^{-1}(\mathbf{u})\right)^{1/2}$$
Now in the special case of the manifold of Green's theorem where
$$\mathbf{R}^{-1}(u_2, \cdots, u_n) = R^*(g(u_2, \cdots, u_n), u_2, \cdots, u_n),$$
show
$$J(\mathbf{u}) = \sqrt{1 + \left(\frac{\partial g}{\partial u_2}\right)^2 + \cdots + \left(\frac{\partial g}{\partial u_n}\right)^2}$$

3. Let $\mathbf{u}_1, \cdots, \mathbf{u}_p$ be vectors in $\mathbb{R}^n$. Show $\det M \geq 0$ where $M_{ij} \equiv \mathbf{u}_i \cdot \mathbf{u}_j$. **Hint:** Show this matrix has all nonnegative eigenvalues and then use the theorem which says the determinant is the product of the eigenvalues. This matrix is called the Grammian matrix. The details follow from noting that M is of the form
$$U^*U \equiv \begin{pmatrix} \mathbf{u}_1^* \\ \vdots \\ \mathbf{u}_p^* \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_p \end{pmatrix}$$
and then showing that U^*U has all nonnegative eigenvalues.

4. Suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ are n vectors in $\mathbb{R}^m$ for $m \geq n$. Show that the only appropriate definition of the volume of the n dimensional parallelepiped determined by these vectors,
$$\left\{\sum_{j=1}^n s_j \mathbf{v}_j : s_j \in [0,1]\right\}$$
is
$$\det(M^*M)^{1/2}$$
where M is the $m \times n$ matrix which has columns $\mathbf{v}_1, \cdots, \mathbf{v}_n$. **Hint:** Show this is clearly true if $n = 1$ because the above just yields the usual length of the vector. Now suppose the formula gives the right thing for $n-1$ vectors and argue it gives

the right thing for n vectors. In doing this, you might want to show that a vector which is perpendicular to the span of $\mathbf{v}_1, \cdots, \mathbf{v}_{n-1}$ is

$$\det \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \\ v_{11} & v_{12} & \cdots & v_{1n} \\ \vdots & \vdots & & \vdots \\ v_{n-1,1} & v_{n-1,2} & \cdots & v_{n-1,n} \end{pmatrix}$$

where $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ is an orthonormal basis for $\text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_n)$ and v_{ij} is the j^{th} component of $\mathbf{v}_i$ with respect to this orthonormal basis. Then argue that if you replace the top line with $v_{n1}, \cdots, v_{nn}$, the absolute value of the resulting determinant is the appropriate definition of the volume of the parallelepiped. Next note you could get this number by taking the determinant of the transpose of the above matrix times that matrix and then take a square root. After this, identify this product with a Grammian matrix and then the desired result follows.

5. Why is the definition of area on a manifold given above reasonable and what is its geometric meaning? Each function

$$u_i \to \mathbf{R}^{-1}(u_1, \cdots, u_n)$$

yields a curve which lies in Ω. Thus $\mathbf{R}^{-1}_{,u_i}$ is a vector tangent to this curve and $\mathbf{R}^{-1}_{,u_i} du_i$ is an "infinitesimal" vector tangent to the curve. Now use the previous problem to see that when you find the area of a set on Ω, you are essentially summing the volumes of infinitesimal parallelepipeds which are "tangent" to Ω.

6. Let Ω be a bounded open set in $\mathbb{R}^n$ with PC^1 boundary or more generally one for which the divergence theorem holds. Let $u, v \in C^2(\overline{\Omega})$. Then

$$\int_\Omega (v\Delta u - u\Delta v)\, dx = \int_{\partial\Omega} \left(v\frac{\partial u}{\partial n} - u\frac{\partial v}{\partial n} \right) d\sigma_{n-1}$$

Here

$$\frac{\partial u}{\partial n} \equiv \nabla u \cdot \mathbf{n}$$

where $\mathbf{n}$ is the unit outer normal described above. Establish this formula which is known as Green's identity. **Hint:** You might establish the following easy identity.

$$\nabla \cdot (v\nabla u) - v\Delta u = \nabla v \cdot \nabla u.$$

Recall $\Delta u \equiv \sum_{k=1}^n u_{x_k x_k}$ and $\nabla u = (u_{x_1}, \cdots, u_{x_n})$ while

$$\nabla \cdot \mathbf{F} \equiv f_{1x_1} + \cdots + f_{nx_n} \equiv \text{div}(\mathbf{F})$$

Chapter 12

The Laplace And Poisson Equations

This material is mostly in the book by Evans [14] which is where I got it. It is really partial differential equations but it is such a nice illustration of the divergence theorem and other advanced calculus theorems, that I am including it here even if it is somewhat out of place and would normally be encountered in a partial differential equations course.

12.1 Balls

Recall, $B(\mathbf{x},r)$ denotes the set of all $\mathbf{y}\in\mathbb{R}^n$ such that $|\mathbf{y}-\mathbf{x}|<r$. By the change of variables formula for multiple integrals or simple geometric reasoning, all balls of radius r have the same volume. Furthermore, simple reasoning or change of variables formula will show that the volume of the ball of radius r equals $\alpha_n r^n$ where α_n will denote the volume of the unit ball in $\mathbb{R}^n$. With the divergence theorem, it is now easy to give a simple relationship between the surface area of the ball of radius r and the volume. By the divergence theorem,

$$\int_{B(\mathbf{0},r)} \operatorname{div}\mathbf{x}\,dx = \int_{\partial B(\mathbf{0},r)} \mathbf{x}\cdot\frac{\mathbf{x}}{|\mathbf{x}|}d\sigma_{n-1}$$

because the unit outward normal on $\partial B(\mathbf{0},r)$ is $\frac{\mathbf{x}}{|\mathbf{x}|}$. Therefore,

$$n\alpha_n r^n = r\sigma_{n-1}\left(\partial B(\mathbf{0},r)\right)$$

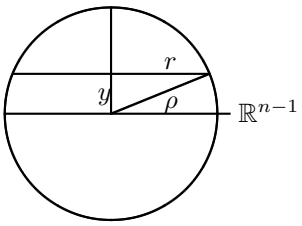
and so

$$\sigma_{n-1}\left(\partial B(\mathbf{0},r)\right) = n\alpha_n r^{n-1}.$$

You recall the surface area of $S^2\equiv\left\{\mathbf{x}\in\mathbb{R}^3 : |\mathbf{x}|=r\right\}$ is given by $4\pi r^2$ while the volume of the ball, $B(\mathbf{0},r)$ is $\frac{4}{3}\pi r^3$. This follows the above pattern. You just take the derivative with respect to the radius of the volume of the ball of radius r to get the area of the surface of this ball. Let ω_n denote the area of the sphere $S^{n-1}=\{\mathbf{x}\in\mathbb{R}^n : |\mathbf{x}|=1\}$. I just showed that

$$\omega_n = n\alpha_n. \tag{12.1}$$

I want to find α_n now and also to get a relationship between ω_n and ω_{n-1}. Consider the following picture of the ball of radius ρ seen on the side.



Taking slices at height y as shown and using that these slices have $n-1$ dimensional area equal to $\alpha_{n-1}r^{n-1}$, it follows from Fubini's theorem

$$\alpha_n \rho^n = 2\int_0^\rho \alpha_{n-1}\left(\rho^2 - y^2\right)^{(n-1)/2} dy \tag{12.2}$$

Lemma 12.1.1 $\Gamma(1/2) = \sqrt{\pi}$

Proof:

$$\Gamma\left(\frac{1}{2}\right) \equiv \int_0^\infty e^{-t}t^{-1/2}dt$$

Now change the variables letting $t = s^2$ so $dt = 2sds$ and the integral becomes

$$2\int_0^\infty e^{-s^2}ds = \int_{-\infty}^\infty e^{-s^2}ds$$

Thus $\Gamma\left(\frac{1}{2}\right) = \int_{-\infty}^\infty e^{-x^2}dx$ so $\Gamma\left(\frac{1}{2}\right)^2 = \int_{-\infty}^\infty \int_{-\infty}^\infty e^{-\left(x^2+y^2\right)}dxdy$ and by polar coordinates and changing the variables, this is just

$$\int_0^{2\pi}\int_0^\infty e^{-r^2}rdrd\theta = \pi$$

Therefore, $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ as claimed. This proves the lemma. ■

Theorem 12.1.2 $\alpha_n = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}$ *where* Γ *denotes the gamma function, defined for* $\alpha > 0$ *by*

$$\Gamma(\alpha) \equiv \int_0^\infty e^{-t}t^{\alpha-1}dt.$$

Proof: Recall that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$. (Establish this by integrating by parts.) This is proved by induction using 12.2. When $n = 1$, the right answer should be 2 because in this case the ball is just $(-1,1)$. Is this what is obtained from the formula? Is

$$\alpha_1 \equiv 2 = \frac{\pi^{1/2}}{\Gamma(3/2)}?$$

Using the identity $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, the above equals

$$\frac{\pi^{1/2}}{(1/2)\Gamma(1/2)} = 2$$

from the above lemma. Now suppose the theorem is true for n. Then letting $\rho = 1$, 12.2 implies

$$\alpha_{n+1} = 2\frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}\int_0^1 \left(1-y^2\right)^{n/2}dy$$

Now change variables. Let $u = y^2$. Then

$$\begin{aligned}\alpha_{n+1} &= \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}\int_0^1 u^{-1/2}\left(1-u\right)^{\frac{1}{2}n}\,du \\ &= \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}\int_0^1 u^{(1/2)-1}\left(1-u\right)^{\frac{n+2}{2}-1}\,du \\ &= \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}B\left(\frac{1}{2},\frac{n+2}{2}\right)\end{aligned}$$

At this point, use the result of Problem 8 on Page 222 to simplify the messy integral which equals the beta function. Thus the above equals

$$\begin{aligned}&\frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}\frac{\Gamma\left(1/2\right)\Gamma\left(\frac{n+2}{2}\right)}{\Gamma\left(\frac{n+2}{2}+\frac{1}{2}\right)} \\ = \;& \pi^{n/2}\frac{\pi^{1/2}}{\Gamma\left(\frac{1}{2}n+\frac{3}{2}\right)} = \frac{\pi^{(n+1)/2}}{\Gamma\left(\frac{n+1}{2}+1\right)}\end{aligned}$$

and this gives the correct formula for α_{n+1}. This proves the theorem. ■

12.2 Poisson's Problem

The Poisson problem is to find u satisfying the two conditions

$$\Delta u = f, \text{ in } U, \; u = g \text{ on } \partial U. \tag{12.3}$$

Here U is an open bounded set for which the divergence theorem holds. For example, it could be a PC^1 manifold. When $f = 0$ this is called Laplace's equation and the boundary condition given is called a Dirichlet boundary condition. When $\Delta u = 0$, the function, u is said to be a harmonic function. When $f \neq 0$, it is called Poisson's equation. I will give a way of representing the solution to these problems. When this has been done, great and marvelous conclusions may be drawn about the solutions. Before doing anything else however, it is wise to prove a fundamental result called the weak maximum principle.

Theorem 12.2.1 *Suppose U is an open bounded set and*

$$u \in C^2\left(U\right) \cap C\left(\overline{U}\right)$$

and

$$\Delta u \geq 0 \text{ in } U.$$

Then

$$\max\left\{u\left(\mathbf{x}\right) : \mathbf{x} \in \overline{U}\right\} = \max\left\{u\left(\mathbf{x}\right) : \mathbf{x} \in \partial U\right\}.$$

Proof: Suppose not. Then there exists $\mathbf{x}_0 \in U$ such that

$$u\left(\mathbf{x}_0\right) > \max\left\{u\left(\mathbf{x}\right) : \mathbf{x} \in \partial U\right\}.$$

Consider $w_\varepsilon\left(\mathbf{x}\right) \equiv u\left(\mathbf{x}\right) + \varepsilon\left|\mathbf{x}\right|^2$. I claim that for small enough $\varepsilon > 0$, the function w_ε also has this property. If not, there exists $\mathbf{x}_\varepsilon \in \partial U$ such that $w_\varepsilon\left(\mathbf{x}_\varepsilon\right) \geq w_\varepsilon\left(\mathbf{x}\right)$ for all $\mathbf{x} \in U$. But since U is bounded, it follows the points, $\mathbf{x}_\varepsilon$ are in a compact set and so there exists a subsequence, still denoted by $\mathbf{x}_\varepsilon$ such that as $\varepsilon \to 0, \mathbf{x}_\varepsilon \to \mathbf{x}_1 \in \partial U$. But then for any $\mathbf{x} \in U$,

$$u\left(\mathbf{x}_0\right) \leq w_\varepsilon\left(\mathbf{x}_0\right) \leq w_\varepsilon\left(\mathbf{x}_\varepsilon\right)$$

and taking a limit as $\varepsilon \to 0$ yields

$$u(\mathbf{x}_0) \leq u(\mathbf{x}_1)$$

contrary to the property of $\mathbf{x}_0$ above. It follows that my claim is verified. Pick such an ε. Then w_ε assumes its maximum value in U say at $\mathbf{x}_2$. Then by the second derivative test,

$$\Delta w_\varepsilon(\mathbf{x}_2) = \Delta u(\mathbf{x}_2) + 2\varepsilon \leq 0$$

which requires $\Delta u(\mathbf{x}_2) \leq -2\varepsilon$, contrary to the assumption that $\Delta u \geq 0$. This proves the theorem. ■

The theorem makes it very easy to verify the following uniqueness result.

Corollary 12.2.2 *Suppose U is an open bounded set and*

$$u \in C^2(U) \cap C\left(\overline{U}\right)$$

and

$$\Delta u = 0 \text{ in } U, u = 0 \text{ on } \partial U.$$

Then $u = 0$.

Proof: From the weak maximum principle, $u \leq 0$. Now apply the weak maximum principle to $-u$ which satisfies the same conditions as u. Thus $-u \leq 0$ and so $u \geq 0$. Therefore, $u = 0$ as claimed. This proves the corollary. ■

Define

$$r_n(\mathbf{x}) \equiv \begin{cases} \ln|\mathbf{x}| & \text{if } n = 2 \\ \frac{1}{|\mathbf{x}|^{n-2}} & \text{if } n > 2 \end{cases}.$$

Then it is fairly routine to verify the following Lemma.

Lemma 12.2.3 *For r_n given above,*

$$\Delta r_n = 0.$$

Proof: I will verify the case where $n \geq 3$ and leave the other case for you.

$$D_{x_i}\left(\sum_{i=1}^n x_i^2\right)^{-(n-2)/2} = -(n-2)\, x_i \left(\sum_j x_j^2\right)^{-n/2}$$

Therefore,

$$D_{x_i}(D_{x_i}(r_n)) = \left(\sum_j x_j^2\right)^{-(n+2)/2} (n-2)\left[n x_i^2 - \sum_{j=1}^n x_j^2\right].$$

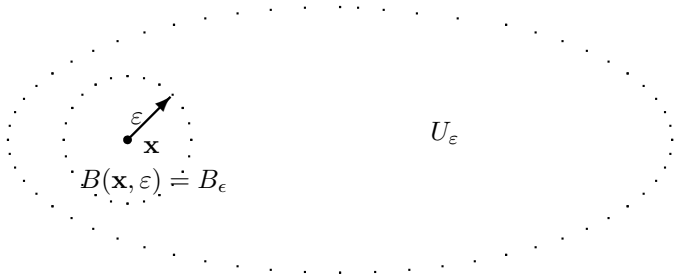
It follows

$$\Delta r_n = \left(\sum_j x_j^2\right)^{\frac{-(n+2)}{2}} (n-2)\left(n\sum_{i=1}^n x_i^2 - \sum_{i=1}^n \sum_{j=1}^n x_j^2\right) = 0.$$

This proves the lemma. ■

From now on assume ∂U is PC^1 to be specific.

Now let U_ε be as indicated in the following picture. I have taken out a ball of radius ε which is centered at the point, $\mathbf{x} \in U$.



Then the divergence theorem will continue to hold for U_ε (why?) and so I can use Green's identity, Problem 6 on Page 333 to write the following for $u, v \in C^2\left(\overline{U}\right)$.

$$\int_{U_\varepsilon}\left(u\Delta v - v\Delta u\right)dx = \int_{\partial U}\left(u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n}\right)d\sigma - \int_{\partial B_\varepsilon}\left(u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n}\right)d\sigma \tag{12.4}$$

Now, letting $\mathbf{x} \in U$, I will pick for v the function,

$$v\left(\mathbf{y}\right) \equiv r_n\left(\mathbf{y}-\mathbf{x}\right) - \psi^{\mathbf{x}}\left(\mathbf{y}\right) \tag{12.5}$$

where $\psi^{\mathbf{x}}$ is a function which is chosen such that on ∂U,

$$\psi^{\mathbf{x}}\left(\mathbf{y}\right) = r_n\left(\mathbf{y}-\mathbf{x}\right)$$

so that 12.5 vanishes for $\mathbf{y} \in \partial U$ and $\psi^{\mathbf{x}}$ is in $C^2\left(\overline{U}\right)$ and also satisfies

$$\Delta\psi^{\mathbf{x}} = 0.$$

The existence of such a function is another issue. For now, assume such a function exists.[1] Then assuming such a function exists, 12.4 reduces to

$$-\int_{U_\varepsilon} v\Delta u dx = \int_{\partial U} u\frac{\partial v}{\partial n}d\sigma - \int_{\partial B_\varepsilon}\left(u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n}\right)d\sigma. \tag{12.6}$$

The idea now is to let $\varepsilon \to 0$ and see what happens. Consider the term

$$\int_{\partial B_\varepsilon} v\frac{\partial u}{\partial n}d\sigma.$$

The area is $O\left(\varepsilon^{n-1}\right)$ while the integrand is $O\left(\varepsilon^{-(n-2)}\right)$ in the case where $n \geq 3$. In the case where $n = 2$, the area is $O\left(\varepsilon\right)$ and the integrand is $O\left(\left|\ln\left|\varepsilon\right|\right|\right)$. Now you know that $\lim_{\varepsilon\to 0}\varepsilon\ln\left|\varepsilon\right| = 0$ and so in the case $n = 2$, this term converges to 0 as $\varepsilon \to 0$. In the case that $n \geq 3$, it also converges to zero because in this case the integral is $O\left(\varepsilon\right)$.

Next consider the term

$$-\int_{\partial B_\varepsilon} u\frac{\partial v}{\partial n}d\sigma = -\int_{\partial B_\varepsilon} u\left(\mathbf{y}\right)\left(\frac{\partial r_n}{\partial n}\left(\mathbf{y}-\mathbf{x}\right) - \frac{\partial\psi^{\mathbf{x}}}{\partial n}\left(\mathbf{y}\right)\right)d\sigma.$$

[1] In fact, if the boundary of U is smooth enough, such a function will always exist, although this requires more work to show but this is not the point. The point is to explicitly find it and this will only be possible for certain simple choices of U.

This term does not disappear as $\varepsilon \to 0$. First note that since $\psi^{\mathbf{x}}$ has bounded derivatives,

$$\lim_{\varepsilon\to 0} -\int_{\partial B_\varepsilon} u(\mathbf{y})\left(\frac{\partial r_n}{\partial n}(\mathbf{y}-\mathbf{x}) - \frac{\partial \psi^{\mathbf{x}}}{\partial n}(\mathbf{y})\right) d\sigma = \lim_{\varepsilon\to 0}\left(-\int_{\partial B_\varepsilon} u(\mathbf{y})\frac{\partial r_n}{\partial n}(\mathbf{y}-\mathbf{x})\, d\sigma\right) \tag{12.7}$$

and so it is just this last item which is of concern.

First consider the case that $n = 2$. In this case,

$$\nabla r_2(\mathbf{y}) = \left(\frac{y_1}{|\mathbf{y}|^2}, \frac{y_2}{|\mathbf{y}|^2}\right)$$

Also, on ∂B_ε, the exterior unit normal, $\mathbf{n}$, equals

$$\frac{1}{\varepsilon}(y_1 - x_1, y_2 - x_2).$$

It follows that on ∂B_ε,

$$\frac{\partial r_2}{\partial n}(\mathbf{y}-\mathbf{x}) = \frac{1}{\varepsilon}(y_1 - x_1, y_2 - x_2)\cdot\left(\frac{y_1 - x_1}{|\mathbf{y}-\mathbf{x}|^2}, \frac{y_2 - x_2}{|\mathbf{y}-\mathbf{x}|^2}\right) = \frac{1}{\varepsilon}.$$

Therefore, this term in 12.7 converges to

$$-u(\mathbf{x})\, 2\pi. \tag{12.8}$$

Next consider the case where $n \geq 3$. In this case,

$$\nabla r_n(\mathbf{y}) = -(n-2)\left(\frac{y_1}{|\mathbf{y}|^n}, \cdots, \frac{y_n}{|\mathbf{y}|}\right)$$

and the unit outer normal, $\mathbf{n}$, equals

$$\frac{1}{\varepsilon}(y_1 - x_1, \cdots, y_n - x_n).$$

Therefore,

$$\frac{\partial r_n}{\partial n}(\mathbf{y}-\mathbf{x}) = -\frac{(n-2)}{\varepsilon}\frac{|\mathbf{y}-\mathbf{x}|^2}{|\mathbf{y}-\mathbf{x}|^n} = \frac{-(n-2)}{\varepsilon^{n-1}}.$$

Letting ω_n denote the $n-1$ dimensional surface area of the unit sphere, S^{n-1}, it follows that the last term in 12.7 converges to

$$u(\mathbf{x})(n-2)\omega_n \tag{12.9}$$

Finally consider the integral,

$$\int_{B_\varepsilon} v\Delta u dx.$$

$$\begin{aligned}\int_{B_\varepsilon} |v\Delta u|\, dx &\leq C\int_{B_\varepsilon} |r_n(\mathbf{y}-\mathbf{x}) - \psi^{\mathbf{x}}(\mathbf{y})|\, dy \\ &\leq C\int_{B_\varepsilon} |r_n(\mathbf{y}-\mathbf{x})|\, dy + O(\varepsilon^n)\end{aligned}$$

Using polar coordinates to evaluate this improper integral in the case where $n \geq 3$,

$$\begin{aligned}C\int_{B_\varepsilon} |r_n(\mathbf{y}-\mathbf{x})|\, dx &= C\int_0^\varepsilon \int_{S^{n-1}} \frac{1}{\rho^{n-2}}\rho^{n-1} d\sigma d\rho \\ &= C\int_0^\varepsilon \int_{S^{n-1}} \rho\, d\sigma d\rho\end{aligned}$$

which converges to 0 as $\varepsilon \to 0$. In the case where $n = 2$

$$C\int_{B_\varepsilon} |r_n(\mathbf{y}-\mathbf{x})|\,dx = C\int_0^\varepsilon \int_{S^{n-1}} \ln(\rho)\,\rho d\sigma d\rho$$

which also converges to 0 as $\varepsilon \to 0$. Therefore, returning to 12.6 and using the above limits, yields in the case where $n \geq 3$,

$$-\int_U v\Delta u dx = \int_{\partial U} u\frac{\partial v}{\partial n}d\sigma + u(\mathbf{x})(n-2)\omega_n, \tag{12.10}$$

and in the case where $n = 2$,

$$-\int_U v\Delta u dx = \int_{\partial U} u\frac{\partial v}{\partial n}d\sigma - u(\mathbf{x})2\pi. \tag{12.11}$$

These two formulas show that it is possible to represent the solutions to Poisson's problem provided the function, $\psi^{\mathbf{x}}$ can be determined. I will show you can determine this function in the case that $U = B(\mathbf{0}, r)$.

12.2.1 Poisson's Problem For A Ball

Lemma 12.2.4 *When* $|\mathbf{y}| = r$ *and* $\mathbf{x} \neq \mathbf{0}$,

$$\left|\frac{\mathbf{y}|\mathbf{x}|}{r} - \frac{r\mathbf{x}}{|\mathbf{x}|}\right| = |\mathbf{x}-\mathbf{y}|,$$

and for $|\mathbf{x}|, |\mathbf{y}| < r, \mathbf{x} \neq \mathbf{0}$,

$$\left|\frac{\mathbf{y}|\mathbf{x}|}{r} - \frac{r\mathbf{x}}{|\mathbf{x}|}\right| \neq 0.$$

Proof: Suppose first that $|\mathbf{y}| = r$. Then

$$\begin{aligned}\left|\frac{\mathbf{y}|\mathbf{x}|}{r} - \frac{r\mathbf{x}}{|\mathbf{x}|}\right|^2 &= \left(\frac{\mathbf{y}|\mathbf{x}|}{r} - \frac{r\mathbf{x}}{|\mathbf{x}|}\right)\cdot\left(\frac{\mathbf{y}|\mathbf{x}|}{r} - \frac{r\mathbf{x}}{|\mathbf{x}|}\right)\\ &= \frac{|\mathbf{x}|^2}{r^2}|\mathbf{y}|^2 - 2\mathbf{y}\cdot\mathbf{x} + r^2\frac{|\mathbf{x}|^2}{|\mathbf{x}|^2}\\ &= |\mathbf{x}|^2 - 2\mathbf{x}\cdot\mathbf{y} + |\mathbf{y}|^2 = |\mathbf{x}-\mathbf{y}|^2.\end{aligned}$$

This proves the first claim. Next suppose $|\mathbf{x}|, |\mathbf{y}| < r$ and suppose, contrary to what is claimed, that

$$\frac{\mathbf{y}|\mathbf{x}|}{r} - \frac{r\mathbf{x}}{|\mathbf{x}|} = \mathbf{0}.$$

Then

$$\mathbf{y}|\mathbf{x}|^2 = r^2\mathbf{x}$$

and so $|\mathbf{y}||\mathbf{x}|^2 = r^2|\mathbf{x}|$ which implies

$$|\mathbf{y}||\mathbf{x}| = r^2$$

contrary to the assumption that $|\mathbf{x}|, |\mathbf{y}| < r$. ■

Let

$$\psi^{\mathbf{x}}(\mathbf{y}) \equiv \begin{cases} \left|\frac{\mathbf{y}|\mathbf{x}|}{r} - \frac{r\mathbf{x}}{|\mathbf{x}|}\right|^{-(n-2)}, \ r^{-(n-2)} \text{ for } \mathbf{x} = \mathbf{0} \text{ if } n \geq 3 \\ \ln\left|\frac{\mathbf{y}|\mathbf{x}|}{r} - \frac{r\mathbf{x}}{|\mathbf{x}|}\right|, \ \ln(r) \text{ if } \mathbf{x} = \mathbf{0} \text{ if } n = 2 \end{cases}$$

Note that

$$\lim_{\mathbf{x}\to\mathbf{0}}\left|\frac{\mathbf{y}\,|\mathbf{x}|}{r}-\frac{r\mathbf{x}}{|\mathbf{x}|}\right|=r.$$

Then $\psi^{\mathbf{x}}(\mathbf{y})=r_n(\mathbf{y}-\mathbf{x})$ if $|\mathbf{y}|=r$, and $\Delta\psi^{\mathbf{x}}=0$. This last claim is obviously true if $\mathbf{x}\neq\mathbf{0}$. If $\mathbf{x}=\mathbf{0}$, then $\psi^{\mathbf{0}}(\mathbf{y})$ equals a constant and so it is also obvious in this case that $\Delta\psi^{\mathbf{x}}=0$.

The following lemma is easy to obtain.

Lemma 12.2.5 *Let*

$$f(\mathbf{y})=\begin{cases}|\mathbf{y}-\mathbf{x}|^{-(n-2)} & \text{if } n\geq 3\\ \ln|\mathbf{y}-\mathbf{x}| & \text{if } n=2\end{cases}.$$

Then

$$\nabla f(\mathbf{y})=\begin{cases}\frac{-(n-2)(\mathbf{y}-\mathbf{x})}{|\mathbf{y}-\mathbf{x}|^{n}} & \text{if } n\geq 3\\ \frac{\mathbf{y}-\mathbf{x}}{|\mathbf{y}-\mathbf{x}|^{2}} & \text{if } n=2\end{cases}.$$

Also, the outer normal on $\partial B(\mathbf{0},r)$ is $\mathbf{y}/r$.

From Lemma 12.2.5 it follows easily that for $v(\mathbf{y})=r_n(\mathbf{y}-\mathbf{x})-\psi^{\mathbf{x}}(\mathbf{y})$ and $\mathbf{y}\in\partial B(\mathbf{0},r)$, then for $n\geq 3$,

$$\begin{aligned}\frac{\partial v}{\partial n} &= \frac{\mathbf{y}}{r}\cdot\left[\frac{-(n-2)(\mathbf{y}-\mathbf{x})}{|\mathbf{y}-\mathbf{x}|^{n}}+\left(\frac{|\mathbf{x}|}{r}\right)^{-(n-2)}(n-2)\frac{\left(\mathbf{y}-\frac{r^2}{|\mathbf{x}|^2}\mathbf{x}\right)}{\left|\mathbf{y}-\frac{r^2}{|\mathbf{x}|^2}\mathbf{x}\right|^{n}}\right]\\ &= \frac{-(n-2)}{r}\frac{\left(r^2-\mathbf{y}\cdot\mathbf{x}\right)}{|\mathbf{y}-\mathbf{x}|^{n}}+\frac{\frac{|\mathbf{x}|^2}{r^2}}{\left(\frac{|\mathbf{x}|}{r}\right)^{n}}\frac{(n-2)}{r}\frac{\left(r^2-\frac{r^2}{|\mathbf{x}|^2}\mathbf{x}\cdot\mathbf{y}\right)}{\left|\mathbf{y}-\frac{r^2}{|\mathbf{x}|^2}\mathbf{x}\right|^{n}}\\ &= \frac{-(n-2)}{r}\frac{\left(r^2-\mathbf{y}\cdot\mathbf{x}\right)}{|\mathbf{y}-\mathbf{x}|^{n}}+\frac{(n-2)}{r}\frac{\left(\frac{|\mathbf{x}|^2}{r^2}r^2-\mathbf{x}\cdot\mathbf{y}\right)}{\left|\frac{|\mathbf{x}|}{r}\mathbf{y}-\frac{r}{|\mathbf{x}|}\mathbf{x}\right|^{n}}\end{aligned}$$

which by Lemma 12.2.4 equals

$$\begin{aligned}&\frac{-(n-2)}{r}\frac{\left(r^2-\mathbf{y}\cdot\mathbf{x}\right)}{|\mathbf{y}-\mathbf{x}|^{n}}+\frac{(n-2)}{r}\frac{\left(\frac{|\mathbf{x}|^2}{r^2}r^2-\mathbf{x}\cdot\mathbf{y}\right)}{|\mathbf{y}-\mathbf{x}|^{n}}\\ =\;&\frac{-(n-2)}{r}\frac{r^2}{|\mathbf{y}-\mathbf{x}|^{n}}+\frac{(n-2)}{r}\frac{|\mathbf{x}|^2}{|\mathbf{y}-\mathbf{x}|^{n}}\\ =\;&\frac{(n-2)}{r}\frac{|\mathbf{x}|^2-r^2}{|\mathbf{y}-\mathbf{x}|^{n}}.\end{aligned}$$

In the case where $n=2$, and $|\mathbf{y}|=r$, then Lemma 12.2.4 implies

$$\begin{aligned}\frac{\partial v}{\partial n} &= \frac{\mathbf{y}}{r}\cdot\left[\frac{(\mathbf{y}-\mathbf{x})}{|\mathbf{y}-\mathbf{x}|^{2}}-\left(\frac{|\mathbf{x}|}{r}\right)\frac{\left(\frac{\mathbf{y}|\mathbf{x}|}{r}-\frac{r\mathbf{x}}{|\mathbf{x}|}\right)}{\left|\frac{\mathbf{y}|\mathbf{x}|}{r}-\frac{r\mathbf{x}}{|\mathbf{x}|}\right|^{2}}\right]\\ &= \frac{\mathbf{y}}{r}\cdot\left[\frac{(\mathbf{y}-\mathbf{x})}{|\mathbf{y}-\mathbf{x}|^{2}}-\frac{\left(\frac{\mathbf{y}|\mathbf{x}|^2}{r^2}-\mathbf{x}\right)}{|\mathbf{y}-\mathbf{x}|^{2}}\right]\\ &= \frac{1}{r}\frac{r^2-|\mathbf{x}|^2}{|\mathbf{y}-\mathbf{x}|^{2}}.\end{aligned}$$

Referring to 12.10 and 12.11, we would hope a solution, u to Poisson's problem satisfies for $n \geq 3$

$$\begin{aligned} & -\int_U \left(r_n\left(\mathbf{y}-\mathbf{x}\right)-\psi^{\mathbf{x}}\left(\mathbf{y}\right)\right) f\left(\mathbf{y}\right) dy \\ = & \int_{\partial U} g\left(\mathbf{y}\right)\left(\frac{\left(n-2\right)}{r}\frac{\left|\mathbf{x}\right|^2-r^2}{\left|\mathbf{y}-\mathbf{x}\right|^n}\right) d\sigma\left(y\right)+u\left(\mathbf{x}\right)\left(n-2\right)\omega_n. \end{aligned}$$

Thus

$$u\left(\mathbf{x}\right)=\frac{1}{\omega_n\left(n-2\right)}\cdot$$

$$\left[\int_U \left(\psi^{\mathbf{x}}\left(\mathbf{y}\right)-r_n\left(\mathbf{y}-\mathbf{x}\right)\right) f\left(\mathbf{y}\right) dy+\int_{\partial U} g\left(\mathbf{y}\right)\left(\frac{\left(n-2\right)}{r}\frac{r^2-\left|\mathbf{x}\right|^2}{\left|\mathbf{y}-\mathbf{x}\right|^n}\right) d\sigma\left(y\right)\right]. \quad (12.12)$$

In the case where $n=2$,

$$-\int_U \left(r_2\left(\mathbf{y}-\mathbf{x}\right)-\psi^{\mathbf{x}}\left(\mathbf{y}\right)\right) f\left(\mathbf{y}\right) dx=\int_{\partial U} g\left(\mathbf{y}\right)\left(\frac{1}{r}\frac{r^2-\left|\mathbf{x}\right|^2}{\left|\mathbf{y}-\mathbf{x}\right|^2}\right) d\sigma\left(\mathbf{y}\right)-u\left(\mathbf{x}\right)2\pi$$

and so in this case,

$$u\left(\mathbf{x}\right)=\frac{1}{2\pi}\left[\int_U \left(r_2\left(\mathbf{y}-\mathbf{x}\right)-\psi^{\mathbf{x}}\left(\mathbf{y}\right)\right) f\left(\mathbf{y}\right) dx+\int_{\partial U} g\left(\mathbf{y}\right)\left(\frac{1}{r}\frac{r^2-\left|\mathbf{x}\right|^2}{\left|\mathbf{y}-\mathbf{x}\right|^2}\right) d\sigma\left(\mathbf{y}\right)\right]. \quad (12.13)$$

12.2.2 Does It Work In Case $f=0$?

It turns out these formulas work better than you might expect. In particular, they work in the case where g is only continuous. In deriving these formulas, more was assumed on the function than this. In particular, it would have been the case that g was equal to the restriction of a function in $C^2\left(\mathbb{R}^n\right)$ to $\partial B\left(\mathbf{0},r\right)$. The problem considered here is

$$\Delta u=0 \text{ in } U,\ u=g \text{ on } \partial U$$

From 12.12 it follows that if u solves the above problem, known as the Dirichlet problem, then

$$u\left(\mathbf{x}\right)=\frac{r^2-\left|\mathbf{x}\right|^2}{\omega_n r}\int_{\partial U} g\left(\mathbf{y}\right)\frac{1}{\left|\mathbf{y}-\mathbf{x}\right|^n} d\sigma\left(\mathbf{y}\right).$$

I have shown this in case $u \in C^2\left(\overline{U}\right)$ which is more specific than to say $u \in C^2\left(U\right) \cap C\left(\overline{U}\right)$. Nevertheless, it is enough to give the following lemma.

Lemma 12.2.6 *The following holds for $n \geq 3$.*

$$1=\int_{\partial U} \frac{r^2-\left|\mathbf{x}\right|^2}{r\omega_n\left|\mathbf{y}-\mathbf{x}\right|^n} d\sigma\left(y\right).$$

For $n=2$,

$$1=\int_{\partial U} \frac{1}{2\pi r}\frac{r^2-\left|\mathbf{x}\right|^2}{\left|\mathbf{y}-\mathbf{x}\right|^2} d\sigma\left(\mathbf{y}\right).$$

Proof: Consider the problem

$$\Delta u = 0 \text{ in } U,\ u = 1 \text{ on } \partial U.$$

I know a solution to this problem which is in $C^2\left(\overline{U}\right)$, namely $u \equiv 1$. Therefore, by Corollary 12.2.2 this is the only solution and since it is in $C^2\left(\overline{U}\right)$, it follows from 12.12 that in case $n \geq 3$,

$$1 = u(\mathbf{x}) = \int_{\partial U} \frac{r^2 - |\mathbf{x}|^2}{r\omega_n |\mathbf{y} - \mathbf{x}|^n} d\sigma(y)$$

and in case $n = 2$, the other formula claimed above holds. ■

Theorem 12.2.7 *Let $U = B(\mathbf{0}, r)$ and let $g \in C(\partial U)$. Then there exists a unique solution $u \in C^2(U) \cap C\left(\overline{U}\right)$ to the problem*

$$\Delta u = 0 \text{ in } U,\ u = g \text{ on } \partial U.$$

This solution is given by the formula,

$$u(\mathbf{x}) = \frac{1}{\omega_n r} \int_{\partial U} g(\mathbf{y}) \frac{r^2 - |\mathbf{x}|^2}{|\mathbf{y} - \mathbf{x}|^n} d\sigma(\mathbf{y}) \tag{12.14}$$

for every $n \geq 2$. Here $\omega_2 \equiv 2\pi$.

Proof: That $\Delta u = 0$ in U follows from the observation that the difference quotients used to compute the partial derivatives converge uniformly in $\mathbf{y} \in \partial U$ for any given $\mathbf{x} \in U$. To see this note that for $\mathbf{y} \in \partial U$, the partial derivatives of the expression,

$$\frac{r^2 - |\mathbf{x}|^2}{|\mathbf{y} - \mathbf{x}|^n}$$

taken with respect to x_k are uniformly bounded and continuous. In fact, this is true of all partial derivatives. Therefore you can take the differential operator inside the integral and write

$$\Delta_x \frac{1}{\omega_n r} \int_{\partial U} g(\mathbf{y}) \frac{r^2 - |\mathbf{x}|^2}{|\mathbf{y} - \mathbf{x}|^n} d\sigma(\mathbf{y}) = \frac{1}{\omega_n r} \int_{\partial U} g(\mathbf{y}) \Delta_x \left(\frac{r^2 - |\mathbf{x}|^2}{|\mathbf{y} - \mathbf{x}|^n} \right) d\sigma(\mathbf{y}) = 0.$$

It only remains to verify that it achieves the desired boundary condition. Let $\mathbf{x}_0 \in \partial U$. From Lemma 12.2.6,

$$\begin{aligned} |g(\mathbf{x}_0) - u(\mathbf{x})| &\leq \frac{1}{\omega_n r} \int_{\partial U} |g(\mathbf{y}) - g(\mathbf{x}_0)| \left(\frac{r^2 - |\mathbf{x}|^2}{|\mathbf{y} - \mathbf{x}|^n} \right) d\sigma(\mathbf{y}) && (12.15) \\ &\leq \frac{1}{\omega_n r} \int_{[|\mathbf{y} - \mathbf{x}_0| < \delta]} |g(\mathbf{y}) - g(\mathbf{x}_0)| \left(\frac{r^2 - |\mathbf{x}|^2}{|\mathbf{y} - \mathbf{x}|^n} \right) d\sigma(\mathbf{y}) + && (12.16) \\ &\quad \frac{1}{\omega_n r} \int_{[|\mathbf{y} - \mathbf{x}_0| \geq \delta]} |g(\mathbf{y}) - g(\mathbf{x}_0)| \left(\frac{r^2 - |\mathbf{x}|^2}{|\mathbf{y} - \mathbf{x}|^n} \right) d\sigma(\mathbf{y}) && (12.17) \end{aligned}$$

where δ is a positive number. Letting $\varepsilon > 0$ be given, choose δ small enough that if $|\mathbf{y} - \mathbf{x}_0| < \delta$, then $|g(\mathbf{y}) - g(\mathbf{x}_0)| < \frac{\varepsilon}{2}$. Then for such δ,

$$\begin{aligned} &\frac{1}{\omega_n r} \int_{[|\mathbf{y} - \mathbf{x}_0| < \delta]} |g(\mathbf{y}) - g(\mathbf{x}_0)| \left(\frac{r^2 - |\mathbf{x}|^2}{|\mathbf{y} - \mathbf{x}|^n} \right) d\sigma(\mathbf{y}) \\ \leq\ &\frac{1}{\omega_n r} \int_{[|\mathbf{y} - \mathbf{x}_0| < \delta]} \frac{\varepsilon}{2} \left(\frac{r^2 - |\mathbf{x}|^2}{|\mathbf{y} - \mathbf{x}|^n} \right) d\sigma(\mathbf{y}) \\ \leq\ &\frac{1}{\omega_n r} \int_{\partial U} \frac{\varepsilon}{2} \left(\frac{r^2 - |\mathbf{x}|^2}{|\mathbf{y} - \mathbf{x}|^n} \right) d\sigma(\mathbf{y}) = \frac{\varepsilon}{2}. \end{aligned}$$

Denoting by M the maximum value of g on ∂U, the integral in 12.17 is dominated by

$$\begin{aligned}
&\frac{2M}{\omega_n r}\int_{[|\mathbf{y}-\mathbf{x}_0|\geq\delta]}\left(\frac{r^2-|\mathbf{x}|^2}{|\mathbf{y}-\mathbf{x}|^n}\right)d\sigma(\mathbf{y})\\
\leq\ &\frac{2M}{\omega_n r}\int_{[|\mathbf{y}-\mathbf{x}_0|\geq\delta]}\left(\frac{r^2-|\mathbf{x}|^2}{[|\mathbf{y}-\mathbf{x}_0|-|\mathbf{x}-\mathbf{x}_0|]^n}\right)d\sigma(\mathbf{y})\\
\leq\ &\frac{2M}{\omega_n r}\int_{[|\mathbf{y}-\mathbf{x}_0|\geq\delta]}\left(\frac{r^2-|\mathbf{x}|^2}{\left[\delta-\frac{\delta}{2}\right]^n}\right)d\sigma(\mathbf{y})\\
\leq\ &\frac{2M}{\omega_n r}\left(\frac{2}{\delta}\right)^n\int_{\partial U}\left(r^2-|\mathbf{x}|^2\right)d\sigma(\mathbf{y})
\end{aligned}$$

when $|\mathbf{x}-\mathbf{x}_0|$ is sufficiently small. Then taking $|\mathbf{x}-\mathbf{x}_0|$ still smaller, if necessary, this last expression is less than $\varepsilon/2$ because $|\mathbf{x}_0|=r$ and so $\lim_{\mathbf{x}\to\mathbf{x}_0}\left(r^2-|\mathbf{x}|^2\right)=0$. This proves $\lim_{\mathbf{x}\to\mathbf{x}_0}u(\mathbf{x})=g(\mathbf{x}_0)$ and this proves the existence part of this theorem. The uniqueness part follows from Corollary 12.2.2. ■

Actually, I could have said a little more about the boundary values in Theorem 12.2.7. Since g is continuous on ∂U, it follows g is uniformly continuous and so the above proof shows that actually $\lim_{\mathbf{x}\to\mathbf{x}_0}u(\mathbf{x})=g(\mathbf{x}_0)$ uniformly for $\mathbf{x}_0\in\partial U$.

Not surprisingly, it is not necessary to have the ball centered at $\mathbf{0}$ for the above to work.

Corollary 12.2.8 *Let $U=B(\mathbf{x}_0,r)$ and let $g\in C(\partial U)$. Then there exists a unique solution $u\in C^2(U)\cap C(\overline{U})$ to the problem*

$$\Delta u=0 \text{ in } U,\ u=g \text{ on } \partial U.$$

This solution is given by the formula,

$$u(\mathbf{x})=\frac{1}{\omega_n r}\int_{\partial U}g(\mathbf{y})\frac{r^2-|\mathbf{x}-\mathbf{x}_0|^2}{|\mathbf{y}-\mathbf{x}|^n}d\sigma(\mathbf{y}) \tag{12.18}$$

for every $n\geq 2$. Here $\omega_2=2\pi$.

This corollary implies the following.

Corollary 12.2.9 *Let u be a harmonic function defined on an open set, $U\subseteq\mathbb{R}^n$ and let $\overline{B(\mathbf{x}_0,r)}\subseteq U$. Then*

$$u(\mathbf{x}_0)=\frac{1}{\omega_n r^{n-1}}\int_{\partial B(\mathbf{x}_0,r)}u(\mathbf{y})\,d\sigma$$

The representation formula, 12.14 is called Poisson's integral formula. I have now shown it works better than you had a right to expect for the Laplace equation. What happens when $f\neq 0$?

12.2.3 The Case Where $f\neq 0$, Poisson's Equation

I will verify the results for the case $n\geq 3$. The case $n=2$ is entirely similar. This is still in the context that $U=B(\mathbf{0},r)$. Thus

$$\psi^{\mathbf{x}}(\mathbf{y})\equiv\begin{cases}\left|\frac{\mathbf{y}|\mathbf{x}|}{r}-\frac{r\mathbf{x}}{|\mathbf{x}|}\right|^{-(n-2)}, & r^{-(n-2)} \text{ for } \mathbf{x}=\mathbf{0} \text{ if } n\geq 3\\ \ln\left|\frac{\mathbf{y}|\mathbf{x}|}{r}-\frac{r\mathbf{x}}{|\mathbf{x}|}\right|, & \ln(r) \text{ if } \mathbf{x}=\mathbf{0} \text{ if } n=2\end{cases}$$

Recall that $r_n(\mathbf{y}-\mathbf{x})=\psi^{\mathbf{x}}(\mathbf{y})$ whenever $\mathbf{y}\in\partial U$.

Lemma 12.2.10 *Let $f \in C\left(\overline{U}\right)$ or in $L^p(U)$ for $p > n/2$ [2]. Then for $\mathbf{x} \in U$, and $\mathbf{x}_0 \in \partial U$,*

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \frac{1}{\omega_n (n-2)} \int_U \left(\psi^{\mathbf{x}}(\mathbf{y}) - r_n(\mathbf{y}-\mathbf{x})\right) f(\mathbf{y})\, dy = 0.$$

Proof: There are two parts to this lemma. First the following claim is shown in which an integral is taken over $B(\mathbf{x}_0, \delta)$. After this, the integral over $U \setminus B(\mathbf{x}_0, \delta)$ will be considered. First note that

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \psi^{\mathbf{x}}(\mathbf{y}) - r_n(\mathbf{y}-\mathbf{x}) = 0$$

Claim:

$$\lim_{\delta\to 0} \int_{B(\mathbf{x}_0,\delta)} \psi^{\mathbf{x}}(\mathbf{y}) |f(\mathbf{y})|\, dy = 0, \ \lim_{\delta\to 0} \int_{B(\mathbf{x}_0,\delta)} r_n(\mathbf{y}-\mathbf{x}) |f(\mathbf{y})|\, dy = 0.$$

Proof of the claim: Using polar coordinates,

$$\begin{aligned} & \int_{B(\mathbf{x}_0,\delta)} \psi^{\mathbf{x}}(\mathbf{y}) |f(\mathbf{y})|\, dy \\ = & \int_{B(\mathbf{0},\delta)} r_n\left(\frac{(\mathbf{x}_0+\mathbf{z})|\mathbf{x}|}{r} - \frac{r\mathbf{x}}{|\mathbf{x}|}\right) |f(\mathbf{x}_0+\mathbf{z})|\, dz \\ = & \int_0^\delta \int_{S^{n-1}} r_n\left(\frac{(\mathbf{x}_0+\rho\mathbf{w})|\mathbf{x}|}{r} - \frac{r\mathbf{x}}{|\mathbf{x}|}\right) |f(\mathbf{x}_0+\rho\mathbf{w})|\, \rho^{n-1} d\sigma d\rho \end{aligned}$$

Now from the formula for r_n, there exists $\delta_0 > 0$ such that for $\rho \in [0, \delta_0]$,

$$r_n\left(\frac{(\mathbf{x}_0+\rho\mathbf{w})|\mathbf{x}|}{r} - \frac{r\mathbf{x}}{|\mathbf{x}|}\right) \rho^{n-2}$$

is bounded. Therefore,

$$\int_{B(\mathbf{x}_0,\delta)} \psi^{\mathbf{x}}(\mathbf{y}) |f(\mathbf{y})|\, dy \le C \int_0^\delta \int_{S^{n-1}} |f(\mathbf{x}_0+\rho\mathbf{w})|\, \rho d\sigma d\rho.$$

If f is continuous, this is dominated by an expression of the form

$$C' \int_0^\delta \int_{S^{n-1}} \rho d\sigma d\rho$$

which converges to 0 as $\delta \to 0$.

If $f \in L^p(U)$, then by Holder's inequality, (Problem 3 on Page 256) for $\frac{1}{p} + \frac{1}{q} = 1$,

$$\begin{aligned} & \int_0^\delta \int_{S^{n-1}} |f(\mathbf{x}_0+\rho\mathbf{w})|\, \rho d\sigma d\rho \\ = & \int_0^\delta \int_{S^{n-1}} |f(\mathbf{x}_0+\rho\mathbf{w})|\, \rho^{2-n} \rho^{n-1} d\sigma d\rho \\ \le & \left(\int_0^\delta \int_{S^{n-1}} |f(\mathbf{x}_0+\rho\mathbf{w})|^p\, \rho^{n-1} d\sigma d\rho\right)^{1/p} \cdot \\ & \left(\int_0^\delta \int_{S^{n-1}} \left(\rho^{2-n}\right)^q \rho^{n-1} d\sigma d\rho\right)^{1/q} \\ \le & C\, ||f||_{L^p(U)}. \end{aligned}$$

[2] This means f is measurable and $|f|^p$ has finite integral

Similar reasoning shows that

$$\lim_{\delta\to 0}\int_{B(\mathbf{x}_0,\delta)} |r_n(\mathbf{y}-\mathbf{x})|\,|f(\mathbf{y})|\,dy = 0.$$

This proves the claim.

Let $\varepsilon > 0$ be given and choose $\delta > 0$ such that $r/2 > \delta > 0$ where r is the radius of the ball U and small enough that

$$\int_{B(\mathbf{x}_0,2\delta)} |f(\mathbf{y})|\,|\psi^{\mathbf{x}}(\mathbf{y}) - r_n(\mathbf{y}-\mathbf{x})|\,dy < \varepsilon.$$

Then consider $\mathbf{x} \in B(\mathbf{x}_0,\delta)$ so that for $\mathbf{y} \notin B(\mathbf{x}_0,2\delta), |\mathbf{y}-\mathbf{x}| > \delta$ and so for such $\mathbf{y}$,

$$|\psi^{\mathbf{x}}(\mathbf{y}) - r_n(\mathbf{y}-\mathbf{x})| \le C\delta^{-(n-2)}$$

for some constant C. Thus the integrand in

$$\begin{aligned}
&\left|\int_U f(\mathbf{y})\left(\psi^{\mathbf{x}}(\mathbf{y}) - r_n(\mathbf{y}-\mathbf{x})\right)dy\right| \\
\le\ &\int_{U\setminus B(\mathbf{x}_0,2\delta)} |f(\mathbf{y})|\,|(\psi^{\mathbf{x}}(\mathbf{y}) - r_n(\mathbf{y}-\mathbf{x}))|\,dy \\
&+\int_{B(\mathbf{x}_0,2\delta)} |f(\mathbf{y})|\,|\psi^{\mathbf{x}}(\mathbf{y}) - r_n(\mathbf{y}-\mathbf{x})|\,dy \\
\le &\int_{U\setminus B(\mathbf{x}_0,2\delta)} f(\mathbf{y})\left(\psi^{\mathbf{x}}(\mathbf{y}) - r_n(\mathbf{y}-\mathbf{x})\right)dy + \varepsilon
\end{aligned}$$

Now apply the dominated convergence theorem in this last integral to conclude it converges to 0 as $\mathbf{x}\to\mathbf{x}_0$. This proves the lemma. ■

The following lemma follows from this one and Theorem 12.2.7.

Lemma 12.2.11 *Let $f \in C(\overline{U})$ or in $L^p(U)$ for $p > n/2$ and let $g \in C(\partial U)$. Then if u is given by 12.12 in the case where $n \ge 3$ or by 12.13 in the case where $n = 2$, then if $\mathbf{x}_0 \in \partial U$,*

$$\lim_{\mathbf{x}\to\mathbf{x}_0} u(\mathbf{x}) = g(\mathbf{x}_0).$$

Not surprisingly, you can relax the condition that $g \in C(\partial U)$ but I won't do so here.

The next question is about the partial differential equation satisfied by u for u given by 12.12 in the case where $n \ge 3$ or by 12.13 for $n = 2$. This is going to introduce a new idea. I will just sketch the main ideas and leave you to work out the details, most of which have already been considered in a similar context.

Let $\phi \in C_c^\infty(U)$ and let $\mathbf{x} \in U$. Let U_ε denote the open set which has $\overline{B(\mathbf{y},\varepsilon)}$ deleted from it, much as was done earlier. In what follows I will denote with a subscript of $\mathbf{x}$ things for which $\mathbf{x}$ is the variable. Then denoting by $G(\mathbf{y},\mathbf{x})$ the expression $\psi^{\mathbf{x}}(\mathbf{y}) - r_n(\mathbf{y}-\mathbf{x})$, it is easy to verify that $\Delta_x G(\mathbf{y},\mathbf{x}) = 0$ and so by Fubini's theorem,

$$\int_U \frac{1}{\omega_n(n-2)}\left[\int_U \left(\psi^{\mathbf{x}}(\mathbf{y}) - r_n(\mathbf{y}-\mathbf{x})\right) f(\mathbf{y})\,dy\right]\Delta_x\phi(\mathbf{x})\,dx$$

$$\begin{aligned}
&= \lim_{\varepsilon\to 0}\int_{U_\varepsilon} \frac{1}{\omega_n(n-2)}\left[\int_U \left(\psi^{\mathbf{x}}(\mathbf{y}) - r_n(\mathbf{y}-\mathbf{x})\right) f(\mathbf{y})\,dy\right]\Delta_x\phi(\mathbf{x})\,dx \\
&= \lim_{\varepsilon\to 0}\int_U \left(\int_{U_\varepsilon} \frac{1}{\omega_n(n-2)}\left(\psi^{\mathbf{x}}(\mathbf{y}) - r_n(\mathbf{y}-\mathbf{x})\right)\Delta_x\phi(\mathbf{x})\,dx\right) f(\mathbf{y})\,dy
\end{aligned}$$

$$= \lim_{\varepsilon\to 0}\int_U \left(\int_{U_\varepsilon} \frac{1}{\omega_n (n-2)} \left(\overbrace{\psi^{\mathbf{x}}(\mathbf{y}) - r_n(\mathbf{y}-\mathbf{x})}^{G(\mathbf{y},\mathbf{x})}\right) \Delta_x \phi(\mathbf{x})\, dx\right) f(\mathbf{y})\, dy$$

$$\begin{aligned} &= \lim_{\varepsilon\to 0} \frac{1}{\omega_n (n-2)} \int_U f(\mathbf{y}) \left[-\int_{\partial B(\mathbf{y},\varepsilon)} \left(G\frac{\partial \phi}{\partial n_x} - \phi \frac{\partial G}{\partial n_x}\right) d\sigma(x)\right] dy \\ &= \lim_{\varepsilon\to 0} \frac{1}{\omega_n (n-2)} \int_U f(\mathbf{y}) \int_{\partial B(\mathbf{y},\varepsilon)} \phi \frac{\partial G}{\partial n_x} d\sigma(x)\, dy \end{aligned}$$

Now $\mathbf{x} \to \psi^{\mathbf{x}}(\mathbf{y})$ and its partial derivatives are continuous and so the above reduces to

$$\begin{aligned} &= \lim_{\varepsilon\to 0} \frac{1}{\omega_n (n-2)} \int_U f(\mathbf{y}) \int_{\partial B(\mathbf{y},\varepsilon)} \phi \frac{\partial r_n}{\partial n_x}(\mathbf{x}-\mathbf{y})\, d\sigma(x)\, dy \\ &= \lim_{\varepsilon\to 0} \frac{1}{\omega_n} \int_U f(\mathbf{y}) \int_{\partial B(\mathbf{y},\varepsilon)} \phi \frac{1}{\varepsilon^{n-1}} d\sigma(x)\, dy = \int_U f(\mathbf{y})\phi(\mathbf{y})\, dy. \end{aligned}$$

Similar but easier reasoning shows that

$$\int_U \left(\frac{1}{\omega_n r}\int_{\partial U} g(\mathbf{y}) \frac{r^2 - |\mathbf{x}|^2}{|\mathbf{y}-\mathbf{x}|^n} d\sigma(\mathbf{y})\right) \Delta_x \phi(\mathbf{x})\, dx = 0.$$

Therefore, if $n \geq 3$, and u is given by 12.12, then whenever $\phi \in C_c^\infty(U)$,

$$\int_U u\Delta\phi dx = \int_U f\phi dx. \tag{12.19}$$

The same result holds for $n = 2$.

Definition 12.2.12 *$\Delta u = f$ on U in the weak sense or in the sense of distributions if for all $\phi \in C_c^\infty(U)$, 12.19 holds.*

This with Lemma 12.2.11 proves the following major theorem.

Theorem 12.2.13 *Let $f \in C(\overline{U})$ or in $L^p(U)$ for $p > n/2$ and let $g \in C(\partial U)$. Then if u is given by 12.12 in the case where $n \geq 3$ or by 12.13 in the case where $n = 2$, then u solves the differential equation of the Poisson problem in the sense of distributions along with the boundary conditions.*

12.3 Properties Of Harmonic Functions

Consider the problem for $g \in C(\partial U)$.

$$\Delta u = 0 \text{ in } U,\ u = g \text{ on } \partial U.$$

When $U = B(\mathbf{x}_0, r)$, it has now been shown there exists a unique solution to the above problem satisfying $u \in C^2(U) \cap C(\overline{U})$ and it is given by the formula

$$u(\mathbf{x}) = \frac{r^2 - |\mathbf{x}-\mathbf{x}_0|^2}{\omega_n r} \int_{\partial B(\mathbf{x}_0,r)} \frac{g(\mathbf{y})}{|\mathbf{y}-\mathbf{x}|^n} d\sigma(y) \tag{12.20}$$

It was also noted that this formula implies the mean value property for harmonic functions,

$$u(\mathbf{x}_0) = \frac{1}{\omega_n r^{n-1}} \int_{\partial B(\mathbf{x}_0,r)} u(\mathbf{y})\, d\sigma(y). \tag{12.21}$$

The mean value property can also be formulated in terms of an integral taken over $B(\mathbf{x}_0, r)$.

Lemma 12.3.1 *Let u be harmonic and C^2 on an open set, V and let $\overline{B(\mathbf{x}_0,r)} \subseteq V$. Then*

$$u(\mathbf{x}_0) = \frac{1}{m_n(B(\mathbf{x}_0,r))}\int_{B(\mathbf{x}_0,r)} u(\mathbf{y})\,dy$$

where here $m_n(B(\mathbf{x}_0,r))$ denotes the volume of the ball.

Proof: From the method of polar coordinates and the mean value property given in 12.21, along with the observation that $m_n(B(\mathbf{x}_0,r)) = \frac{\omega_n}{n}r^n$,

$$\begin{aligned}\int_{B(\mathbf{x}_0,r)} u(\mathbf{y})\,dy &= \int_0^r \int_{S^{n-1}} u(\mathbf{x}_0+\mathbf{y})\,\rho^{n-1}d\sigma(\mathbf{y})\,d\rho\\ &= \int_0^r \int_{\partial B(\mathbf{0},\rho)} u(\mathbf{x}_0+\mathbf{y})\,d\sigma(y)\,d\rho\\ &= u(\mathbf{x}_0)\int_0^r \omega_n\rho^{n-1}d\rho = u(\mathbf{x}_0)\frac{\omega_n}{n}r^n\\ &= u(\mathbf{x}_0)\,m_n(B(\mathbf{x}_0,r)).\end{aligned}$$

This proves the lemma. ■

There is a very interesting theorem which says roughly that the values of a nonnegative harmonic function are all comparable. It is known as Harnack's inequality.

Theorem 12.3.2 *Let U be an open set and let $u \in C^2(U)$ be a nonnegative harmonic function. Also let U_1 be a connected open set which is bounded and satisfies $\overline{U_1} \subseteq U$. Then there exists a constant, C, depending only on U_1 such that*

$$\max\left\{u(\mathbf{x}) : \mathbf{x} \in \overline{U_1}\right\} \le C\min\left\{u(\mathbf{x}) : \mathbf{x} \in \overline{U_1}\right\}$$

Proof: There is a positive distance between $\overline{U_1}$ and U^C because of compactness of $\overline{U_1}$. Therefore there exists $r > 0$ such that whenever $\mathbf{x} \in \overline{U_1}$, $B(\mathbf{x},2r) \subseteq U$. Then consider $\mathbf{x} \in \overline{U_1}$ and let $|\mathbf{x}-\mathbf{y}| < r$. Then from Lemma 12.3.1

$$\begin{aligned}u(\mathbf{x}) &= \frac{1}{m_n(B(\mathbf{x},2r))}\int_{B(\mathbf{x},2r)} u(\mathbf{z})\,dz\\ &= \frac{1}{2^n m_n(B(\mathbf{x},r))}\int_{B(\mathbf{x},2r)} u(\mathbf{z})\,dz\\ &\ge \frac{1}{2^n m_n(B(\mathbf{y},r))}\int_{B(\mathbf{y},r)} u(\mathbf{z})\,dz = \frac{1}{2^n}u(\mathbf{y}).\end{aligned}$$

The fact that $u \ge 0$ is used in going to the last line. Since $\overline{U_1}$ is compact, there exist finitely many balls having centers in $\overline{U_1}$, $\{B(\mathbf{x}_i,r)\}_{i=1}^m$ such that

$$\overline{U_1} \subseteq \cup_{i=1}^m B(\mathbf{x}_i, r/2).$$

Furthermore each of these balls must have nonempty intersection with at least one of the others because if not, it would follow that $\overline{U_1}$ would not be connected. Letting $\mathbf{x},\mathbf{y} \in U_1$, there must be a sequence of these balls, $B_1, B_2, \cdots, B_k$ such that $\mathbf{x} \in B_1, \mathbf{y} \in B_k$, and $B_i \cap B_{i+1} \ne \emptyset$ for $i = 1,2,\cdots,k-1$. Therefore, picking a point, $\mathbf{z}_{i+1} \in B_i \cap B_{i+1}$, the above estimate implies

$$u(\mathbf{x}) \ge \frac{1}{2^n}u(\mathbf{z}_2),\ u(\mathbf{z}_2) \ge \frac{1}{2^n}u(\mathbf{z}_3),\ u(\mathbf{z}_3) \ge \frac{1}{2^n}u(\mathbf{z}_4),\cdots,\ u(\mathbf{z}_k) \ge \frac{1}{2^n}u(\mathbf{y}).$$

Therefore,

$$u(\mathbf{x}) \ge \left(\frac{1}{2^n}\right)^k u(\mathbf{y}) \ge \left(\frac{1}{2^n}\right)^m u(\mathbf{y}).$$

Therefore, for all $\mathbf{x} \in \overline{U_1}$,

$$\sup\{u(\mathbf{y}) : \mathbf{y} \in U_1\} \leq (2^n)^m u(\mathbf{x})$$

and so

$$\max\left\{u(\mathbf{x}) : \mathbf{x} \in \overline{U_1}\right\} = \sup\{u(\mathbf{y}) : \mathbf{y} \in U_1\}$$
$$\leq (2^n)^m \inf\{u(\mathbf{x}) : \mathbf{x} \in U_1\} = (2^n)^m \min\left\{u(\mathbf{x}) : \mathbf{x} \in \overline{U_1}\right\}.$$

This proves the inequality. ■

The next theorem comes from the representation formula for harmonic functions given above.

Theorem 12.3.3 *Let U be an open set and suppose $u \in C^2(U)$ and u is harmonic. Then in fact, $u \in C^\infty(U)$. That is, u possesses all partial derivatives and they are all continuous.*

Proof: Let $B(\mathbf{x}_0,r) \subseteq U$. I will show that $u \in C^\infty(B(\mathbf{x}_0,r))$. From 12.20, it follows that for $\mathbf{x} \in B(\mathbf{x}_0,r)$,

$$\frac{r^2 - |\mathbf{x}-\mathbf{x}_0|^2}{\omega_n r} \int_{\partial B(\mathbf{x}_0,r)} \frac{u(\mathbf{y})}{|\mathbf{y}-\mathbf{x}|^n} d\sigma(y) = u(\mathbf{x}).$$

It is obvious that $\mathbf{x} \to \frac{r^2-|\mathbf{x}-\mathbf{x}_0|^2}{\omega_n r}$ is infinitely differentiable. Therefore, consider

$$\mathbf{x} \to \int_{\partial B(\mathbf{x}_0,r)} \frac{u(\mathbf{y})}{|\mathbf{y}-\mathbf{x}|^n} d\sigma(y). \tag{12.22}$$

Take $\mathbf{x} \in B(\mathbf{x}_0, r)$ and consider a difference quotient for $t \neq 0$.

$$\left(\int_{\partial B(\mathbf{x}_0,r)} u(\mathbf{y}) \frac{1}{t}\left(\frac{1}{|\mathbf{y}-(\mathbf{x}+t\mathbf{e}_k)|^n} - \frac{1}{|\mathbf{y}-\mathbf{x}|^n}\right) d\sigma(y)\right)$$

Then by the mean value theorem, the term

$$\frac{1}{t}\left(\frac{1}{|\mathbf{y}-(\mathbf{x}+t\mathbf{e}_k)|^n} - \frac{1}{|\mathbf{y}-\mathbf{x}|^n}\right)$$

equals

$$-n\,|\mathbf{x}+t\theta(t)\,\mathbf{e}_k - \mathbf{y}|^{-(n+2)}\,(x_k + \theta(t)\,t - y_k)$$

and as $t \to 0$, this converges uniformly for $\mathbf{y} \in \partial B(\mathbf{x}_0, r)$ to

$$-n\,|\mathbf{x}-\mathbf{y}|^{-(n+2)}\,(x_k - y_k).$$

This uniform convergence implies you can take a partial derivative of the function of $\mathbf{x}$ given in 12.22 obtaining the partial derivative with respect to x_k equals

$$\int_{\partial B(\mathbf{x}_0,r)} \frac{-n\,(x_k - y_k)\,u(\mathbf{y})}{|\mathbf{y}-\mathbf{x}|^{n+2}} d\sigma(y).$$

Now exactly the same reasoning applies to this function of $\mathbf{x}$ yielding a similar formula. The continuity of the integrand as a function of $\mathbf{x}$ implies continuity of the partial derivatives. The idea is there is never any problem because $\mathbf{y} \in \partial B(\mathbf{x}_0, r)$ and $\mathbf{x}$ is a given point not on this boundary. This proves the theorem. ■

Liouville's theorem is a famous result in complex variables which asserts that an entire bounded function is constant. A similar result holds for harmonic functions.

Theorem 12.3.4 *(Liouville's theorem) Suppose u is harmonic on $\mathbb{R}^n$ and is bounded. Then u is constant.*

Proof: From the Poisson formula

$$\frac{r^2-|\mathbf{x}|^2}{\omega_n r}\int_{\partial B(\mathbf{0},r)}\frac{u(\mathbf{y})}{|\mathbf{y}-\mathbf{x}|^n}d\sigma(y)=u(\mathbf{x}).$$

Now from the discussion above,

$$\frac{\partial u(\mathbf{x})}{\partial x_k}=\frac{-2x_k}{\omega_n r}\int_{\partial B(\mathbf{x}_0,r)}\frac{u(\mathbf{y})}{|\mathbf{y}-\mathbf{x}|^n}d\sigma(y)+\frac{r^2-|\mathbf{x}|^2}{\omega_n r}\int_{\partial B(\mathbf{0},r)}\frac{u(\mathbf{y})(y_k-x_k)}{|\mathbf{y}-\mathbf{x}|^{n+2}}d\sigma(y)$$

Therefore, letting $|u(\mathbf{y})|\le M$ for all $\mathbf{y}\in\mathbb{R}^n$,

$$\begin{aligned}\left|\frac{\partial u(\mathbf{x})}{\partial x_k}\right| &\le \frac{2|\mathbf{x}|}{\omega_n r}\int_{\partial B(\mathbf{x}_0,r)}\frac{M}{(r-|\mathbf{x}|)^n}d\sigma(y)+\frac{\left(r^2-|\mathbf{x}|^2\right)M}{\omega_n r}\int_{\partial B(\mathbf{0},r)}\frac{1}{(r-|\mathbf{x}|)^{n+1}}d\sigma(y)\\ &= \frac{2|\mathbf{x}|}{\omega_n r}\frac{M}{(r-|\mathbf{x}|)^n}\omega_n r^{n-1}+\frac{\left(r^2-|\mathbf{x}|^2\right)M}{\omega_n r}\frac{1}{(r-|\mathbf{x}|)^{n+1}}\omega_n r^{n-1}\end{aligned}$$

and these terms converge to 0 as $r\to\infty$. Since the inequality holds for all $r>|\mathbf{x}|$, it follows $\frac{\partial u(\mathbf{x})}{\partial x_k}=0$. Similarly all the other partial derivatives equal zero as well and so u is a constant. This proves the theorem. ■

12.4 Laplace's Equation For General Sets

Here I will consider the Laplace equation with Dirichlet boundary conditions on a general bounded open set, U. Thus the problem of interest is

$$\Delta u=0 \text{ on } U, \text{ and } u=g \text{ on } \partial U.$$

I will be presenting Perron's method for this problem. This method is based on exploiting properties of subharmonic functions which are functions satisfying the following definition.

Definition 12.4.1 *Let U be an open set and let u be a function defined on U. Then u is subharmonic if it is continuous and for all $\mathbf{x}\in U$,*

$$u(\mathbf{x})\le\frac{1}{\omega_n r^{n-1}}\int_{\partial B(\mathbf{x},r)}u(\mathbf{y})\,d\sigma \tag{12.23}$$

whenever r is small enough.

Compare with Corollary 12.2.9.

12.4.1 Properties Of Subharmonic Functions

The first property is a maximum principle. Compare to Theorem 12.2.1.

Theorem 12.4.2 *Suppose U is a bounded open set and u is subharmonic on U and continuous on $\overline{U}$. Then*

$$\max\left\{u(\mathbf{y}):\mathbf{y}\in\overline{U}\right\}=\max\left\{u(\mathbf{y}):\mathbf{y}\in\partial U\right\}.$$

Proof: Suppose $\mathbf{x} \in U$ and $u(\mathbf{x}) = \max\left\{u(\mathbf{y}) : \mathbf{y} \in \overline{U}\right\} \equiv M$. Let V denote the connected component of U which contains $\mathbf{x}$. Then since u is subharmonic on V, it follows that for all small $r > 0$, $u(\mathbf{y}) = M$ for all $\mathbf{y} \in \partial B(\mathbf{x}, r)$. Therefore, there exists some $r_0 > 0$ such that $u(\mathbf{y}) = M$ for all $\mathbf{y} \in B(\mathbf{x}, r_0)$ and this shows $\{\mathbf{x} \in V : u(\mathbf{x}) = M\}$ is an open subset of V. However, since u is continuous, it is also a closed subset of V. Therefore, since V is connected,

$$\{\mathbf{x} \in V : u(\mathbf{x}) = M\} = V$$

and so by continuity of u, it must be the case that $u(\mathbf{y}) = M$ for all $\mathbf{y} \in \partial V \subseteq \partial U$. This proves the theorem because $M = u(\mathbf{y})$ for some $\mathbf{y} \in \partial U$. ■

As a simple corollary, the proof of the above theorem shows the following startling result.

Corollary 12.4.3 *Suppose U is a connected open set and that u is subharmonic on U. Then either*

$$u(\mathbf{x}) < \sup\{u(\mathbf{y}) : \mathbf{y} \in U\}$$

for all $\mathbf{x} \in U$ or

$$u(\mathbf{x}) \equiv \sup\{u(\mathbf{y}) : \mathbf{y} \in U\}$$

for all $\mathbf{x} \in U$.

The next result indicates that the maximum of any finite list of subharmonic functions is also subharmonic.

Lemma 12.4.4 *Let U be an open set and let $u_1, u_2, \cdots, u_p$ be subharmonic functions defined on U. Then letting*

$$v \equiv \max(u_1, u_2, \cdots, u_p),$$

it follows that v is also subharmonic.

Proof: Let $\mathbf{x} \in U$. Then whenever r is small enough to satisfy the subharmonicity condition for each u_i.

$$\begin{aligned} v(\mathbf{x}) &= \max(u_1(\mathbf{x}), u_2(\mathbf{x}), \cdots, u_p(\mathbf{x})) \\ &\leq \max\left(\frac{1}{\omega_n r^{n-1}}\int_{\partial B(\mathbf{x},r)} u_1(\mathbf{y})\, d\sigma(y), \cdots, \frac{1}{\omega_n r^{n-1}}\int_{\partial B(\mathbf{x},r)} u_p(\mathbf{y})\, d\sigma(y)\right) \\ &\leq \frac{1}{\omega_n r^{n-1}}\int_{\partial B(\mathbf{x},r)} \max(u_1, u_2, \cdots, u_p)(\mathbf{y})\, d\sigma(y) = \frac{1}{\omega_n r^{n-1}}\int_{\partial B(\mathbf{x},r)} v(\mathbf{y})\, d\sigma(y). \end{aligned}$$

This proves the lemma. ■

The next lemma concerns modifying a subharmonic function on an open ball in such a way as to make the new function harmonic on the ball. Recall Corollary 12.2.8 which I will list here for convenience.

Corollary 12.4.5 *Let $U = B(\mathbf{x}_0, r)$ and let $g \in C(\partial U)$. Then there exists a unique solution $u \in C^2(U) \cap C\left(\overline{U}\right)$ to the problem*

$$\Delta u = 0 \text{ in } U,\ u = g \text{ on } \partial U.$$

This solution is given by the formula,

$$u(\mathbf{x}) = \frac{1}{\omega_n r}\int_{\partial U} g(\mathbf{y})\frac{r^2 - |\mathbf{x} - \mathbf{x}_0|^2}{|\mathbf{y} - \mathbf{x}|^n} d\sigma(\mathbf{y}) \tag{12.24}$$

for every $n \geq 2$. Here $\omega_2 = 2\pi$.

Definition 12.4.6 *Let U be an open set and let u be subharmonic on U. Then for $\overline{B(\mathbf{x}_0,r)} \subseteq U$ define*

$$u_{\mathbf{x}_0,r}(\mathbf{x}) \equiv \begin{cases} u(\mathbf{x}) & \text{if } \mathbf{x} \notin B(\mathbf{x}_0,r) \\ \frac{1}{\omega_n r}\int_{\partial B(\mathbf{x}_0,r)} u(\mathbf{y}) \frac{r^2-|\mathbf{x}-\mathbf{x}_0|^2}{|\mathbf{y}-\mathbf{x}|^n} d\sigma(\mathbf{y}) & \text{if } \mathbf{x} \in B(\mathbf{x}_0,r) \end{cases}$$

Thus $u_{\mathbf{x}_0,r}$ is harmonic on $B(\mathbf{x}_0,r)$, and equals to u off $B(\mathbf{x}_0,r)$. The wonderful thing about this is that $u_{\mathbf{x}_0,r}$ is still subharmonic on all of U. Also note that from Corollary 12.2.9 on Page 345 every harmonic function is subharmonic.

Lemma 12.4.7 *Let U be an open set and $\overline{B(\mathbf{x}_0,r)} \subseteq U$ as in the above definition. Then $u_{\mathbf{x}_0,r}$ is subharmonic on U and $u \leq u_{\mathbf{x}_0,r}$.*

Proof: First I show that $u \leq u_{\mathbf{x}_0,r}$. This follows from the maximum principle. Here is why. The function $u - u_{\mathbf{x}_0,r}$ is subharmonic on $B(\mathbf{x}_0,r)$ and equals zero on $\partial B(\mathbf{x}_0,r)$. Here is why: For $\mathbf{z} \in B(\mathbf{x}_0,r)$,

$$u(\mathbf{z}) - u_{\mathbf{x}_0 r}(\mathbf{z}) = u(\mathbf{z}) - \frac{1}{\omega\rho^{n-1}}\int_{\partial B(\mathbf{z},\rho)} u_{\mathbf{x}_0,r}(\mathbf{y})\, d\sigma(\mathbf{y})$$

for all ρ small enough. This is by the mean value property of harmonic functions and the observation that $u_{\mathbf{x}_0 r}$ is harmonic on $B(\mathbf{x}_0,r)$. Therefore, from the fact that u is subharmonic,

$$u(\mathbf{z}) - u_{\mathbf{x}_0 r}(\mathbf{z}) \leq \frac{1}{\omega\rho^{n-1}}\int_{\partial B(\mathbf{z},\rho)} (u(\mathbf{y}) - u_{\mathbf{x}_0,r}(\mathbf{y}))\, d\sigma(\mathbf{y})$$

Therefore, for all $\mathbf{x} \in B(\mathbf{x}_0,r)$,

$$u(\mathbf{x}) - u_{\mathbf{x}_0,r}(\mathbf{x}) \leq 0.$$

The two functions are equal off $B(\mathbf{x}_0,r)$.

The condition for being subharmonic is clearly satisfied at every point, $\mathbf{x} \notin \overline{B(\mathbf{x}_0,r)}$. It is also satisfied at every point of $B(\mathbf{x}_0,r)$ thanks to the mean value property, Corollary 12.2.9 on Page 345. It is only at the points of $\partial B(\mathbf{x}_0,r)$ where the condition needs to be checked. Let $\mathbf{z} \in \partial B(\mathbf{x}_0,r)$. Then since u is given to be subharmonic, it follows that for all r small enough,

$$\begin{aligned} u_{\mathbf{x}_0,r}(\mathbf{z}) &= u(\mathbf{z}) \leq \frac{1}{\omega_n r^{n-1}}\int_{\partial B(\mathbf{x}_0,r)} u(\mathbf{y})\, d\sigma \\ &\leq \frac{1}{\omega_n r^{n-1}}\int_{\partial B(\mathbf{x}_0,r)} u_{\mathbf{x}_0,r}(\mathbf{y})\, d\sigma. \end{aligned}$$

This proves the lemma. ■

Definition 12.4.8 *For U a bounded open set and $g \in C(\partial U)$, define*

$$w_g(\mathbf{x}) \equiv \sup\{u(\mathbf{x}) : u \in S_g\}$$

where S_g consists of those functions u which are subharmonic with $u(\mathbf{y}) \leq g(\mathbf{y})$ for all $\mathbf{y} \in \partial U$ and $u(\mathbf{y}) \geq \min\{g(\mathbf{y}) : \mathbf{y} \in \partial U\} \equiv m$.

Note that $S_g \neq \emptyset$ because $u(\mathbf{x}) \equiv m$ is a member of S_g. Also all functions in S_g have values between m and $\max\{g(\mathbf{y}) : \mathbf{y} \in \partial U\}$. The fundamental result is the following absolutely amazing incredible result.

Proposition 12.4.9 *Let U be a bounded open set and let $g \in C(\partial U)$. Then $w_g \in S_g$ and in addition to this, w_g is harmonic.*

Proof: Let $\overline{B(\mathbf{x}_0, 2r)} \subseteq U$ and let $\{\mathbf{x}_k\}_{k=1}^{\infty}$ denote a countable dense subset of $\overline{B(\mathbf{x}_0, r)}$. Let $\{u_{1k}\}$ denote a sequence of functions of S_g with the property that

$$\lim_{k\to\infty} u_{1k}(\mathbf{x}_1) = w_g(\mathbf{x}_1).$$

By Lemma 12.4.7, it can be assumed each u_{1k} is a harmonic function in $B(\mathbf{x}_0, 2r)$ since otherwise, you could use the process of replacing u with $u_{\mathbf{x}_0,2r}$. Similarly, for each l, there exists a sequence of harmonic functions in S_g, $\{u_{lk}\}$ with the property that

$$\lim_{k\to\infty} u_{lk}(\mathbf{x}_l) = w_g(\mathbf{x}_l).$$

Now define

$$w_k = (\max(u_{1k}, \cdots, u_{kk}))_{\mathbf{x}_0,2r}.$$

Then each $w_k \in S_g$, each w_k is harmonic in $B(\mathbf{x}_0, 2r)$, and for each $\mathbf{x}_l$,

$$\lim_{k\to\infty} w_k(\mathbf{x}_l) = w_g(\mathbf{x}_l).$$

For $\mathbf{x} \in \overline{B(\mathbf{x}_0, r)}$

$$w_k(\mathbf{x}) = \frac{1}{\omega_n 2r}\int_{\partial B(\mathbf{x}_0,2r)} w_k(\mathbf{y})\frac{r^2 - |\mathbf{x}-\mathbf{x}_0|^2}{|\mathbf{y}-\mathbf{x}|^n}d\sigma(\mathbf{y}) \tag{12.25}$$

and so there exists a constant, C which is independent of k such that for all $i = 1, 2, \cdots, n$ and $\mathbf{x} \in \overline{B(\mathbf{x}_0, r)}$,

$$\left|\frac{\partial w_k(\mathbf{x})}{\partial x_i}\right| \leq C$$

Therefore, this set of functions, $\{w_k\}$ is equicontinuous on $\overline{B(\mathbf{x}_0, r)}$ as well as being uniformly bounded and so by the Ascoli Arzela theorem, it has a subsequence which converges uniformly on $\overline{B(\mathbf{x}_0, r)}$ to a continuous function I will denote by w which has the property that for all k,

$$w(\mathbf{x}_k) = w_g(\mathbf{x}_k) \tag{12.26}$$

Also since each w_k is harmonic,

$$w_k(\mathbf{x}) = \frac{1}{\omega_n r}\int_{\partial B(\mathbf{x}_0,r)} w_k(\mathbf{y})\frac{r^2 - |\mathbf{x}-\mathbf{x}_0|^2}{|\mathbf{y}-\mathbf{x}|^n}d\sigma(\mathbf{y}) \tag{12.27}$$

Passing to the limit in 12.27 using the uniform convergence, it follows

$$w(\mathbf{x}) = \frac{1}{\omega_n r}\int_{\partial B(\mathbf{x}_0,r)} w(\mathbf{y})\frac{r^2 - |\mathbf{x}-\mathbf{x}_0|^2}{|\mathbf{y}-\mathbf{x}|^n}d\sigma(\mathbf{y}) \tag{12.28}$$

which shows that w is also harmonic. I have shown that $w = w_g$ on a dense set. Also, it follows that $w(\mathbf{x}) \leq w_g(\mathbf{x})$ for all $\mathbf{x} \in \overline{B(\mathbf{x}_0, r)}$. It remains to verify these two functions are in fact equal.

Claim: w_g is lower semicontinuous on U.

Proof of claim: Suppose $\mathbf{z}_k \to \mathbf{z}$. I need to verify that

$$\liminf_{k\to\infty} w_g(\mathbf{z}_k) \geq w_g(\mathbf{z}).$$

Let $\varepsilon > 0$ be given and pick $u \in S_g$ such that $w_g(\mathbf{z}) - \varepsilon < u(\mathbf{z})$. Then

$$w_g(\mathbf{z}) - \varepsilon < u(\mathbf{z}) = \liminf_{k\to\infty} u(\mathbf{z}_k) \leq \liminf_{k\to\infty} w_g(\mathbf{z}_k).$$

Since ε is arbitrary, this proves the claim.

Using the claim, let $\mathbf{x} \in \overline{B(\mathbf{x}_0, r)}$ and pick $\mathbf{x}_{k_l} \to \mathbf{x}$ where $\{\mathbf{x}_{k_l}\}$ is a subsequence of the dense set, $\{\mathbf{x}_k\}$. Then

$$w_g(\mathbf{x}) \geq w(\mathbf{x}) = \liminf_{l\to\infty} w(\mathbf{x}_{k_l}) = \liminf_{l\to\infty} w_g(\mathbf{x}_{k_l}) \geq w_g(\mathbf{x}).$$

This proves $w = w_g$ and since w is harmonic, so is w_g. This proves the proposition. ■

It remains to consider whether the boundary values are assumed. This requires an additional assumption on the set, U. It is a remarkably mild assumption, however.

Definition 12.4.10 *A bounded open set, U has the barrier condition at $\mathbf{z} \in \partial U$, if there exists a function, $b_{\mathbf{z}}$ called a barrier function which has the property that $b_{\mathbf{z}}$ is subharmonic on U, $b_{\mathbf{z}}(\mathbf{z}) = 0$, and for all $\mathbf{x} \in \partial U \setminus \{\mathbf{z}\}$, $b_{\mathbf{z}}(\mathbf{x}) < 0$.*

The main result is the following remarkable theorem.

Theorem 12.4.11 *Let U be a bounded open set which has the barrier condition at $\mathbf{z} \in \partial U$ and let $g \in C(\partial U)$. Then the function, w_g, defined above is in $C^2(U)$ and satisfies*

$$\Delta w_g = 0 \text{ in } U,$$
$$\lim_{\mathbf{x}\to\mathbf{z}} w_g(\mathbf{x}) = g(\mathbf{z}).$$

Proof: From Proposition 12.4.9 it follows $\Delta w_g = 0$. Let $\mathbf{z} \in \partial U$ and let $b_{\mathbf{z}}$ be the barrier function at $\mathbf{z}$. Then letting $\varepsilon > 0$ be given, the function

$$u_-(\mathbf{x}) \equiv \max(g(\mathbf{z}) - \varepsilon + Kb_{\mathbf{z}}(\mathbf{x}), m)$$

is subharmonic for all $K > 0$.

Claim: For K large enough, $g(\mathbf{z}) - \varepsilon + Kb_{\mathbf{z}}(\mathbf{x}) \leq g(\mathbf{x})$ for all $\mathbf{x} \in \partial U$.

Proof of claim: Let $\delta > 0$ and let $B_\delta = \max\{b_{\mathbf{z}}(\mathbf{x}) : \mathbf{x} \in \partial U \setminus B(\mathbf{z}, \delta)\}$. Then $B_\delta < 0$ by assumption and the compactness of $\partial U \setminus B(\mathbf{z}, \delta)$. Choose $\delta > 0$ small enough that if $|\mathbf{x} - \mathbf{z}| < \delta$, then $g(\mathbf{x}) - g(\mathbf{z}) + \varepsilon > 0$. Then for $|\mathbf{x} - \mathbf{z}| < \delta$,

$$b_{\mathbf{z}}(\mathbf{x}) \leq \frac{g(\mathbf{x}) - g(\mathbf{z}) + \varepsilon}{K}$$

for any choice of positive K. Now choose K large enough that $B_\delta < \frac{g(\mathbf{x})-g(\mathbf{z})+\varepsilon}{K}$ for all $\mathbf{x} \in \partial U$. This can be done because $B_\delta < 0$. It follows the above inequality holds for all $\mathbf{x} \in \partial U$. This proves the claim.

Let K be large enough that the conclusion of the above claim holds. Then, for all $\mathbf{x}$, $u_-(\mathbf{x}) \leq g(\mathbf{x})$ for all $\mathbf{x} \in \partial U$ and so $u_- \in S_g$ which implies $u_- \leq w_g$ and so

$$g(\mathbf{z}) - \varepsilon + Kb_{\mathbf{z}}(\mathbf{x}) \leq w_g(\mathbf{x}). \tag{12.29}$$

This is a very nice inequality and I would like to say

$$\begin{aligned} \lim_{\mathbf{x}\to\mathbf{z}} g(\mathbf{z}) - \varepsilon + Kb_{\mathbf{z}}(\mathbf{x}) &= g(\mathbf{z}) - \varepsilon \\ &\leq \liminf_{\mathbf{x}\to\mathbf{z}} w_g(\mathbf{x}) \\ &\leq \limsup_{\mathbf{x}\to\mathbf{z}} w_g(\mathbf{x}) = w_g(\mathbf{z}) \leq g(\mathbf{z}) \end{aligned}$$

but this would be wrong because I do not know that w_g is continuous at a boundary point. I only have shown that it is harmonic in U. Therefore, a little more is required. Let

$$u_+(\mathbf{x}) \equiv g(\mathbf{z}) + \varepsilon - Kb_{\mathbf{z}}(\mathbf{x}).$$

Then $-u_+$ is subharmonic and also if K is large enough, it follows from reasoning similar to that of the above claim that

$$-u_+(\mathbf{x}) = -g(\mathbf{z}) - \varepsilon + Kb_{\mathbf{z}}(\mathbf{x}) \leq -g(\mathbf{x})$$

on ∂U. Therefore, letting $u \in S_g, u - u_+$ is a subharmonic function which satisfies for $\mathbf{x} \in \partial U$,

$$u(\mathbf{x}) - u_+(\mathbf{x}) \leq g(\mathbf{x}) - g(\mathbf{x}) = 0.$$

Consequently, the maximum principle implies $u \leq u_+$ and so since this holds for every $u \in S_g$, it follows

$$w_g(\mathbf{x}) \leq u_+(\mathbf{x}) = g(\mathbf{z}) + \varepsilon - Kb_{\mathbf{z}}(\mathbf{x}).$$

It follows that

$$g(\mathbf{z}) - \varepsilon + Kb_{\mathbf{z}}(\mathbf{x}) \leq w_g(\mathbf{x}) \leq g(\mathbf{z}) + \varepsilon - Kb_{\mathbf{z}}(\mathbf{x})$$

and so,

$$g(\mathbf{z}) - \varepsilon \leq \liminf_{\mathbf{x}\to\mathbf{z}} w_g(\mathbf{x}) \leq \limsup_{\mathbf{x}\to\mathbf{z}} w_g(\mathbf{x}) \leq g(\mathbf{z}) + \varepsilon.$$

Since ε is arbitrary, this shows

$$\lim_{\mathbf{x}\to\mathbf{z}} w_g(\mathbf{x}) = g(\mathbf{z}).$$

This proves the theorem. ■

12.4.2 Poisson's Problem Again

Corollary 12.4.12 *Let U be a bounded open set which has the barrier condition and let $f \in C(\overline{U}), g \in C(\partial U)$. Then there exists at most one solution, $u \in C^2(U) \cap C(\overline{U})$ to Poisson's problem. If there is a solution, then it is of the form*

$$u(\mathbf{x}) = \frac{-1}{(n-2)\omega_n}\left[\int_U G(\mathbf{x},\mathbf{y}) f(\mathbf{y})\,dy + \int_{\partial U} g(\mathbf{y}) \frac{\partial G}{\partial n_y}(\mathbf{x},\mathbf{y})\,d\sigma(\mathbf{y})\right], \text{ if } n \geq 3 \tag{12.30}$$

$$u(\mathbf{x}) = \frac{1}{2\pi}\left[\int_{\partial U} g(\mathbf{y}) \frac{\partial G}{\partial n_y}(\mathbf{x},\mathbf{y})\,d\sigma + \int_U G(\mathbf{x},\mathbf{y}) f(\mathbf{y})\,dx\right], \text{ if } n = 2 \tag{12.31}$$

for $G(\mathbf{x},\mathbf{y}) = r_n(\mathbf{y}-\mathbf{x}) - \psi^{\mathbf{x}}(\mathbf{y})$ where $\psi^{\mathbf{x}}$ is a function which satisfies $\psi^{\mathbf{x}} \in C^2(U) \cap C(\overline{U})$

$$\Delta\psi^{\mathbf{x}} = 0,\ \psi^{\mathbf{x}}(\mathbf{y}) = r_n(\mathbf{x}-\mathbf{y}) \text{ for } \mathbf{y} \in \partial U.$$

Furthermore, if u is given by the above representations, then u is a weak solution to Poisson's problem.

Proof: Uniqueness follows from Corollary 12.2.2 on Page 338. If u_1 and u_2 both solve the Poisson problem, then their difference, w satisfies

$$\Delta w = 0, \text{ in } U,\ w = 0 \text{ on } \partial U.$$

The same arguments used earlier show that the representations in 12.30 and 12.31 both yield a weak solution to Poisson's problem. ■

The function, G in the above representation is called Green's function. Much more can be said about the Green's function.

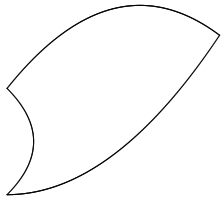
How can you recognize that a bounded open set, U has the barrier condition? One way would be to check the following condition.

Condition 12.4.13 *For each* $\mathbf{z} \in \partial U$, *there exists* $\mathbf{x_z} \notin \overline{U}$ *such that* $|\mathbf{x_z} - \mathbf{z}| < |\mathbf{x_z} - \mathbf{y}|$ *for every* $\mathbf{y} \in \partial U \setminus \{\mathbf{z}\}$.

Proposition 12.4.14 *Suppose Condition 12.4.13 holds. Then* U *satisfies the barrier condition.*

Proof: For $n \geq 3$, let $b_{\mathbf{z}}(\mathbf{y}) \equiv r_n(\mathbf{y} - \mathbf{x_z}) - r_n(\mathbf{z} - \mathbf{x_z})$. Then $b_{\mathbf{z}}(\mathbf{z}) = 0$ and if $\mathbf{y} \in \partial U$ with $\mathbf{y} \neq \mathbf{z}$, then clearly $b_{\mathbf{z}}(\mathbf{y}) < 0$. For $n = 2$, let $b_{\mathbf{z}}(\mathbf{y}) = -\ln|\mathbf{y} - \mathbf{x_z}| + \ln|\mathbf{z} - \mathbf{x_z}|$. This works out the same way. ■

Here is a picture of a domain which satisfies the barrier condition.



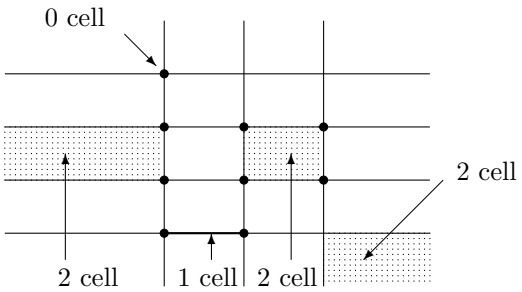
In fact, you have to have a fairly pathological example in order to find something which does not satisfy the barrier condition. You might try to think of some examples. Think of $B(\mathbf{0}, 1) \setminus \{z \text{ axis}\}$ for example. The points on the z axis which are in $B(\mathbf{0}, 1)$ become boundary points of this new set. Thus this set can't satisfy the above condition. Could this set have the barrier property?

Chapter 13

The Jordan Curve Theorem

This short chapter is devoted to giving an elementary proof of the Jordan curve theorem which is independent of the chapter on degree theory. I am following lecture notes from a topology course given by Fernley at BYU in the 1970's. The ideas used in this presentation are elementary and also lead to more general notions in algebraic topology. In addition to this, these techniques are very useful in complex analysis.

Definition 13.0.15 *A grating G is a **finite** set of horizontal and vertical lines, each of which separate the plane. The grating divides the plane into two dimensional domains the closures of which are called 2 cells of G. The 1 cells of G are the edges of the 2 cells and the 0 cells of G are the end points of the 1 cells.*



For $k = 0, 1, 2$, one speaks of k chains. For $\{a_j\}_{j=1}^n$ a set of k cells, the k chain is denoted as a formal sum

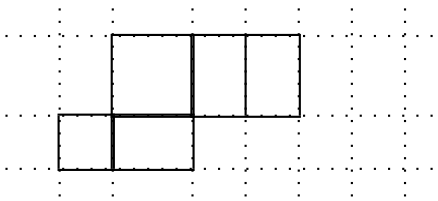
$$C = a_1 + a_2 + \cdots + a_n$$

where the sum is taken modulo 2. The sums are just formal expressions like the above. Thus for a a k cell, $a + a = 0, 0 + a = a$, the summation sign is commutative. In other words, if a k cell is repeated an even number of times in the formal sum, it disappears resulting in 0 defined by $0 + a = a + 0 = a$. For a a k cell, $|a|$ denotes the points of the plane which are contained in a. For a k chain, C as above,

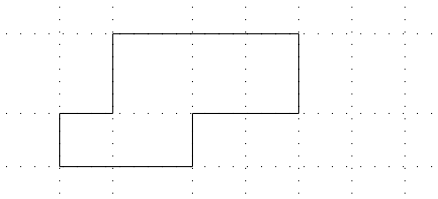
$$|C| \equiv \{x : x \in |a_j| \text{ for some } a_j\}$$

so $|C|$ is the union of the k cells in the sum remembering that when a k cell occurs twice, it is gone and does not contribute to $|C|$.

The following picture illustrates the above definition. The following is a picture of the 2 cells in a 2 chain. The dotted lines indicate the lines in the grating.



Now the following is a picture of the 1 chain consisting of the sum of the 1 cells which are the edges of the above 2 cells. Remember when a 1 cell is added to itself, it disappears from the chain. Thus if you add up the 1 cells which are the edges of the above 2 cells, lots of them cancel off. In fact all the edges which are shared between two 2 cells disappear. The following is what results.



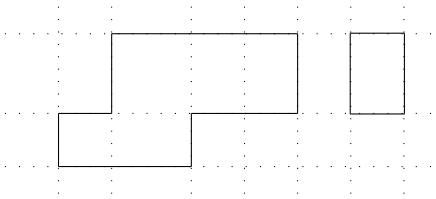
Definition 13.0.16 *Next the boundary operator is defined. This is denoted by ∂. ∂ takes k cells to $k-1$ chains. If a is a 2 cell, then ∂a consists of the edges of a. If a is a 1 cell, then ∂a consists of the ends of the 1 cell. If a is a 0 cell, then $\partial a \equiv 0$. This extends in a natural way to k chains. For*

$$C = a_1 + a_2 + \cdots + a_n,$$

$$\partial C \equiv \partial a_1 + \partial a_2 + \cdots + \partial a_n$$

A k chain C is called a cycle if $\partial C = 0$.

In the second of the above pictures, you have a 1 cycle. Here is a picture of another one in which the boundary of another 2 cell has been included over on the right.



This 1 cycle shown above is the boundary of exactly two 2 chains. What are they? C_1 consists of the 2 cells in the first picture above along with the 2 cell whose boundary is the 1 cycle over on the right. C_2 is all the other 2 cells of the grating. You see this clearly works. Could you make that 2 cell on the right be in C_2? No, you couldn't do it. This is because the 1 cells which are shown would disappear, being listed twice.

This illustrates the fundamental lemma of the plane which comes next.

Lemma 13.0.17 *If C is a bounded 1 cycle ($\partial C = 0$), then there are exactly two 2 chains D_1, D_2 such that*

$$C = \partial D_1 = \partial D_2.$$

Proof: The lemma is vacuously true unless there are at least two vertical lines and at least two horizontal lines in the grating G. It is also obviously true if there are exactly two vertical lines and two horizontal lines in G. Suppose the theorem is true for n lines in G. Then as just mentioned, there is nothing to prove unless there are either 2 or more

vertical lines and two or more horizontal lines. Suppose without loss of generality there are at least as many veritical lines are there are horizontal lines and that this number is at least 3. If it is only two, there is nothing left to show. Let l be the second vertical line from the left. Let $\{e_1, \cdots, e_m\}$ be the 1 cells of C with the property that $|e_j| \subseteq l$. Note that e_j occurs only once in C since if it occurred twice, it would disappear because of the rule for addition. Pick one of the 2 cells adjacent to e_j, b_j and add in ∂b_j which is a 1 cycle. Thus

$$C + \sum_j \partial b_j$$

is a bounded 1 cycle and it has the property that it has no 1 cells contained in l. Thus you could eliminate l from the grating G and all the 1 cells of the above 1 chain are edges of the grating $G \setminus \{l\}$. By induction, there are exactly two 2 chains D_1, D_2 composed of 2 cells of $G \setminus \{l\}$ such that for $i = 1, 2$,

$$\partial D_i = C + \sum_j \partial b_j \tag{13.1}$$

Since none of the 2 cells of D_i have any edges on l, one can add l back in and regard D_1 and D_2 as 2 chains in G. Therefore, adding $\sum_j \partial b_j$ to both sides of the above yields

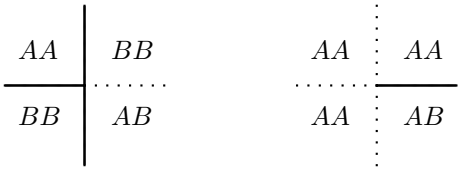
$$C = \partial D_i + \sum_j \partial b_j = \partial \left(D_i + \sum_j b_j \right), i = 1, 2.$$

and this shows there exist two 2 chains which have C as the boundary. If $\partial D_i' = C$, then

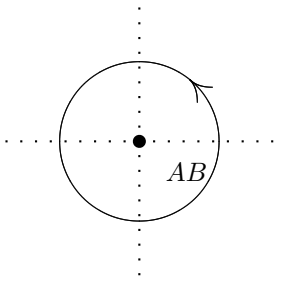
$$\partial D_i' + \sum_j \partial b_j = \partial \left(D_i' + \sum_j b_j \right) = C + \sum_j \partial b_j$$

and by induction, there are exactly two 2 chains which $D_i' + \sum_j b_j$ can equal. Thus adding $\sum_j b_j$ there are exactly two 2 chains which D_i' can equal.

Here is another proof which is not by induction. This proof also gives an algorithm for identifying the two 2 chains. The 1 cycle is bounded and so every 1 cell in it is part of the boundary of a 2 cell which is bounded. For the unbounded 2 cells on the left, label them all as A. Now starting from the left and moving toward the right, toggle between A and B every time you hit a vertical 1 cell of C. This will label every 2 cell with either A or B. Next, starting at the top, label all the unbounded 2 cells as A and move down and toggle between A and B every time you encounter a horizontal 1 cell of C. This also labels every 2 cell as either A or B. Suppose there is a contradiction in the labeling. Pick the first column in which a contradiction occurs and then pick the top contradictory 2 cell in this column. There are various cases which can occur, each leading to the existence of a vertex of C which is contained in an odd number of 1 cells of C, thus contradicting the conclusion that C is a 1 cycle. In the following picture, AB will mean the labeling from the left to right gives A and the labeling from top to bottom yields B with similar modification for AA and BB.



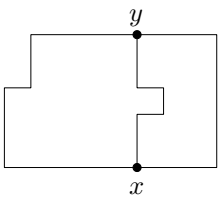
A solid line indicates the corresponding 1 cell is in C. It is there because a change took place either from top to bottom or from left to right. Note that in both of those situations the vertex right in the middle of the crossed lines will occur in ∂C and so C is not a 1 cycle. There are 8 similar pictures you can draw and in each case this happens. The vertex in the center gets added in an odd number of times. You can also notice that if you start with the contradictory 2 cell and move counter clockwise, crossing 1 cells as you go and starting with B, you must end up at A as a result of crossing 1 cells of C and this requires crossing either one or three of these 1 cells of C.



Thus that center vertex is a boundary point of C and so C is not a 1 cycle after all. Similar considerations would hold if the contradictory 2 cell were labeled BA. Thus there can be no contradiction in the two labeling schemes. They label the 2 cells in G either A or B in an unambiguous manner.

The labeling algorithm encounters every 1 cell of C (in fact of G) and gives a label to every 2 cell of G. Define the two 2 chains as A and B where A consists of those labeled as A and B those labeled as B. The 1 cells which cause a change to take place in the labeling are exactly those in C and each is contained in one 2 cell from A and one 2 cell from B. Therefore, each of these 1 cells of C appears in ∂A and ∂B which shows $C \subseteq \partial A$ and $C \subseteq \partial B$. On the other hand, if l is a 1 cell in ∂A, then it can only occur in a single 2 cell of A and so the 2 cell adjacent to that one along l must be in B and so l is one of the 1 cells of C by definition. As to uniqueness, in moving from left to right, you must assign adjacent 2 cells joined at a 1 cell of C to different 2 chains or else the 1 cell would not appear when you take the boundary of either A or B since it would be added in twice. Thus there are exactly two 2 chains with the desired property. ■

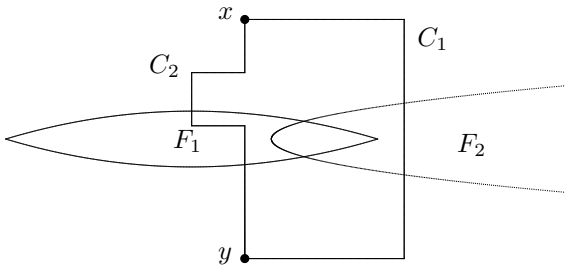
The next lemma is interesting because it gives the existence of a continuous curve joining two points.



Lemma 13.0.18 *Let C be a bounded 1 chain such that $\partial C = x + y$. Then both x, y are contained in a continuous curve which is a subset of $|C|$.*

Proof: There are an odd number of 1 cells of C which have x at one end. Otherwise $\partial C \neq x + y$. Begin at x and move along an edge leading away from x. Continue till there is no new edge to travel along. You must be at y since otherwise, you would have found another boundary point of C. This point would be in either one or three one cells of C. It can't be x because x is contained in either one or three one cells of C. Thus, there is always a way to leave x if the process returns to it. IT follows that there is a continuous curve in $|C|$ joining x to y. ■

The next lemma gives conditions under which you can go around a couple of closed sets. It is called Alexander's lemma. The following picture is a rough illustration of the situation. Roughly, it says that if you can mis F_1 and you can mis F_2 in going from x to y, then you can mis both F_1 and F_2 by climbing around F_1.



Lemma 13.0.19 *Let F_1 be compact and F_2 closed. Suppose C_1, C_2 are two bounded 1 chains in a grating which has no unbounded two cells having nonempty intersection with F_1.Suppose $\partial C_i = x + y$ where $x, y \notin F_1 \cup F_2$. Suppose C_2 does not intersect F_2 and C_1 does not intersect F_1. Also suppose the 1 cycle $C_1 + C_2$ bounds a 2 chain D for which $|D| \cap F_1 \cap F_2 = \emptyset$. Then there exists a 1 chain C such that $\partial C = x + y$ and $|\partial C| \cap (F_1 \cup F_2) = \emptyset$. In particular x, y cannot be in different components of the complement of $F_1 \cup F_2$.*

Proof: Let $a_1, a_2, \cdots, a_m$ be the 2 cells of D which intersect the compact set F_1. Consider

$$C \equiv C_2 + \sum_k \partial a_k.$$

This is a 1 chain and $\partial C = x + y$ because $\partial \partial a_k = 0$. Then $|a_k| \cap F_2 = \emptyset$. This is because $|a_k| \cap F_1 \neq \emptyset$ and none of the 2 cells of D intersect both F_1 and F_2 by assumption. Therefore, C is a bounded 1 chain which avoids intersecting F_2.

Does it also avoid F_1? Suppose to the contrary that l is a one cell of C which does intersect F_1. If $|l| \subseteq |C_1 + C_2|$, then it would be an edge of some 2 cell of D and would have to be a 1 cell of C_2 since it intersects F_1 so it would have been added twice, once from C_2 and once from

$$\sum_k \partial a_k$$

and therefore could not be a summand in C. Therefore, $|l|$ is not in $|C_1 + C_2|$. It follows l must be an edge of some $a_k \in D$ and it is not a 1 cell of $C_1 + C_2$. Therefore, if b is the 2 cell adjacent to a_k, it must follow $b \in D$ since otherwise l would be a 1 cell of $C_1 + C_2$ the boundary of D and this was just ruled out. But now it would follow that l would occur twice in the above sum so l cannot be a summand of C. Therefore C misses F_1 also.

Here is another argument. Suppose $|l| \cap F_1 \neq \emptyset$. $l \in C = C_2 + \sum_k \partial a_k$. First note that $l \notin C_1$ since $|C_1| \cap F_1 = \emptyset$.

Case 1: $l \in C_2$.

In this case it is in $C_1 + C_2$ because, as just noted it is not in C_1. Therefore, there exists $a \in D$ such that l is an edge of a and is not in the two cell adjacent to a. But this would require l to disappear since it would occur in both C_2 and $\sum_k \partial a_k$. Hence $l \notin C_2$.

Case 2: In this case $l \notin C_2$. Then l is the edge of some $a \in D$ which intersects F_1. Letting b be the two cell adjacent to a sharing l, then b cannot be in D since otherwise l would occur twice in the above sum and would then disappear. Hence $b \notin D$ and $l \in \partial D = C_1 + C_2$ but this cannot happen because $l \notin C_1$ and in this case $l \notin C_2$ either. ■

Lemma 13.0.20 *Let C be a bounded 1 cycle such that $|C| \cap H = \emptyset$ where H is a connected set. Also let D, E be the 2 chains with $\partial D = C = \partial E$. Then either $|H| \subseteq |E|$ or $|H| \subseteq |D|$.*

Proof: If p is a limit point of $|E|$ and $p \in |D|$, then p must be contained in an edge of some 2 cell of D since otherwise it could not be a limit point, being contained in an open set whose intersection with $|E|$ is empty. If p is a point of an even number of edges of 2 cells of D, then it is likewise an interior point of $|D|$ which cannot be a limit point of $|E|$. Therefore, if p is a limit point of $|E|$, it must be the case that $p \in |C|$. A similar observation holds for the case where $p \in |E|$ and is a limit point of $|D|$. Thus if $H \cap |D|$ and $H \cap |E|$ are both nonempty, then they separate the connected set H and so H must be a subset of one of $|D|$ or $|E|$. ■

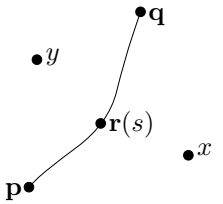
Definition 13.0.21 *A Jordan arc is a set of points of the form $\Gamma \equiv \mathbf{r}([a,b])$ where $\mathbf{r}$ is a one to one map from $[a,b]$ to the plane. For $p, q \in \Gamma$, say $p < q$ if $p = \mathbf{r}(t_1), q = \mathbf{r}(t_2)$ for $t_1 < t_2$. Also let pq denote the arc $\mathbf{r}([t_1, t_2])$.*

Theorem 13.0.22 *Let Γ be a Jordan arc. Then its complement is connected.*

Proof: Suppose this is not so. Then there exists x, y points in Γ^C which are in different components of Γ^C. Let G be a grating having x, y as points of intersection of a horizontal line and a vertical line of G and let p, q be the points at the ends of the Jordan arc. Also let G be such that no unbounded two cell has nonempty intersection with Γ. Let $p = \mathbf{r}(a)$ and $q = \mathbf{r}(b)$. Now let $z = \mathbf{r}\left(\frac{a+b}{2}\right)$ and consider the two arcs pz and zq. If $\partial C = x + y$ then it is required $|C| \cap \Gamma \neq \emptyset$ since otherwise these two points would not be in different components. Suppose there exists $C_1, \partial C_1 = x + y$ and $|C_1| \cap zq = \emptyset$ and $C_2, \partial C_2 = x + y$ but $|C_2| \cap pz = \emptyset$. Then $C_1 + C_2$ is a 1 cycle and so by Lemma 13.0.17 there are exactly two 2 chains whose boundaries are $C_1 + C_2$. Since $z \notin |C_i|$, it follows $z = pz \cap zq$ can only be in one of these 2 chains because it is a single point. Then by Lemma 13.0.19, Alexander's lemma, there exists C a 1 chain with $\partial C = x + y$ and $|C| \cap (pz \cup zq) = \emptyset$ so by Lemma 13.0.18 x, y are not in different components of Γ^C contrary to the assumption they are in different components. Hence one of pz, zq has the property that every 1 chain, $\partial C = x + y$ goes through it. Say every such 1 chain goes through zq. Then let zq play the role of pq and conclude every 1 chain C such that $\partial C = x + y$ goes through either zw or wq there

$$w = \mathbf{r}\left(\left(\frac{a+b}{2} + b\right)\frac{1}{2}\right)$$

Thus, continuing this way, there is a sequence of Jordan arcs $p_k q_k$ where $\mathbf{r}(t_k) = q_k$ and $\mathbf{r}(s_k) = p_k$ with $|t_k - s_k| < \frac{b-a}{2^k}, [s_k, t_k] \subseteq [a,b]$ such that every C with $\partial C = x + y$ has nonempty intersection with $p_k q_k$. The intersection of these arcs is $\mathbf{r}(s)$ where $s = \cap_{k=1}^{\infty} [s_k, t_k]$. Then all such C must go through $\mathbf{r}(s)$ because such C with $\partial C = x + y$ must intersect $p_k q_k$ for each k and their intersection is $\mathbf{r}(s)$. But now there is an obvious contradiction to having every 1 chain whose boundary is $x + y$ intersecting $\mathbf{r}(s)$.



Pick a 1 chain whose boundary is $x+y$. Let D be the two chain of at most four 2 cells consisting of those two cells which have $\mathbf{r}(s)$ on some edge. Then $\partial(C + \partial D) = \partial C =$

$x+y$ but $\mathbf{r}(s) \notin |C+\partial D|$. Therefore, this contradiction shows Γ^C must be connected after all. ■

The other important observation about a Jordan arc is that it has no interior points. This will follow later from a harder result but it is also easy to prove.

Lemma 13.0.23 *Let $\Gamma = \mathbf{r}([a,b])$ be a Jordan arc where $\mathbf{r}$ is as above, one to one, onto and continuous. Then Γ has no interior points.*

Proof: Suppose to the contrary that Γ has an interior point p. Then for some $r > 0$,

$$B(p,r) \subseteq \Gamma.$$

Consider the circles of radius $\delta < r$ centered at p. Denoting as C_δ one of these, it follows the C_δ are disjoint. Therefore, since $\mathbf{r}$ is one to one, the sets $\mathbf{r}^{-1}(C_\delta)$ are also disjoint. Now $\mathbf{r}$ is continuous and one to one mapping to a compact set. Therefore, $\mathbf{r}^{-1}$ is also continuous. It follows $\mathbf{r}^{-1}(C_\delta)$ is connected and compact. Thus by Theorem 5.3.8 each of these sets is a closed interval of positive length since $\mathbf{r}$ is one to one. It follows there exist disjoint open nonempty intervals consisting of the interiors of $\mathbf{r}^{-1}(C_\delta), \{I_\delta\}_{\delta<r}$. This is a contradiction to the density of $\mathbb{Q}$ and the fact that $\mathbb{Q}$ is at most countable. ■

Definition 13.0.24 *Let $\mathbf{r}$ map $[a,b]$ to the plane such that $\mathbf{r}$ is one to one on $[a,b)$ and $(a,b]$ but $\mathbf{r}(a) = \mathbf{r}(b)$. Then $J = \mathbf{r}([a,b])$ is called a simple closed curve. It is also called a Jordan curve. Also since the term "boundary" has been given a specialized meaning relative to chains of various sizes, we say x is in the frontier of S if every open ball containing x contains points of S as well as points of S^C.*

Note that if J is a Jordan curve, then it is the union of two Jordan arcs whose intersection is two distinct points of J. You could pick $z \in (a,b)$ and consider $\mathbf{r}([a,z])$ and $\mathbf{r}([z,b])$ as the two Jordan arcs.

The next lemma gives a probably more convenient way of thinking about a Jordan curve. It says essentially that a Jordan curve is a wriggly circle. First consider the following simple lemma.

Lemma 13.0.25 *Let K be a compact set in $\mathbb{R}^n$ and let $\mathbf{f}: K \to \mathbb{R}^m$ be continuous and one to one. Then $\mathbf{f}^{-1}: \mathbf{f}(K) \to K$ is also continuous.*

Proof: Suppose $\{\mathbf{f}(k_n)\}$ is a convergent sequence in $\mathbf{f}(K)$ converging to $\mathbf{f}(k)$. Does it follow that $k_n \to k$? If not, there exists a subsequence $\{k_{n_k}\}$ which converges as $k \to \infty$ to $l \neq k$. Then by continuity of $\mathbf{f}$ it follows $\mathbf{f}(k_{n_k}) \to \mathbf{f}(l)$. Hence $\mathbf{f}(l) = \mathbf{f}(k)$ which violates the condition that $\mathbf{f}$ is one to one.

Lemma 13.0.26 *J is a simple closed curve if and only if there exists a mapping $\boldsymbol{\theta}: S^1 \to J$ where S^1 is the unit circle*

$$\left\{(x,y): x^2+y^2 = 1\right\},$$

such that $\boldsymbol{\theta}$ is one to one and continuous.

Proof: Suppose that J is a simple closed curve so there is a parameterization $\mathbf{r}$ and an interval $[a,b]$ such that $\mathbf{r}$ is continuous and one to one on $[a,b)$ and $(a,b]$ with $\mathbf{r}(a) = \mathbf{r}(b)$. Let $C_0 = \mathbf{r}((a,b)), C_\delta = \mathbf{r}([a+\delta, b-\delta])$, and let S^1 denote the unit circle. Let l be a linear one to one map from $[a,b]$ onto $[0,2\pi]$. Consider the following diagram.

$$\begin{array}{ccc} [a,b] & \stackrel{l}{\to} & [0,2\pi] \\ \downarrow \mathbf{r} & & \downarrow \mathbf{R} \\ C & & S^1 \end{array}$$

where $\mathbf{R}(\theta) \equiv (\cos\theta, \sin\theta)$. Then clearly $\mathbf{R}$ is continuous. It is also the case that, from the above lemma, $\mathbf{r}^{-1}$ is continuous on C_δ. Therefore, since $\delta > 0$ is arbitrary, $\boldsymbol{\theta} \equiv \mathbf{R} \circ l \circ \mathbf{r}^{-1}$ is a one to one and onto mapping from C_0 to $S^1 \setminus (1,0)$. Also, letting $\mathbf{p} = \mathbf{r}(a) = \mathbf{r}(b)$, it follows that $\boldsymbol{\theta}(\mathbf{p}) = (1,0)$. It remains to verify that $\boldsymbol{\theta}$ is continuous at $\mathbf{p}$. Suppose then that $\mathbf{r}(x_n) \to \mathbf{p} = \mathbf{r}(a) = \mathbf{r}(b)$. If $\boldsymbol{\theta}(\mathbf{r}(x_n))$ fails to converge to $(1,0) = \boldsymbol{\theta}(\mathbf{p})$, then there is a subsequence, still denoted as x_n and $\varepsilon > 0$ such that $|\boldsymbol{\theta}(\mathbf{r}(x_n)) - \boldsymbol{\theta}(\mathbf{p})| \geq \varepsilon$. In particular $x_n \notin \{a, b\}$. By the above lemma, $\mathbf{r}^{-1}$ is continuous on $\mathbf{r}([a,b))$ since this is true for $\mathbf{r}([a, b-\eta])$ for each $\eta > 0$. Since $\mathbf{p} = \mathbf{r}(a)$, it follows that

$$|\boldsymbol{\theta}(\mathbf{r}(x_n)) - \boldsymbol{\theta}(\mathbf{r}(a))| = |\mathbf{R} \circ l(x_n) - \mathbf{R} \circ l(a)| \geq \varepsilon$$

Hence there is some $\delta > 0$ such that $|x_n - a| \geq \delta_1$. Similarly, $|x_n - b| \geq \delta_2 > 0$. Letting $\delta = \min(\delta_1, \delta_2)$, it follows that $x_n \in [a+\delta, b-\delta]$. Taking a convergent subsequence, still denoted as $\{x_n\}$, there exists $x \in [a+\delta, b-\delta]$ such that $x_n \to x$. However, this implies that $\mathbf{r}(x_n) \to \mathbf{r}(x)$ and so $\mathbf{r}(x) = \mathbf{r}(a) = \mathbf{p}$, a contradiction to the fact that $\mathbf{r}$ is one to one on $[a,b)$.

Next suppose J is the image of the unit circle as just explained. Then let $\mathbf{R} : [0, 2\pi] \to S^1$ be defined as $\mathbf{R}(t) \equiv (\cos(t), \sin(t))$. Then consider $\mathbf{r}(t) \equiv \boldsymbol{\theta}(\mathbf{R}(t))$. $\mathbf{r}$ is one to one on $[0, 2\pi)$ and $(0, 2\pi]$ with $\mathbf{r}(0) = \mathbf{r}(2\pi)$ and is continuous, being the composition of continuous functions. ■

Before the proof of the Jordan curve theorem, recall Theorem 5.3.14 which says that the connected components of an open sets are open and that an open connected set is arcwise connected. If J is a Jordan curve then it is the continuous image of the compact set S^1 and so J is also compact. Therefore, its complement is open and the connected components of J^C are connected. The following lemma is a fairly obvious conclusion of this. A square curve is a continuous curve which consists entirely of line segments which are either horizontal or vertical.

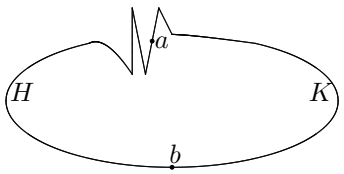
Lemma 13.0.27 *Let U be a connected open set and let x, y be points of U. Then there is a square curve which joins x and y.*

Proof: Let V denote those points of U which can be joined to x by a square curve. Then if $z \in V$, there exists $B(z,r) \subseteq U$. It is clear that every point of $B(z,r)$ can be joined to z by a square curve. Also V^C must be open since if $z \in V^C, B(z,r) \subseteq U$ for some r. Then if any $w \in B(z,r)$ is in V, one could join w to z by a square curve and conclude that $z \in V$ after all. The fact that both V, V^C are both open would result in a contradiction unless both $x, y \in V$ since otherwise, U is separated by V, V^C. ■

Theorem 13.0.28 *Let J be a Jordan curve in the plane. Then J^C consists of exactly two components, a bounded component, and an unbounded component, and J is the frontier of both of these components. Furthermore, J has empty interior.*

Proof: To begin with consider the claim there are no more than two components. Suppose this is not so. Then there exist x, y, z each of which is in a different component of J^C. Let $J = H \cup K$ where H and K are two Jordan arcs joined at the points a and b. If the Jordan curve is $\mathbf{r}([c,d])$ where $\mathbf{r}(c) = \mathbf{r}(d)$ as described above, you could take $H = \mathbf{r}\left(\left[c, \frac{c+d}{2}\right]\right)$ and $K = \mathbf{r}\left(\left[\frac{c+d}{2}, d\right]\right)$. Thus the points on the Jordan curve illustrated in the following picture could be

$$a = \mathbf{r}(c), b = \mathbf{r}\left(\frac{c+d}{2}\right)$$



First we show that there is at most two components in J^C. Suppose to the contrary that there exists x, y, z, each in a different component. By the Jordan arc theorem above, and the above lemma about square curves, there exists a square curve C_{xyH} such that $\partial C_{xyH} = x + y$ and $|C_{x,yH}| \cap H = \emptyset$. Using the same notation in relation to the other points, there exist square curves in the following list.

$$\begin{array}{rcl} C_{xyH},\ \partial C_{xyH} & = & x+y,\ C_{yzH},\ \partial C_{yzH} = y+z \\ C_{xyK},\ \partial C_{xyH} & = & x+y,\ C_{yzK},\ \partial C_{yzK} = y+z \end{array}$$

Let these square curves be part of a grating which includes all vertices of all these square curves and contains the compact set J in the bounded two cells. First note that $C_{xyH} + C_{xyK}$ is a one cycle and that

$$|C_{xyH} + C_{xyK}| \cap (H \cap K) = \emptyset$$

Also note that $H \cap K = \{a, b\}$ since $\mathbf{r}$ is one to one on $[c, d)$ and $(c, d]$. Therefore, there exist unique two chains D, E such that $\partial D = \partial E = C_{xyH} + C_{xyK}$. Now if one of these two chains contains both a, b then then the other two chain does not contain either a nor b. Then by Alexander's lemma, Lemma 13.0.19, there would exist a square curve C such that $|C| \cap (H \cup K) = |C| \cap J = \emptyset$ and $\partial C = x + y$ which is assumed not to happen. Therefore, one of the two chains contains a and the other contains b. Say $a \in |D|$ and $b \in |E|$. Similarly there exist unique two chains P, Q such that

$$\partial P = \partial Q = C_{yzH} + C_{yzK}$$

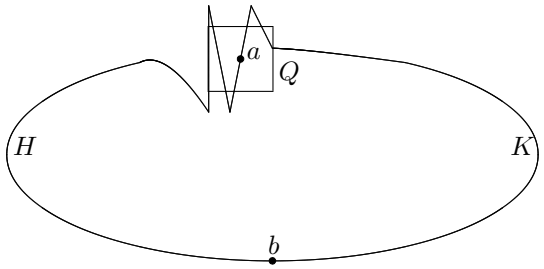
where $a \in |P|$ and $b \in |Q|$. Now consider

$$\partial (D + Q) = C_{xyH} + C_{xyK} + C_{yzH} + C_{yzK}$$

$$= (C_{xyH} + C_{yzH}) + (C_{xyK} + C_{yzK})$$

This is a one cycle because its boundary is $x + y + y + z + x + y + y + z = 0$. By Lemma 13.0.17, the fundamental lemma of the plane, there are exactly two two chains whose boundaries equal this one cycle. Therefore, $D + Q$ must be one of them. Also $b \in |Q|$ and is not in $|D|$. Hence $b \in |D + Q|$. Similarly $a \in |D + Q|$. It follows that the other two chain whose boundary equals the above one cycle contains neither a nor b. In addition to this, $C_{xyH} + C_{yzH}$ misses H and $C_{xyK} + C_{yzK}$ misses K. Both of these one chains have boundary equal to $x + z$. By Alexander's lemma, there exists a one chain C which misses both H and K (all of J) such that $\partial C = x + z$ which contradicts the assertion that x, z are in different components. This proves the assertion that there are only two components to J^C.

Next, why are there at least two components in J^C? Suppose there is only one and let a, b be the points of J described above and H, K also as above. Let Q be a small square 1 cycle which encloses a on its inside such that b is not inside Q. Thus a is on the inside of Q and b is on the outside of Q as shown in the picture.



Now let G be a grating which has the corners of Q as points of intersection of horizontal and vertical lines and also has all the 2 cells so small that none of them can intersect both of the disjoint compact sets $H \cap |Q|$ and $|Q| \cap K$. Let P be the 1 cells contained in Q which have nonempty intersection with H. Some of them must have nonempty intersection with H because if not, then H would fail to be connected, having points inside Q, a, and points outside Q, b, but no points on Q. Similarly some of these one cells in Q have nonempty intersection with K. Let $\partial P = x_1 + \cdots + x_m$. Then it follows each $x_k \notin H$. Could $\partial P = 0$? Suppose $\partial P = 0$. If l is a one cell of P, then since its ends are not in ∂P, the two adjacent one cells to l which are in Q must also intersect H. Moving counterclockwise around Q, it would follow that all the one cells contained in Q would intersect H. However, at least one must intersect K because if not, a is a point of K inside the square Q while b is a point of K outside Q thus separating K which is a connected set. However, this contradicts the choice of the grating. Therefore, $\partial P \neq 0$. Now this violates the assumption that no 2 cell of G can intersect both of those disjoint compact sets $H \cap |Q|$ and $|Q| \cap K$. Starting with a one cell of Q which does not intersect H, move counter clockwise till you obtain the first one which intersects H. This will produce a point of ∂P. Then the next point of ∂P will occur when the first one cell of P which does not intersect H is encountered. Thus a pair of points in ∂P are obtained. Now you are in the same position as before, continue moving counter clockwise and obtaining pairs of points of ∂P till there are no more one cells of Q which intersect H. You must have encountered an even number of points for ∂P.

Since it is assumed there is only one component of J^C, it follows upon refining G if necessary, there exist 1 chains B_k contained in J^C such that $\partial B_k = x_1 + x_k$ **and it is the existence of these B_k which will give the desired contradiction.** Let $B = B_2 + \cdots + B_m$. Then $P + B$ is a 1 cycle which misses K. It is a one cycle because m is even.

$$\partial (P + B) = \sum_{k=2}^{m=2l} x_1 + x_k + \sum_{k=1}^{2l} x_k = \sum_{k=1}^{2l} x_k + \sum_{k=2}^{2l} x_k + x_k = 0$$

It misses K because B misses J and all the 1 cells of P in the original grating G intersect $H \cap |Q|$ so they cannot intersect K. Also $P + Q + B$ is a 1 cycle which misses H. This is because B misses J and every 1 cell of P which intersects H disappears because $P + Q$ causes them to be added twice. Since H and K are connected, it follows from Lemma 13.0.20, 13.0.17 that $P + B$ bounds a 2 chain D which is contained entirely in K^C(the one which does not contain K). Similarly $P + Q + B$ bounds a 2 chain E which is contained in H^C(the 2 chain which does not contain H). Thus $D + E$ is a 2 chain which does not contain either a or b. (D misses K and E misses H and $\{a, b\} = H \cap K$) However,

$$\partial (D + E) = P + B + P + Q + B = Q$$

and so $D+E$ is one of the 2 chains of Lemma 13.0.17 which have Q as boundary. However, Q bounds the 2 chain of 2 cells which are inside Q which contains a and the 2 chain of 2 cells which are outside Q which contains b. This is a contradiction because neither of these 2 chains miss both a and b and this shows there are two components of J^C.

In the above argument, if each pair $\{x_1, x_i\}$ can be joined by a square curve B_i which lies in J^C, then the contradiction was obtained. Therefore, there must exist a pair $\{x_1, x_i\}$ which can't be joined by any square curve in J^C and this requires these points to be in different components by Lemma 13.0.27 above. Since they are both on Q and Q could be as small as desired, this shows a is in the frontier of both components of J^C. Furthermore, a was arbitrary so every point of J is a frontier point of both the components of J^C. These are the only frontier points because the components of J^C are open.

By Lemma 13.0.26, J is the continuous image of the compact set S^1 so it follows J is bounded. The unbounded component of J^C is the one which contains the connected set $B(0,R)^C$ where $J \subseteq B(0,R)$. Thus there are two components for J^C, the unbounded one which contains $B(0,R)^C$ and the bounded one which must be contained in $B(0,R)$. This proves the theorem. ■

Chapter 14

Line Integrals

14.1 Basic Properties

14.1.1 Length

I will give a discussion of what is meant by a line integral which is independent of the earlier material on Lebesgue integration. Line integrals are of fundamental importance in physics and in the theory of functions of a complex variable.

Definition 14.1.1 *Let $\boldsymbol{\gamma} : [a,b] \to \mathbb{R}^n$ be a function. Then $\boldsymbol{\gamma}$ is of bounded variation if*

$$\sup\left\{\sum_{i=1}^{n} |\boldsymbol{\gamma}(t_i) - \boldsymbol{\gamma}(t_{i-1})| : a = t_0 < \cdots < t_n = b\right\} \equiv V(\boldsymbol{\gamma}, [a,b]) < \infty$$

where the sums are taken over all possible lists, $\{a = t_0 < \cdots < t_n = b\}$. The set of points traced out will be denoted by $\boldsymbol{\gamma}^* \equiv \boldsymbol{\gamma}([a,b])$. The function $\boldsymbol{\gamma}$ is called a parameterization of $\boldsymbol{\gamma}^*$. The set of points $\boldsymbol{\gamma}^*$ is called a rectifiable curve. If a set of points $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}([a,b])$ where $\boldsymbol{\gamma}$ is continuous and $\boldsymbol{\gamma}$ is one to one on $[a,b)$ and also one to one on $(a,b]$, then $\boldsymbol{\gamma}^*$ is called a simple curve. A closed curve is one which has a parameterization $\boldsymbol{\gamma}$ defined on an interval $[a,b]$ such that $\boldsymbol{\gamma}(a) = \boldsymbol{\gamma}(b)$. It is a simple closed curve if there is a parameterization g such that $\boldsymbol{\gamma}$ is one to one on $[a,b)$ and one to one on $(a,b]$ with $\boldsymbol{\gamma}(a) = \boldsymbol{\gamma}(b)$.

The case of most interest is for simple curves. It turns out that in this case, the above concept of length is a property which $\boldsymbol{\gamma}^*$ possesses independent of the parameterization $\boldsymbol{\gamma}$ used to describe the set of points $\boldsymbol{\gamma}^*$. To show this, it is helpful to use the following lemma.

Lemma 14.1.2 *Let $\phi : [a,b] \to \mathbb{R}$ be a continuous function and suppose ϕ is $1-1$ on (a,b). Then ϕ is either strictly increasing or strictly decreasing on $[a,b]$. Furthermore, ϕ^{-1} is continuous.*

Proof: First it is shown that ϕ is either strictly increasing or strictly decreasing on (a,b).

If ϕ is not strictly decreasing on (a,b), then there exists $x_1 < y_1$, $x_1, y_1 \in (a,b)$ such that

$$(\phi(y_1) - \phi(x_1))(y_1 - x_1) > 0.$$

If for some other pair of points, $x_2 < y_2$ with $x_2, y_2 \in (a,b)$, the above inequality does not hold, then since ϕ is $1-1$,

$$(\phi(y_2) - \phi(x_2))(y_2 - x_2) < 0.$$

Let $x_t \equiv tx_1 + (1-t)x_2$ and $y_t \equiv ty_1 + (1-t)y_2$. Then $x_t < y_t$ for all $t \in [0,1]$ because

$$tx_1 \leq ty_1 \text{ and } (1-t)x_2 \leq (1-t)y_2$$

with strict inequality holding for at least one of these inequalities since not both t and $(1-t)$ can equal zero. Now define

$$h(t) \equiv (\phi(y_t) - \phi(x_t))(y_t - x_t).$$

Since h is continuous and $h(0) < 0$, while $h(1) > 0$, there exists $t \in (0,1)$ such that $h(t) = 0$. Therefore, both x_t and y_t are points of (a,b) and $\phi(y_t) - \phi(x_t) = 0$ contradicting the assumption that ϕ is one to one. It follows ϕ is either strictly increasing or strictly decreasing on (a,b).

This property of being either strictly increasing or strictly decreasing on (a,b) carries over to $[a,b]$ by the continuity of ϕ. Suppose ϕ is strictly increasing on (a,b), a similar argument holding for ϕ strictly decreasing on (a,b). If $x > a$, then pick $y \in (a,x)$ and from the above, $\phi(y) < \phi(x)$. Now by continuity of ϕ at a,

$$\phi(a) = \lim_{x \to a+} \phi(z) \leq \phi(y) < \phi(x).$$

Therefore, $\phi(a) < \phi(x)$ whenever $x \in (a,b)$. Similarly $\phi(b) > \phi(x)$ for all $x \in (a,b)$.

It only remains to verify ϕ^{-1} is continuous. Suppose then that $s_n \to s$ where s_n and s are points of $\phi([a,b])$. It is desired to verify that $\phi^{-1}(s_n) \to \phi^{-1}(s)$. If this does not happen, there exists $\varepsilon > 0$ and a subsequence, still denoted by s_n such that $\left|\phi^{-1}(s_n) - \phi^{-1}(s)\right| \geq \varepsilon$. Using the sequential compactness of $[a,b]$ there exists a further subsequence, still denoted by n, such that $\phi^{-1}(s_n) \to t_1 \in [a,b], t_1 \neq \phi^{-1}(s)$. Then by continuity of ϕ, it follows $s_n \to \phi(t_1)$ and so $s = \phi(t_1)$. Therefore, $t_1 = \phi^{-1}(s)$ after all. This proves the lemma. ■

Now suppose $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ are two parameterizations of the simple curve $\boldsymbol{\gamma}^*$ as described above. Thus $\boldsymbol{\gamma}([a,b]) = \boldsymbol{\gamma}^* = \boldsymbol{\eta}([c,d])$ and the two continuous functions $\boldsymbol{\gamma}, \boldsymbol{\eta}$ are one to one on their respective open intervals. I need to show the two definitions of length yield the same thing with either parameterization. Since $\boldsymbol{\gamma}^*$ is compact, it follows from Theorem 5.1.3 on Page 90, both $\boldsymbol{\gamma}^{-1}$ and $\boldsymbol{\eta}^{-1}$ are continuous. Thus $\boldsymbol{\gamma}^{-1} \circ \boldsymbol{\eta} : [c,d] \to [a,b]$ is continuous. It is also uniformly continuous because $[c,d]$ is compact. Let $\mathcal{P} \equiv \{t_0, \cdots, t_n\}$ be a partition of $[a,b], t_0 < t_1 < \cdots < t_n$ such that for $L < V(\boldsymbol{\gamma}, [a,b])$,

$$L < \sum_{k=1}^{n} |\boldsymbol{\gamma}(t_k) - \boldsymbol{\gamma}(t_{k-1})| \leq V(\boldsymbol{\gamma}, [a,b])$$

Note the sums approximating the total variation are all no larger than the total variation because when another point is added in to the partition, it is an easy exercise in the triangle inequality to show the corresponding sum either becomes larger or stays the same.

Let $\boldsymbol{\gamma}^{-1} \circ \boldsymbol{\eta}(s_k) = t_k$ so that $\{s_0, \cdots, s_n\}$ is a partition of $[c,d]$. By the lemma, the s_k are either strictly decreasing or strictly increasing as a function of k, depending on whether $\boldsymbol{\gamma}^{-1} \circ \boldsymbol{\eta}$ is increasing or decreasing. Thus $\boldsymbol{\gamma}(t_k) = \boldsymbol{\eta}(s_k)$ and so

$$L < \sum_{k=1}^{n} |\boldsymbol{\eta}(s_k) - \boldsymbol{\eta}(s_{k-1})| \leq V(\boldsymbol{\gamma}, [a,b])$$

It follows that whenever $L < V(\boldsymbol{\gamma}, [a,b])$, there exists a partition of $[c,d]$, $\{s_0, \cdots, s_n\}$ such that

$$L < \sum_{k=1}^{n} |\boldsymbol{\eta}(s_k) - \boldsymbol{\eta}(s_{k-1})|$$

It follows that for every $L < V(\boldsymbol{\gamma}, [a,b])$, $V(\boldsymbol{\eta}, [c,d]) \geq L$ which requires $V(\boldsymbol{\eta}, [c,d]) \geq V(\boldsymbol{\gamma}, [a,b])$. Turning the argument around, it follows

$$V(\boldsymbol{\eta}, [c,d]) = V(\boldsymbol{\gamma}, [a,b]).$$

This proves the following fundamental theorem.

Theorem 14.1.3 *Let Γ be a simple curve and let $\boldsymbol{\gamma}$ be a parameterization for Γ where $\boldsymbol{\gamma}$ is one to one on (a,b), continuous on $[a,b]$ and of bounded variation. Then the total variation*

$$V(\boldsymbol{\gamma}, [a,b])$$

can be used as a definition for the length of Γ in the sense that if $\Gamma = \boldsymbol{\eta}([c,d])$ where $\boldsymbol{\eta}$ is a continuous function which is one to one on (c,d) with $\boldsymbol{\eta}([c,d]) = \Gamma$,

$$V(\boldsymbol{\gamma}, [a,b]) = V(\boldsymbol{\eta}, [c,d]).$$

This common value can be denoted by $V(\Gamma)$ and is called the length of Γ.

The length is not dependent on parameterization. Simple curves which have such parameterizations are called rectifiable.

14.1.2 Orientation

There is another notion called orientation. For simple rectifiable curves, you can think of it as a direction of motion over the curve but what does this really mean for a wriggly curve? A precise description is needed.

Definition 14.1.4 *Let $\boldsymbol{\eta}, \boldsymbol{\gamma}$ be continuous one to one parameterizations for a simple rectifiable curve. If $\boldsymbol{\eta}^{-1} \circ \boldsymbol{\gamma}$ is increasing, then $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ are said to be equivalent parameterizations and this is written as $\boldsymbol{\gamma} \sim \boldsymbol{\eta}$. It is also said that the two parameterizations give the same orientation for the curve when $\boldsymbol{\gamma} \sim \boldsymbol{\eta}$.*

When the parameterizations are equivalent, they preserve the direction of motion along the curve and this also shows there are exactly two orientations of the curve since either $\boldsymbol{\eta}^{-1} \circ \boldsymbol{\gamma}$ is increasing or it is decreasing thanks to Lemma 14.1.2. In simple language, the message is that there are exactly two directions of motion along a simple curve.

Lemma 14.1.5 *The following hold for $\sim$.*

$$\boldsymbol{\gamma} \sim \boldsymbol{\gamma}, \tag{14.1}$$

$$\text{If } \boldsymbol{\gamma} \sim \boldsymbol{\eta} \text{ then } \boldsymbol{\eta} \sim \boldsymbol{\gamma}, \tag{14.2}$$

$$\text{If } \boldsymbol{\gamma} \sim \boldsymbol{\eta} \text{ and } \boldsymbol{\eta} \sim \boldsymbol{\theta}, \text{ then } \boldsymbol{\gamma} \sim \boldsymbol{\theta}. \tag{14.3}$$

Proof: Formula 14.1 is obvious because $\boldsymbol{\gamma}^{-1} \circ \boldsymbol{\gamma}(t) = t$ so it is clearly an increasing function. If $\boldsymbol{\gamma} \sim \boldsymbol{\eta}$ then $\boldsymbol{\gamma}^{-1} \circ \boldsymbol{\eta}$ is increasing. Now $\boldsymbol{\eta}^{-1} \circ \boldsymbol{\gamma}$ must also be increasing because it is the inverse of $\boldsymbol{\gamma}^{-1} \circ \boldsymbol{\eta}$. This verifies 14.2. To see 14.3, $\boldsymbol{\gamma}^{-1} \circ \boldsymbol{\theta} = \left(\boldsymbol{\gamma}^{-1} \circ \boldsymbol{\eta}\right) \circ \left(\boldsymbol{\eta}^{-1} \circ \boldsymbol{\theta}\right)$ and so since both of these functions are increasing, it follows $\boldsymbol{\gamma}^{-1} \circ \boldsymbol{\theta}$ is also increasing. This proves the lemma. ■

Definition 14.1.6 *Let Γ be a simple rectifiable curve and let $\boldsymbol{\gamma}$ be a parameterization for Γ. Denoting by $[\boldsymbol{\gamma}]$ the equivalence class of parameterizations determined by the above equivalence relation, the following pair will be called an oriented curve.*

$$(\Gamma, [\boldsymbol{\gamma}])$$

In simple language, an oriented curve is one which has a direction of motion specified.

Actually, people usually just write Γ and there is understood a direction of motion or orientation on Γ. How can you identify which orientation is being considered?

Proposition 14.1.7 *Let $(\Gamma, [\boldsymbol{\gamma}])$ be an oriented simple curve and let $\mathbf{p}, \mathbf{q}$ be any two distinct points of Γ. Then $[\boldsymbol{\gamma}]$ is determined by the order of $\boldsymbol{\gamma}^{-1}(\mathbf{p})$ and $\boldsymbol{\gamma}^{-1}(\mathbf{q})$. This means that $\boldsymbol{\eta} \in [\boldsymbol{\gamma}]$ if and only if $\boldsymbol{\eta}^{-1}(\mathbf{p})$ and $\boldsymbol{\eta}^{-1}(\mathbf{q})$ occur in the same order as $\boldsymbol{\gamma}^{-1}(\mathbf{p})$ and $\boldsymbol{\gamma}^{-1}(\mathbf{q})$.*

Proof: Suppose $\boldsymbol{\gamma}^{-1}(\mathbf{p}) < \boldsymbol{\gamma}^{-1}(\mathbf{q})$ and let $\boldsymbol{\eta} \in [\boldsymbol{\gamma}]$. Is it true that $\boldsymbol{\eta}^{-1}(\mathbf{p}) < \boldsymbol{\eta}^{-1}(\mathbf{q})$? Of course it is because $\boldsymbol{\gamma}^{-1} \circ \boldsymbol{\eta}$ is increasing. Therefore, if $\boldsymbol{\eta}^{-1}(\mathbf{p}) > \boldsymbol{\eta}^{-1}(\mathbf{q})$ it would follow

$$\boldsymbol{\gamma}^{-1}(\mathbf{p}) = \boldsymbol{\gamma}^{-1} \circ \boldsymbol{\eta}\left(\boldsymbol{\eta}^{-1}(\mathbf{p})\right) > \boldsymbol{\gamma}^{-1} \circ \boldsymbol{\eta}\left(\boldsymbol{\eta}^{-1}(\mathbf{q})\right) = \boldsymbol{\gamma}^{-1}(\mathbf{q})$$

which is a contradiction. Thus if $\boldsymbol{\gamma}^{-1}(\mathbf{p}) < \boldsymbol{\gamma}^{-1}(\mathbf{q})$ for one $\boldsymbol{\gamma} \in [\boldsymbol{\gamma}]$, then this is true for all $\boldsymbol{\eta} \in [\boldsymbol{\gamma}]$.

Now suppose $\boldsymbol{\eta}$ is a parameterization for Γ defined on $[c, d]$ which has the property that

$$\boldsymbol{\eta}^{-1}(\mathbf{p}) < \boldsymbol{\eta}^{-1}(\mathbf{q})$$

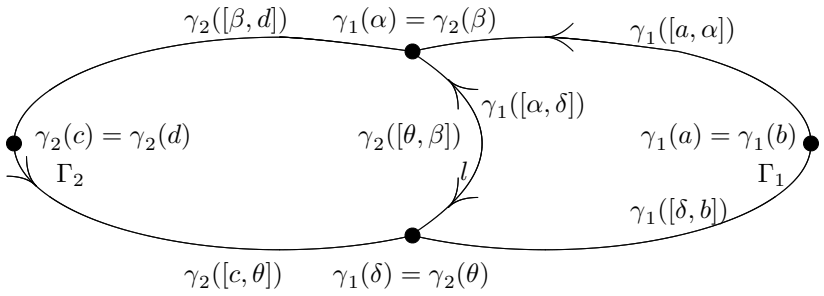
Does it follow $\boldsymbol{\eta} \in [\boldsymbol{\gamma}]$? Is $\boldsymbol{\gamma}^{-1} \circ \boldsymbol{\eta}$ increasing? By Lemma 14.1.2 it is either increasing or decreasing. Thus it suffices to test it on two points of $[c, d]$. Pick the two points $\boldsymbol{\eta}^{-1}(\mathbf{p}), \boldsymbol{\eta}^{-1}(\mathbf{q})$. Is

$$\boldsymbol{\gamma}^{-1} \circ \boldsymbol{\eta}\left(\boldsymbol{\eta}^{-1}(\mathbf{p})\right) < \boldsymbol{\gamma}^{-1} \circ \boldsymbol{\eta}\left(\boldsymbol{\eta}^{-1}(\mathbf{q})\right)?$$

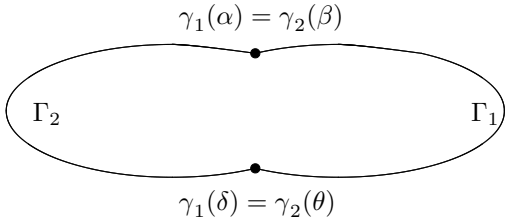
Yes because these reduce to $\boldsymbol{\gamma}^{-1}(\mathbf{p})$ on the left and $\boldsymbol{\gamma}^{-1}(\mathbf{q})$ on the right. It is given that $\boldsymbol{\gamma}^{-1}(\mathbf{p}) < \boldsymbol{\gamma}^{-1}(\mathbf{q})$. This proves the lemma. ■

This shows that the direction of motion on the curve is determined by any two points and the determination of which is encountered first by any parameterization in the equivalence class of parameterizations which determines the orientation. Sometimes people indicate this direction of motion by drawing an arrow.

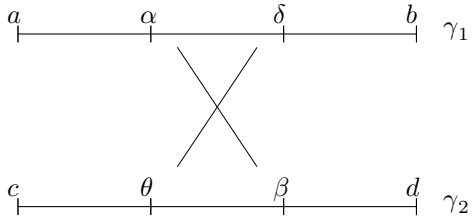
Now here is an interesting observation relative to two simple closed rectifiable curves. The situation is illustrated by the following picture.



Proposition 14.1.8 *Let Γ_1 and Γ_2 be two simple closed rectifiable oriented curves and let their intersection be l. Suppose also that l is itself a simple curve. Also suppose the orientation of l when considered a part of Γ_1 is opposite its orientation when considered a part of Γ_2. Then if the open segment (l except for its endpoints) of l is removed, the result is a simple closed rectifiable curve Γ. This curve has a parameterization $\boldsymbol{\gamma}$ with the property that on $\boldsymbol{\gamma}_j^{-1}(\Gamma \cap \Gamma_j)$, $\boldsymbol{\gamma}^{-1}\boldsymbol{\gamma}_j$ is increasing. In other words, Γ has an orientation consistent with that of Γ_1 and Γ_2. Furthermore, if Γ has such a consistent orientation, then the orientations of l as part of the two simple closed curves, Γ_1 and Γ_2 are opposite.*

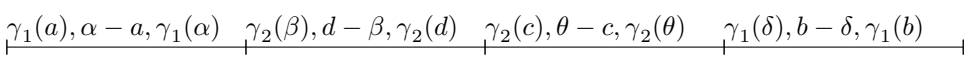


Proof: Let $\Gamma_1 = \boldsymbol{\gamma}_1([a,b]), \boldsymbol{\gamma}_1(a) = \boldsymbol{\gamma}_1(b)$, and $\Gamma_2 = \boldsymbol{\gamma}_2([c,d]), \boldsymbol{\gamma}_2(c) = \boldsymbol{\gamma}_2(d)$, with $l = \boldsymbol{\gamma}_1([\alpha,\delta]) = \boldsymbol{\gamma}_2([\theta,\beta])$. (Recall continuous images of connected sets are connected and the connected sets on the real line are intervals.) By the assumption the two orientations are opposite, something can be said about the relationship of $\alpha, \delta, \theta, \beta$. Suppose without loss of generality that $\alpha < \delta$. Then because of this assumption it follows $\gamma_2(\theta) = \gamma_1(\delta), \gamma_2(\beta) = \gamma_1(\alpha)$. The following diagram might be useful to summarize what was just said.



Note the first of the interval $[\beta, d]$ matches the last of the interval $[a, \alpha]$ and the first of $[\delta, b]$ matches the last of $[c, \theta]$, all this in terms of where these points are sent.

Now I need to describe the parameterization of $\Gamma \equiv \Gamma_1 \cup \Gamma_2$. To verify it is a simple closed curve, I must produce an interval and a mapping from this interval to Γ which satisfies the conditions needed for $\boldsymbol{\gamma}$ to be a simple closed rectifiable curve. The following is the definition as well as a description of which part of Γ_j is being obtained. It is helpful to look at the above picture and the following picture in which there are intervals placed next to each other. Above each is where the left end point starts off followed by its length and finally where it ends up.



Note it ends up where it started, at $\gamma_1(a) = \gamma_1(b)$. The following involved description is nothing but the above picture with the edges of the little intervals computed along with a description of γ which corresponds to the above picture.

Then $\boldsymbol{\gamma}(t)$ is given by

$$\boldsymbol{\gamma}(t) \equiv$$

$$\begin{cases} \boldsymbol{\gamma}_1(t), t \in [a,\alpha], \boldsymbol{\gamma}_1(a) \to \boldsymbol{\gamma}_1(\alpha) = \boldsymbol{\gamma}_2(\beta) \\ \boldsymbol{\gamma}_2(t+\beta-\alpha), t \in [\alpha, \alpha+d-\beta], \boldsymbol{\gamma}_2(\beta) \to \boldsymbol{\gamma}_2(d) = \boldsymbol{\gamma}_2(c) \\ \boldsymbol{\gamma}_2(t+c-\alpha-d+\beta), t \in [\alpha+d-\beta, \alpha+d-\beta+\theta-c], \\ \boldsymbol{\gamma}_2(c) = \boldsymbol{\gamma}_2(d) \to \boldsymbol{\gamma}_2(\theta) = \boldsymbol{\gamma}_1(\delta) \\ \boldsymbol{\gamma}_1(t-\alpha-d+\beta-\theta+c+\delta), t \in [\alpha+d-\beta+\theta-c, \alpha+d-\beta+\theta-c+b-\delta], \\ \boldsymbol{\gamma}_1(\delta) \to \boldsymbol{\gamma}_1(b) = \boldsymbol{\gamma}_1(a) \end{cases}$$

The construction shows $\boldsymbol{\gamma}$ is one to one on

$$(a, \alpha+d-\beta+\theta-c+b-\delta)$$

and if t is in this open interval, then

$$\boldsymbol{\gamma}(t) \neq \boldsymbol{\gamma}(a) = \boldsymbol{\gamma}_1(a)$$

and

$$\boldsymbol{\gamma}(t) \neq \boldsymbol{\gamma}(\alpha+d-\beta+\theta-c+b-\delta) = \boldsymbol{\gamma}_1(b).$$

Also

$$\boldsymbol{\gamma}(a) = \boldsymbol{\gamma}_1(a) = \boldsymbol{\gamma}(\alpha+d-\beta+\theta-c+b-\delta) = \boldsymbol{\gamma}_1(b)$$

so it is a simple closed curve. The claim about preserving the orientation is also obvious from the formula. Note that t is never subtracted.

It only remains to prove the last claim. Suppose then that it is not so and l has the same orientation as part of each Γ_j. Then from a repeat of the above argument, you could change the orientation of l relative to Γ_2 and obtain an orientation of Γ which is consistent with that of Γ_1 and Γ_2. Call a parameterization which has this new orientation $\boldsymbol{\gamma}_n$ while $\boldsymbol{\gamma}$ is the one which is assumed to exist. This new orientation of l changes the orientation of Γ_2 because there are two points in l. Therefore on $\boldsymbol{\gamma}_2^{-1}(\Gamma \cap \Gamma_2)$, $\boldsymbol{\gamma}_n^{-1}\boldsymbol{\gamma}_2$ is decreasing while $\boldsymbol{\gamma}^{-1}\boldsymbol{\gamma}_2$ is assumed to be increasing. Hence $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_n$ are not equivalent. However, the above construction would leave the orientation of both $\boldsymbol{\gamma}_1([a,\alpha])$ and $\boldsymbol{\gamma}_1([\delta,b])$ unchanged and at least one of these must have at least two points. Thus the orientation of Γ must be the same for $\boldsymbol{\gamma}_n$ as for $\boldsymbol{\gamma}$. That is, $\boldsymbol{\gamma} \sim \boldsymbol{\gamma}_n$. This is a contradiction. This proves the proposition. ■

There is a slightly different aspect of the above proposition which is interesting. It involves using the shared segment to orient the simple closed curve Γ.

Corollary 14.1.9 *Let the intersection of simple closed rectifiable curves, Γ_1 and Γ_2 consist of the simple curve l. Then place opposite orientations on l, and use these two different orientations to specify orientations of Γ_1 and Γ_2. Then letting Γ denote the simple closed curve which is obtained from deleting the open segment of l, there exists an orientation for Γ which is consistent with the orientations of Γ_1 and Γ_2 obtained from the given specification of opposite orientations on l.*

14.2 The Line Integral

Now I will return to considering the more general notion of bounded variation parameterizations without worrying about whether $\boldsymbol{\gamma}$ is one to one on the open interval. The line integral and its properties are presented next.

Definition 14.2.1 *Let $\boldsymbol{\gamma} : [a,b] \to \mathbb{R}^n$ be of bounded variation and let $\mathbf{f} : \boldsymbol{\gamma}^* \to \mathbb{R}^n$. Letting $\mathcal{P} \equiv \{t_0, \cdots, t_n\}$ where $a = t_0 < t_1 < \cdots < t_n = b$, define*

$$||\mathcal{P}|| \equiv \max\{|t_j - t_{j-1}| : j = 1, \cdots, n\}$$

and the Riemann Stieltjes sum by

$$S(\mathcal{P}) \equiv \sum_{j=1}^{n} \mathbf{f}(\boldsymbol{\gamma}(\tau_j)) \cdot (\boldsymbol{\gamma}(t_j) - \boldsymbol{\gamma}(t_{j-1}))$$

where $\tau_j \in [t_{j-1}, t_j]$. *(Note this notation is a little sloppy because it does not identify the specific point,* τ_j *used. It is understood that this point is arbitrary.) Define* $\int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma}$ *as the unique number which satisfies the following condition. For all* $\varepsilon > 0$ *there exists a* $\delta > 0$ *such that if* $||\mathcal{P}|| \leq \delta$, *then*

$$\left| \int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma} - S(\mathcal{P}) \right| < \varepsilon.$$

Sometimes this is written as

$$\int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma} \equiv \lim_{||\mathcal{P}|| \to 0} S(\mathcal{P}).$$

Then $\boldsymbol{\gamma}^*$ is a set of points in $\mathbb{R}^n$ and as t moves from a to b, $\boldsymbol{\gamma}(t)$ moves from $\boldsymbol{\gamma}(a)$ to $\boldsymbol{\gamma}(b)$. Thus $\boldsymbol{\gamma}^*$ has a first point and a last point. (In the case of a closed curve these are the same point.) If $\phi : [c, d] \to [a, b]$ is a continuous nondecreasing function, then $\boldsymbol{\gamma} \circ \phi : [c, d] \to \mathbb{R}^n$ is also of bounded variation and yields the same set of points in $\mathbb{R}^n$ with the same first and last points.

Theorem 14.2.2 *Let* ϕ *and* $\boldsymbol{\gamma}$ *be as just described. Then assuming that*

$$\int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma}$$

exists, so does

$$\int_{\boldsymbol{\gamma} \circ \phi} \mathbf{f} \cdot d(\boldsymbol{\gamma} \circ \phi)$$

and

$$\int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma} = \int_{\boldsymbol{\gamma} \circ \phi} \mathbf{f} \cdot d(\boldsymbol{\gamma} \circ \phi). \tag{14.4}$$

Proof: There exists $\delta > 0$ such that if $\mathcal{P}$ is a partition of $[a, b]$ such that $||\mathcal{P}|| < \delta$, then

$$\left| \int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma} - S(\mathcal{P}) \right| < \varepsilon.$$

By continuity of ϕ, there exists $\sigma > 0$ such that if $\mathcal{Q}$ is a partition of $[c, d]$ with $||\mathcal{Q}|| < \sigma, \mathcal{Q} = \{s_0, \cdots, s_n\}$, then $|\phi(s_j) - \phi(s_{j-1})| < \delta$. Thus letting $\mathcal{P}$ denote the points in $[a, b]$ given by $\phi(s_j)$ for $s_j \in \mathcal{Q}$, it follows that $||\mathcal{P}|| < \delta$ and so

$$\left| \int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma} - \sum_{j=1}^{n} \mathbf{f}(\boldsymbol{\gamma}(\phi(\tau_j))) \cdot (\boldsymbol{\gamma}(\phi(s_j)) - \boldsymbol{\gamma}(\phi(s_{j-1}))) \right| < \varepsilon$$

where $\tau_j \in [s_{j-1}, s_j]$. Therefore, from the definition 14.4 holds and

$$\int_{\boldsymbol{\gamma} \circ \phi} \mathbf{f} \cdot d(\boldsymbol{\gamma} \circ \phi)$$

exists. This proves the theorem. ■

This theorem shows that $\int_{\boldsymbol{\gamma}} \mathbf{f}\cdot d\boldsymbol{\gamma}$ is independent of the particular parameterization $\boldsymbol{\gamma}$ used in its computation to the extent that if ϕ is any nondecreasing continuous function from another interval, $[c,d]$, mapping to $[a,b]$, then the same value is obtained by replacing $\boldsymbol{\gamma}$ with $\boldsymbol{\gamma}\circ\phi$. In other words, this line integral depends only on $\boldsymbol{\gamma}^*$ and the order in which $\boldsymbol{\gamma}(t)$ encounters the points of $\boldsymbol{\gamma}^*$ as t moves from one end to the other of the interval. For the case of an oriented rectifiable curve Γ this shows the line integral is dependent only on the set of points and the orientation of Γ. ■

The fundamental result in this subject is the following theorem.

Theorem 14.2.3 *Let $\mathbf{f}:\boldsymbol{\gamma}^*\to\mathbb{R}^n$ be continuous and let $\boldsymbol{\gamma}:[a,b]\to\mathbb{R}^n$ be continuous and of bounded variation. Then $\int_{\boldsymbol{\gamma}}\mathbf{f}\cdot d\boldsymbol{\gamma}$ exists. Also letting $\delta_m>0$ be such that $|t-s|<\delta_m$ implies $|\mathbf{f}(\boldsymbol{\gamma}(t))-\mathbf{f}(\boldsymbol{\gamma}(s))|<\frac{1}{m}$,*

$$\left|\int_{\boldsymbol{\gamma}}\mathbf{f}d\boldsymbol{\gamma}-S(\mathcal{P})\right|\leq\frac{2V(\boldsymbol{\gamma},[a,b])}{m}$$

whenever $||\mathcal{P}||<\delta_m$.

Proof: The function, $\mathbf{f}\circ\boldsymbol{\gamma}$, is uniformly continuous because it is defined on a compact set. Therefore, there exists a decreasing sequence of positive numbers, $\{\delta_m\}$ such that if $|s-t|<\delta_m$, then

$$|\mathbf{f}(\boldsymbol{\gamma}(t))-\mathbf{f}(\boldsymbol{\gamma}(s))|<\frac{1}{m}.$$

Let

$$F_m\equiv\overline{\{S(\mathcal{P}):||\mathcal{P}||<\delta_m\}}.$$

Thus F_m is a closed set. (The symbol, $S(\mathcal{P})$ in the above definition, means to include all sums corresponding to $\mathcal{P}$ for any choice of τ_j.) It is shown that

$$\operatorname{diam}(F_m)\leq\frac{2V(\boldsymbol{\gamma},[a,b])}{m}\tag{14.5}$$

and then it will follow there exists a unique point, $I\in\cap_{m=1}^{\infty}F_m$. This is because $\mathbb{R}$ is complete. It will then follow $I=\int_{\boldsymbol{\gamma}}\mathbf{f}(t)\,d\boldsymbol{\gamma}(t)$. To verify 14.5, it suffices to verify that whenever $\mathcal{P}$ and $\mathcal{Q}$ are partitions satisfying $||\mathcal{P}||<\delta_m$ and $||\mathcal{Q}||<\delta_m$,

$$|S(\mathcal{P})-S(\mathcal{Q})|\leq\frac{2}{m}V(\boldsymbol{\gamma},[a,b]).\tag{14.6}$$

Suppose $||\mathcal{P}||<\delta_m$ and $\mathcal{Q}\supseteq\mathcal{P}$. Then also $||\mathcal{Q}||<\delta_m$. To begin with, suppose that $\mathcal{P}\equiv\{t_0,\cdots,t_p,\cdots,t_n\}$ and $\mathcal{Q}\equiv\{t_0,\cdots,t_{p-1},t^*,t_p,\cdots,t_n\}$. Thus $\mathcal{Q}$ contains only one more point than $\mathcal{P}$. Letting $S(\mathcal{Q})$ and $S(\mathcal{P})$ be Riemann Stieltjes sums,

$$S(\mathcal{Q})\equiv\sum_{j=1}^{p-1}\mathbf{f}(\boldsymbol{\gamma}(\sigma_j))\cdot(\boldsymbol{\gamma}(t_j)-\boldsymbol{\gamma}(t_{j-1}))+\mathbf{f}(\boldsymbol{\gamma}(\sigma_*))(\boldsymbol{\gamma}(t^*)-\boldsymbol{\gamma}(t_{p-1}))$$

$$+\mathbf{f}(\boldsymbol{\gamma}(\sigma^*))\cdot(\boldsymbol{\gamma}(t_p)-\boldsymbol{\gamma}(t^*))+\sum_{j=p+1}^{n}\mathbf{f}(\boldsymbol{\gamma}(\sigma_j))\cdot(\boldsymbol{\gamma}(t_j)-\boldsymbol{\gamma}(t_{j-1})),$$

$$S(\mathcal{P})\equiv\sum_{j=1}^{p-1}\mathbf{f}(\boldsymbol{\gamma}(\tau_j))\cdot(\boldsymbol{\gamma}(t_j)-\boldsymbol{\gamma}(t_{j-1}))+$$

$$\overbrace{\mathbf{f}(\boldsymbol{\gamma}(\tau_p))\cdot(\boldsymbol{\gamma}(t^*)-\boldsymbol{\gamma}(t_{p-1}))+\mathbf{f}(\boldsymbol{\gamma}(\tau_p))\cdot(\boldsymbol{\gamma}(t_p)-\boldsymbol{\gamma}(t^*))}^{=\mathbf{f}(\boldsymbol{\gamma}(\tau_p))\cdot(\boldsymbol{\gamma}(t_p)-\boldsymbol{\gamma}(t_{p-1}))}$$

$$+\sum_{j=p+1}^{n}\mathbf{f}\left(\boldsymbol{\gamma}\left(\tau_j\right)\right)\cdot\left(\boldsymbol{\gamma}\left(t_j\right)-\boldsymbol{\gamma}\left(t_{j-1}\right)\right).$$

Therefore,

$$\left|S\left(\mathcal{P}\right)-S\left(\mathcal{Q}\right)\right|\leq\sum_{j=1}^{p-1}\frac{1}{m}\left|\boldsymbol{\gamma}\left(t_j\right)-\boldsymbol{\gamma}\left(t_{j-1}\right)\right|+\frac{1}{m}\left|\boldsymbol{\gamma}\left(t^*\right)-\boldsymbol{\gamma}\left(t_{p-1}\right)\right|+$$

$$\frac{1}{m}\left|\boldsymbol{\gamma}\left(t_p\right)-\boldsymbol{\gamma}\left(t^*\right)\right|+\sum_{j=p+1}^{n}\frac{1}{m}\left|\boldsymbol{\gamma}\left(t_j\right)-\boldsymbol{\gamma}\left(t_{j-1}\right)\right|\leq\frac{1}{m}V\left(\boldsymbol{\gamma},\left[a,b\right]\right). \tag{14.7}$$

Clearly the extreme inequalities would be valid in 14.7 if $\mathcal{Q}$ had more than one extra point. You simply do the above trick more than one time. Let $S\left(\mathcal{P}\right)$ and $S\left(\mathcal{Q}\right)$ be Riemann Stieltjes sums for which $||\mathcal{P}||$ and $||\mathcal{Q}||$ are less than δ_m and let $\mathcal{R}\equiv\mathcal{P}\cup\mathcal{Q}$. Then from what was just observed,

$$\left|S\left(\mathcal{P}\right)-S\left(\mathcal{Q}\right)\right|\leq\left|S\left(\mathcal{P}\right)-S\left(\mathcal{R}\right)\right|+\left|S\left(\mathcal{R}\right)-S\left(\mathcal{Q}\right)\right|\leq\frac{2}{m}V\left(\boldsymbol{\gamma},\left[a,b\right]\right).$$

and this shows 14.6 which proves 14.5. Therefore, there exists a unique number, $I\in\cap_{m=1}^{\infty}F_m$ which satisfies the definition of $\int_{\boldsymbol{\gamma}}\mathbf{f}\cdot d\boldsymbol{\gamma}$. This proves the theorem. ■

Note this is a general sort of result. It is not assumed that $\boldsymbol{\gamma}$ is one to one anywhere in the proof. The following theorem follows easily from the above definitions and theorem. This theorem is used to establish estimates.

Theorem 14.2.4 *Let* $\mathbf{f}$ *be a continuous function defined on* $\boldsymbol{\gamma}^*$*, denoted as* $f\in C\left(\boldsymbol{\gamma}^*\right)$ *where* $\boldsymbol{\gamma}:\left[a,b\right]\to\mathbb{R}^n$ *is of bounded variation and continuous. Let*

$$M\geq\max\left\{\left|\mathbf{f}\circ\boldsymbol{\gamma}\left(t\right)\right|:t\in\left[a,b\right]\right\}. \tag{14.8}$$

Then

$$\left|\int_{\boldsymbol{\gamma}}\mathbf{f}\cdot d\boldsymbol{\gamma}\right|\leq MV\left(\boldsymbol{\gamma},\left[a,b\right]\right). \tag{14.9}$$

Also if $\left\{\mathbf{f}_m\right\}$ *is a sequence of functions of* $C\left(\boldsymbol{\gamma}^*\right)$ *which is converging uniformly to the function,* $\mathbf{f}$ *on* $\boldsymbol{\gamma}^*$*, then*

$$\lim_{m\to\infty}\int_{\boldsymbol{\gamma}}\mathbf{f}_m\cdot d\boldsymbol{\gamma}=\int_{\boldsymbol{\gamma}}\mathbf{f}\cdot d\boldsymbol{\gamma}. \tag{14.10}$$

In case $\boldsymbol{\gamma}\left(a\right)=\boldsymbol{\gamma}\left(b\right)$ *so the curve is a closed curve and for* f_k *the* k^{th} *component of* $\mathbf{f}$*,*

$$m_k\leq f_k\left(\mathbf{x}\right)\leq M_k$$

for all $\mathbf{x}\in\boldsymbol{\gamma}^*$*, it also follows*

$$\left|\int_{\boldsymbol{\gamma}}\mathbf{f}\cdot d\boldsymbol{\gamma}\right|\leq\frac{1}{2}\left(\sum_{k=1}^{n}\left(M_k-m_k\right)^2\right)^{1/2}V\left(\boldsymbol{\gamma},\left[a,b\right]\right) \tag{14.11}$$

Proof: Let 14.8 hold. From the proof of Theorem 14.2.3, when $||\mathcal{P}||<\delta_m$,

$$\left|\int_{\boldsymbol{\gamma}}\mathbf{f}\cdot d\boldsymbol{\gamma}-S\left(\mathcal{P}\right)\right|\leq\frac{2}{m}V\left(\boldsymbol{\gamma},\left[a,b\right]\right)$$

and so

$$\left|\int_{\boldsymbol{\gamma}}\mathbf{f}\cdot d\boldsymbol{\gamma}\right|\leq\left|S\left(\mathcal{P}\right)\right|+\frac{2}{m}V\left(\boldsymbol{\gamma},\left[a,b\right]\right)$$

Using the Cauchy Schwarz inequality and the above estimate in $S(\mathcal{P})$,

$$\begin{aligned} &\leq \sum_{j=1}^{n} M\left|\boldsymbol{\gamma}(t_j)-\boldsymbol{\gamma}(t_{j-1})\right| + \frac{2}{m}V(\boldsymbol{\gamma},[a,b]) \\ &\leq MV(\boldsymbol{\gamma},[a,b]) + \frac{2}{m}V(\boldsymbol{\gamma},[a,b]). \end{aligned}$$

This proves 14.9 since m is arbitrary.

To verify 14.10 use the above inequality to write

$$\left|\int_{\boldsymbol{\gamma}} \mathbf{f}\cdot d\boldsymbol{\gamma} - \int_{\boldsymbol{\gamma}} \mathbf{f}_m \cdot d\boldsymbol{\gamma}\right| = \left|\int_{\boldsymbol{\gamma}} (\mathbf{f}-\mathbf{f}_m)\cdot d\boldsymbol{\gamma}(t)\right|$$

$$\leq \max\left\{\left|\mathbf{f}\circ\boldsymbol{\gamma}(t)-\mathbf{f}_m\circ\boldsymbol{\gamma}(t)\right| : t\in[a,b]\right\} V(\boldsymbol{\gamma},[a,b]).$$

Since the convergence is assumed to be uniform, this proves 14.10.

Claim: Let γ be closed bounded variation curve. Then if $\mathbf{c}$ is a constant vector,

$$\int_{\gamma} \mathbf{c}\cdot d\boldsymbol{\gamma} = 0$$

Proof of the claim: Let $\mathcal{P}\equiv\{t_0,\cdots,t_p\}$ be a partition with the property that

$$\left|\int_{\gamma} \mathbf{c}\cdot d\boldsymbol{\gamma} - \sum_{k=1}^{p} \mathbf{c}\cdot\left(\boldsymbol{\gamma}(t_k)-\boldsymbol{\gamma}(t_{k-1})\right)\right| < \varepsilon.$$

Consider the sum. It is of the form

$$\begin{aligned} \sum_{k=1}^{p} \mathbf{c}\cdot\boldsymbol{\gamma}(t_k) - \sum_{k=1}^{p} \mathbf{c}\cdot\boldsymbol{\gamma}(t_{k-1}) &= \sum_{k=1}^{p} \mathbf{c}\cdot\boldsymbol{\gamma}(t_k) - \sum_{k=0}^{p-1} \mathbf{c}\cdot\boldsymbol{\gamma}(t_k) \\ &= \mathbf{c}\cdot\boldsymbol{\gamma}(t_p) - \mathbf{c}\cdot\boldsymbol{\gamma}(t_0) = 0 \end{aligned}$$

because it is given that since $\boldsymbol{\gamma}^*$ is a closed curve, $\boldsymbol{\gamma}(t_0)=\boldsymbol{\gamma}(t_p)$. This shows the claim.

It only remains to verify 14.11. In this case $\boldsymbol{\gamma}(a)=\boldsymbol{\gamma}(b)$ and so for each vector $\mathbf{c}$

$$\int_{\boldsymbol{\gamma}} \mathbf{f}\cdot d\boldsymbol{\gamma} = \int_{\boldsymbol{\gamma}} (\mathbf{f}-\mathbf{c})\cdot d\boldsymbol{\gamma}$$

for any constant vector $\mathbf{c}$. Let

$$c_k = \frac{1}{2}(M_k+m_k)$$

Then for $t\in[a,b]$

$$\begin{aligned} \left|\mathbf{f}(\boldsymbol{\gamma}(t))-\mathbf{c}\right|^2 &= \sum_{k=1}^{n}\left|f_k(\boldsymbol{\gamma}(t)) - \frac{1}{2}(M_k+m_k)\right|^2 \\ &\leq \sum_{k=1}^{n}\left(\frac{1}{2}(M_k-m_k)\right)^2 = \frac{1}{4}\sum_{k=1}^{n}(M_k-m_k)^2 \end{aligned}$$

Then with this choice of $\mathbf{c}$, it follows from 14.9 that

$$\left|\int_{\boldsymbol{\gamma}} \mathbf{f}\cdot d\boldsymbol{\gamma}\right| = \left|\int_{\boldsymbol{\gamma}} (\mathbf{f}-\mathbf{c})\cdot d\boldsymbol{\gamma}\right|$$

$$\leq \frac{1}{2}\left(\sum_{k=1}^{n}\left(M_k - m_k\right)^2\right)^{1/2} V\left(\boldsymbol{\gamma}, [a,b]\right)$$

This proves the lemma. ■

It turns out to be much easier to evaluate line integrals in the case where there exists a parameterization $\boldsymbol{\gamma}$ which is in $C^1\left([a,b]\right)$. The following theorem about approximation will be very useful but first here is an easy lemma.

Lemma 14.2.5 *Let $\boldsymbol{\gamma} : [a,b] \to \mathbb{R}^n$ be in $C^1\left([a,b]\right)$. Then $V\left(\boldsymbol{\gamma}, [a,b]\right) < \infty$ so $\boldsymbol{\gamma}$ is of bounded variation.*

Proof: This follows from the following

$$\begin{aligned}
\sum_{j=1}^{n}\left|\boldsymbol{\gamma}\left(t_j\right) - \boldsymbol{\gamma}\left(t_{j-1}\right)\right| &= \sum_{j=1}^{n}\left|\int_{t_{j-1}}^{t_j}\boldsymbol{\gamma}'\left(s\right)ds\right| \\
&\leq \sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\left|\boldsymbol{\gamma}'\left(s\right)\right|ds \\
&\leq \sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\left|\left|\boldsymbol{\gamma}'\right|\right|_\infty ds \\
&= \left|\left|\boldsymbol{\gamma}'\right|\right|_\infty\left(b-a\right).
\end{aligned}$$

where

$$\left|\left|\boldsymbol{\gamma}'\right|\right|_\infty \equiv \max\left\{\left|\boldsymbol{\gamma}'\left(t\right)\right| : t \in [a,b]\right\}$$

which exists because $\boldsymbol{\gamma}'$ is given to be continuous. Therefore it follows $V\left(\boldsymbol{\gamma}, [a,b]\right) \leq \left|\left|\boldsymbol{\gamma}'\right|\right|_\infty\left(b-a\right)$. This proves the lemma. ■

The following is a useful theorem for reducing bounded variation curves to ones which have a C^1 parameterization.

Theorem 14.2.6 *Let $\boldsymbol{\gamma} : [a,b] \to \mathbb{R}^n$ be continuous and of bounded variation. Let Ω be an open set containing $\boldsymbol{\gamma}^*$ and let $\mathbf{f} : \Omega \to \mathbb{R}^n$ be continuous, and let $\varepsilon > 0$ be given. Then there exists $\boldsymbol{\eta} : [a,b] \to \mathbb{R}^n$ such that $\boldsymbol{\eta}\left(a\right) = \boldsymbol{\gamma}\left(a\right)$, $\boldsymbol{\gamma}\left(b\right) = \boldsymbol{\eta}\left(b\right)$, $\boldsymbol{\eta} \in C^1\left([a,b]\right)$, and*

$$\left|\left|\boldsymbol{\gamma} - \boldsymbol{\eta}\right|\right| < \varepsilon, \tag{14.12}$$

where $\left|\left|\boldsymbol{\gamma} - \boldsymbol{\eta}\right|\right| \equiv \max\left\{\left|\boldsymbol{\gamma}\left(t\right) - \boldsymbol{\eta}\left(t\right)\right| : t \in [a,b]\right\}$. Also

$$\left|\int_{\boldsymbol{\gamma}}\mathbf{f}\cdot d\boldsymbol{\gamma} - \int_{\boldsymbol{\eta}}\mathbf{f}\cdot d\boldsymbol{\eta}\right| < \varepsilon, \tag{14.13}$$

$$V\left(\boldsymbol{\eta}, [a,b]\right) \leq V\left(\boldsymbol{\gamma}, [a,b]\right), \tag{14.14}$$

Proof: Extend $\boldsymbol{\gamma}$ to be defined on all $\mathbb{R}$ according to the rule $\boldsymbol{\gamma}\left(t\right) = \boldsymbol{\gamma}\left(a\right)$ if $t < a$ and $\boldsymbol{\gamma}\left(t\right) = \boldsymbol{\gamma}\left(b\right)$ if $t > b$. Now define

$$\boldsymbol{\gamma}_h\left(t\right) \equiv \frac{1}{2h}\int_{-2h+t+\frac{2h}{(b-a)}(t-a)}^{t+\frac{2h}{(b-a)}(t-a)}\boldsymbol{\gamma}\left(s\right)ds.$$

where the integral is defined in the obvious way, that is componentwise. Since $\boldsymbol{\gamma}$ is continuous, this is certainly possible. Then

$$\boldsymbol{\gamma}_h\left(b\right) \equiv \frac{1}{2h}\int_b^{b+2h}\boldsymbol{\gamma}\left(s\right)ds = \frac{1}{2h}\int_b^{b+2h}\boldsymbol{\gamma}\left(b\right)ds = \boldsymbol{\gamma}\left(b\right),$$

$$\boldsymbol{\gamma}_h(a) \equiv \frac{1}{2h}\int_{a-2h}^{a} \boldsymbol{\gamma}(s)\,ds = \frac{1}{2h}\int_{a-2h}^{a} \boldsymbol{\gamma}(a)\,ds = \boldsymbol{\gamma}(a).$$

Also, because of continuity of $\boldsymbol{\gamma}$ and the fundamental theorem of calculus,

$$\boldsymbol{\gamma}_h'(t) = \frac{1}{2h}\left\{\boldsymbol{\gamma}\left(t + \frac{2h}{b-a}(t-a)\right)\left(1+\frac{2h}{b-a}\right) - \boldsymbol{\gamma}\left(-2h + t + \frac{2h}{b-a}(t-a)\right)\left(1+\frac{2h}{b-a}\right)\right\}$$

and so $\boldsymbol{\gamma}_h \in C^1([a,b])$. The following lemma is significant.

Lemma 14.2.7 *$V(\boldsymbol{\gamma}_h,[a,b]) \le V(\boldsymbol{\gamma},[a,b])$.*

Proof: Let $a = t_0 < t_1 < \cdots < t_n = b$. Then using the definition of $\boldsymbol{\gamma}_h$ and changing the variables to make all integrals over $[0,2h]$,

$$\sum_{j=1}^{n} |\boldsymbol{\gamma}_h(t_j) - \boldsymbol{\gamma}_h(t_{j-1})| =$$

$$\sum_{j=1}^{n}\left|\frac{1}{2h}\int_0^{2h}\left[\boldsymbol{\gamma}\left(s-2h+t_j+\frac{2h}{b-a}(t_j-a)\right) - \boldsymbol{\gamma}\left(s-2h+t_{j-1}+\frac{2h}{b-a}(t_{j-1}-a)\right)\right]\right|$$

$$\le \frac{1}{2h}\int_0^{2h}\sum_{j=1}^{n}\left|\boldsymbol{\gamma}\left(s-2h+t_j+\frac{2h}{b-a}(t_j-a)\right) - \boldsymbol{\gamma}\left(s-2h+t_{j-1}+\frac{2h}{b-a}(t_{j-1}-a)\right)\right| ds.$$

For a given $s \in [0,2h]$, the points, $s-2h+t_j+\frac{2h}{b-a}(t_j-a)$ for $j=1,\cdots,n$ form an increasing list of points in the interval $[a-2h, b+2h]$ and so the integrand is bounded above by $V(\boldsymbol{\gamma},[a-2h,b+2h]) = V(\boldsymbol{\gamma},[a,b])$. It follows

$$\sum_{j=1}^{n} |\boldsymbol{\gamma}_h(t_j) - \boldsymbol{\gamma}_h(t_{j-1})| \le V(\boldsymbol{\gamma},[a,b])$$

which proves the lemma.

With this lemma the proof of the theorem can be completed without too much trouble. Let H be an open set containing $\boldsymbol{\gamma}^*$ such that $\overline{H}$ is a compact subset of Ω. Let $0 < \varepsilon < \operatorname{dist}(\boldsymbol{\gamma}^*, H^C)$. Then there exists δ_1 such that if $h < \delta_1$, then for all t,

$$\begin{aligned} |\boldsymbol{\gamma}(t) - \boldsymbol{\gamma}_h(t)| &\le \frac{1}{2h}\int_{-2h+t+\frac{2h}{(b-a)}(t-a)}^{t+\frac{2h}{(b-a)}(t-a)} |\boldsymbol{\gamma}(s) - \boldsymbol{\gamma}(t)|\,ds \\ &< \frac{1}{2h}\int_{-2h+t+\frac{2h}{(b-a)}(t-a)}^{t+\frac{2h}{(b-a)}(t-a)} \varepsilon\,ds = \varepsilon \end{aligned} \tag{14.15}$$

due to the uniform continuity of $\boldsymbol{\gamma}$. This proves 14.12.

Using the estimate from Theorem 14.2.3, 14.5, the uniform continuity of $\mathbf{f}$ on H, and the above lemma, there exists δ such that if $||\mathcal{P}|| < \delta$, then

$$\left|\int_{\boldsymbol{\gamma}} \mathbf{f}\cdot d\boldsymbol{\gamma}(t) - S(\mathcal{P})\right| < \frac{\varepsilon}{3}, \left|\int_{\boldsymbol{\gamma}_h} \mathbf{f}\cdot d\boldsymbol{\gamma}_h(t) - S_h(\mathcal{P})\right| < \frac{\varepsilon}{3}$$

for all $h < 1$. Here $S(\mathcal{P})$ is a Riemann Stieltjes sum of the form

$$\sum_{i=1}^{n} \mathbf{f}(\boldsymbol{\gamma}(\tau_i)) \cdot (\boldsymbol{\gamma}(t_i) - \boldsymbol{\gamma}(t_{i-1}))$$

and $S_h(\mathcal{P})$ is a similar Riemann Stieltjes sum taken with respect to $\boldsymbol{\gamma}_h$ instead of $\boldsymbol{\gamma}$. Because of 14.15 $\boldsymbol{\gamma}_h(t)$ has values in $H \subseteq \Omega$. Therefore, fix the partition $\mathcal{P}$, and choose h small enough that in addition to this, the following inequality is valid.

$$|S(\mathcal{P}) - S_h(\mathcal{P})| < \frac{\varepsilon}{3}$$

This is possible because of 14.15 and the uniform continuity of $\mathbf{f}$ on $\overline{H}$. It follows

$$\left| \int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma}(t) - \int_{\boldsymbol{\gamma}_h} \mathbf{f} \cdot d\boldsymbol{\gamma}_h(t) \right| \leq$$

$$\left| \int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma}(t) - S(\mathcal{P}) \right| + |S(\mathcal{P}) - S_h(\mathcal{P})|$$

$$+ \left| S_h(\mathcal{P}) - \int_{\boldsymbol{\gamma}_h} \mathbf{f} \cdot d\boldsymbol{\gamma}_h(t) \right| < \varepsilon.$$

Let $\boldsymbol{\eta} \equiv \boldsymbol{\gamma}_h$. Formula 14.14 follows from the lemma. This proves the theorem. ■

This is a very useful theorem because if $\boldsymbol{\gamma}$ is $C^1([a,b])$, it is easy to calculate $\int_{\boldsymbol{\gamma}} \mathbf{f} d\boldsymbol{\gamma}$ and the above theorem allows a reduction to the case where $\boldsymbol{\gamma}$ is C^1. The next theorem shows how easy it is to compute these integrals in the case where $\boldsymbol{\gamma}$ is C^1. First note that if $\mathbf{f}$ is continuous and $\boldsymbol{\gamma} \in C^1([a,b])$, then by Lemma 14.2.5 and the fundamental existence theorem, Theorem 14.2.3, $\int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma}$ exists.

Theorem 14.2.8 *If $\mathbf{f} : \boldsymbol{\gamma}^* \to X$ is continuous and $\boldsymbol{\gamma} : [a,b] \to \mathbb{R}^n$ is in $C^1([a,b])$ and is a parameterization, then*

$$\int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma} = \int_a^b \mathbf{f}(\boldsymbol{\gamma}(t)) \cdot \boldsymbol{\gamma}'(t)\, dt. \tag{14.16}$$

Proof: Let $\mathcal{P}$ be a partition of $[a,b]$, $\mathcal{P} = \{t_0, \cdots, t_n\}$ and $||\mathcal{P}||$ is small enough that whenever $|t-s| < ||\mathcal{P}||$,

$$|\mathbf{f}(\boldsymbol{\gamma}(t)) - \mathbf{f}(\boldsymbol{\gamma}(s))| < \varepsilon \tag{14.17}$$

and

$$\left| \int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma} - \sum_{j=1}^{n} \mathbf{f}(\boldsymbol{\gamma}(\tau_j)) \cdot (\boldsymbol{\gamma}(t_j) - \boldsymbol{\gamma}(t_{j-1})) \right| < \varepsilon.$$

Now

$$\begin{aligned} & \sum_{j=1}^{n} \mathbf{f}(\boldsymbol{\gamma}(\tau_j)) \cdot (\boldsymbol{\gamma}(t_j) - \boldsymbol{\gamma}(t_{j-1})) \\ = & \int_a^b \sum_{j=1}^{n} \mathbf{f}(\boldsymbol{\gamma}(\tau_j)) \cdot \mathcal{X}_{[t_{j-1}, t_j]}(s)\, \boldsymbol{\gamma}'(s)\, ds \end{aligned}$$

where here

$$\mathcal{X}_{[p,q]}(s) \equiv \begin{cases} 1 \text{ if } s \in [p,q] \\ 0 \text{ if } s \notin [p,q] \end{cases}.$$

Also,

$$\int_a^b \mathbf{f}(\boldsymbol{\gamma}(s)) \cdot \boldsymbol{\gamma}'(s)\, ds = \int_a^b \sum_{j=1}^n \mathbf{f}(\boldsymbol{\gamma}(s)) \cdot \mathcal{X}_{[t_{j-1},t_j]}(s)\, \boldsymbol{\gamma}'(s)\, ds$$

and thanks to 14.17,

$$\left| \overbrace{\int_a^b \sum_{j=1}^n \mathbf{f}(\boldsymbol{\gamma}(\tau_j)) \cdot \mathcal{X}_{[t_{j-1},t_j]}(s)\, \boldsymbol{\gamma}'(s)\, ds}^{=\sum_{j=1}^n \mathbf{f}(\boldsymbol{\gamma}(\tau_j))\cdot(\boldsymbol{\gamma}(t_j)-\boldsymbol{\gamma}(t_{j-1}))} - \overbrace{\int_a^b \sum_{j=1}^n \mathbf{f}(\boldsymbol{\gamma}(s)) \cdot \mathcal{X}_{[t_{j-1},t_j]}(s)\, \boldsymbol{\gamma}'(s)\, ds}^{=\int_a^b \mathbf{f}(\boldsymbol{\gamma}(s))\cdot\boldsymbol{\gamma}'(s)ds} \right|$$

$$\begin{aligned} &\leq \sum_{j=1}^n \int_{t_{j-1}}^{t_j} |\mathbf{f}(\boldsymbol{\gamma}(\tau_j)) - \mathbf{f}(\boldsymbol{\gamma}(s))|\, |\boldsymbol{\gamma}'(s)|\, ds \\ &\leq ||\boldsymbol{\gamma}'||_\infty \sum_j \varepsilon (t_j - t_{j-1}) \\ &= \varepsilon\, ||\boldsymbol{\gamma}'||_\infty (b-a). \end{aligned}$$

It follows that

$$\begin{aligned} &\left| \int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma} - \int_a^b \mathbf{f}(\boldsymbol{\gamma}(s)) \cdot \boldsymbol{\gamma}'(s)\, ds \right| \\ \leq\ &\left| \int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma} - \sum_{j=1}^n \mathbf{f}(\boldsymbol{\gamma}(\tau_j)) \cdot (\boldsymbol{\gamma}(t_j) - \boldsymbol{\gamma}(t_{j-1})) \right| \\ &+ \left| \sum_{j=1}^n \mathbf{f}(\boldsymbol{\gamma}(\tau_j)) \cdot (\boldsymbol{\gamma}(t_j) - \boldsymbol{\gamma}(t_{j-1})) - \int_a^b \mathbf{f}(\boldsymbol{\gamma}(s)) \cdot \boldsymbol{\gamma}'(s)\, ds \right| \\ \leq\ &\varepsilon\, ||\boldsymbol{\gamma}'||_\infty (b-a) + \varepsilon. \end{aligned}$$

Since ε is arbitrary, this verifies 14.16. ■

You can piece bounded variation curves together to get another bounded variation curve. You can also take the integral in the opposite direction along a given curve. There is also something called a potential.

Definition 14.2.9 *A function $\mathbf{f} : \Omega \to \mathbb{R}^n$ for Ω an open set in $\mathbb{R}^n$ has a potential if there exists a function, F, the potential, such that $\nabla F = \mathbf{f}$. Also if $\boldsymbol{\gamma}_k : [a_k, b_k] \to \mathbb{R}^n$ is continuous and of bounded variation, for $k = 1, \cdots, m$ and $\boldsymbol{\gamma}_k(b_k) = \boldsymbol{\gamma}_{k+1}(a_k)$, define*

$$\int_{\sum_{k=1}^m \boldsymbol{\gamma}_k} \mathbf{f} \cdot d\boldsymbol{\gamma}_k \equiv \sum_{k=1}^m \int_{\boldsymbol{\gamma}_k} \mathbf{f} \cdot d\boldsymbol{\gamma}_k. \tag{14.18}$$

In addition to this, for $\boldsymbol{\gamma} : [a,b] \to \mathbb{R}^n$, define $-\boldsymbol{\gamma} : [a,b] \to \mathbb{R}^n$ by $-\boldsymbol{\gamma}(t) \equiv \boldsymbol{\gamma}(b+a-t)$. Thus $\boldsymbol{\gamma}$ simply traces out the points of $\boldsymbol{\gamma}^*$ in the opposite order.

The following lemma is useful and follows quickly from Theorem 14.2.2.

Lemma 14.2.10 *In the above definition, there exists a continuous bounded variation function, $\boldsymbol{\gamma}$ defined on some closed interval, $[c,d]$, such that $\boldsymbol{\gamma}([c,d]) = \cup_{k=1}^{m}\boldsymbol{\gamma}_k([a_k,b_k])$ and $\boldsymbol{\gamma}(c) = \boldsymbol{\gamma}_1(a_1)$ while $\boldsymbol{\gamma}(d) = \boldsymbol{\gamma}_m(b_m)$. Furthermore,*

$$\int_{\boldsymbol{\gamma}} \mathbf{f}\cdot d\boldsymbol{\gamma} = \sum_{k=1}^{m}\int_{\boldsymbol{\gamma}_k} \mathbf{f}\cdot d\boldsymbol{\gamma}_k.$$

If $\boldsymbol{\gamma}:[a,b]\to\mathbb{R}^n$ is of bounded variation and continuous, then

$$\int_{\boldsymbol{\gamma}} \mathbf{f}\cdot d\boldsymbol{\gamma} = -\int_{-\boldsymbol{\gamma}} \mathbf{f}\cdot d\boldsymbol{\gamma}.$$

The following theorem shows that it is very easy to compute a line integral when the function has a potential.

Theorem 14.2.11 *Let $\boldsymbol{\gamma}:[a,b]\to\mathbb{R}^n$ be continuous and of bounded variation. Also suppose $\nabla F = \mathbf{f}$ on Ω, an open set containing $\boldsymbol{\gamma}^*$ and $\mathbf{f}$ is continuous on Ω. Then*

$$\int_{\boldsymbol{\gamma}} \mathbf{f}\cdot d\boldsymbol{\gamma} = F(\boldsymbol{\gamma}(b)) - F(\boldsymbol{\gamma}(a)).$$

Proof: By Theorem 14.2.6 there exists $\boldsymbol{\eta}\in C^1([a,b])$ such that $\boldsymbol{\gamma}(a) = \boldsymbol{\eta}(a)$, and $\boldsymbol{\gamma}(b) = \boldsymbol{\eta}(b)$ such that

$$\left|\int_{\boldsymbol{\gamma}} \mathbf{f}\cdot d\boldsymbol{\gamma} - \int_{\boldsymbol{\eta}} \mathbf{f}\cdot d\boldsymbol{\eta}\right| < \varepsilon.$$

Then from Theorem 14.2.8, since $\boldsymbol{\eta}$ is in $C^1([a,b])$, it follows from the chain rule and the fundamental theorem of calculus that

$$\begin{aligned}\int_{\boldsymbol{\eta}} \mathbf{f}\cdot d\boldsymbol{\eta} &= \int_a^b \mathbf{f}(\boldsymbol{\eta}(t))\,\boldsymbol{\eta}'(t)\,dt = \int_a^b \frac{d}{dt}F(\boldsymbol{\eta}(t))\,dt\\ &= F(\boldsymbol{\eta}(b)) - F(\boldsymbol{\eta}(a)) = F(\boldsymbol{\gamma}(b)) - F(\boldsymbol{\gamma}(a)).\end{aligned}$$

Therefore,

$$\left|(F(\boldsymbol{\gamma}(b)) - F(\boldsymbol{\gamma}(a))) - \int_{\boldsymbol{\gamma}} \mathbf{f}\cdot d\boldsymbol{\gamma}\right| < \varepsilon$$

and since $\varepsilon > 0$ is arbitrary, This proves the theorem. ■

You can prove this theorem another way without using that approximation result.

Theorem 14.2.12 *Suppose $\boldsymbol{\gamma}$ is continuous and bounded variation, a parametrization of Γ where $t\in[a,b]$. Suppose $\boldsymbol{\gamma}^*\subseteq\Omega$ an open set and that $\mathbf{f}:\Omega\to\mathbb{R}^n$ is continuous and has a potential F. Thus $\nabla F(\mathbf{x}) = \mathbf{f}(\mathbf{x})$ for $\mathbf{x}\in\Omega$. Then*

$$\int_{\gamma} \mathbf{f}\cdot d\boldsymbol{\gamma} = F(\boldsymbol{\gamma}(b)) - F(\boldsymbol{\gamma}(a))$$

Proof: Define the following function

$$h(s,t) \equiv \begin{cases} \frac{F(\boldsymbol{\gamma}(t)) - F(\boldsymbol{\gamma}(s)) - \mathbf{f}(\boldsymbol{\gamma}(s))\cdot(\boldsymbol{\gamma}(t)-\boldsymbol{\gamma}(s))}{|\boldsymbol{\gamma}(t)-\boldsymbol{\gamma}(s)|} \text{ if } \boldsymbol{\gamma}(t)-\boldsymbol{\gamma}(s)\neq\mathbf{0} \\ 0 \text{ if } \boldsymbol{\gamma}(t)-\boldsymbol{\gamma}(s) = \mathbf{0}\end{cases}$$

Then h is continuous at points (s,s). To see this note that the above reduces to

$$\frac{o(\boldsymbol{\gamma}(t)-\boldsymbol{\gamma}(s))}{|\boldsymbol{\gamma}(t)-\boldsymbol{\gamma}(s)|}$$

thus if $(s_n, t_n) \to (s, s)$, the continuity of $\boldsymbol{\gamma}$ requires this to also converge to 0.

Claim: Let $\varepsilon > 0$ be given. There exists $\delta > 0$ such that if $|t - s| < \delta$, then $\|h(s,t)\| < \varepsilon$.

Proof of claim: If not, then for some $\varepsilon > 0$, there exists (s_n, t_n) where $|t_n - s_n| < 1/n$ but $\|h(s_n, t_n)\| \geq \varepsilon$. Then by compactness, there is a subsequence, still called $\{s_n\}$ such that $s_n \to s$. It follows that $t_n \to s$ also. Therefore,

$$0 = \lim_{n\to\infty} |h(s_n, t_n)| \geq \varepsilon$$

a contradiction.

Thus whenever $|t - s| < \delta$,

$$\frac{F(\boldsymbol{\gamma}(t)) - F(\boldsymbol{\gamma}(s)) - \mathbf{f}(\boldsymbol{\gamma}(s)) \cdot (\boldsymbol{\gamma}(t) - \boldsymbol{\gamma}(s))}{|\boldsymbol{\gamma}(t) - \boldsymbol{\gamma}(s)|} < \varepsilon$$

and so

$$|F(\boldsymbol{\gamma}(t)) - F(\boldsymbol{\gamma}(s)) - \mathbf{f}(\boldsymbol{\gamma}(s)) \cdot (\boldsymbol{\gamma}(t) - \boldsymbol{\gamma}(s))| \leq \varepsilon |\boldsymbol{\gamma}(t) - \boldsymbol{\gamma}(s)|$$

So let $\|\mathcal{P}\| < \delta$ and also let δ be small enough that

$$\left| \int_\gamma \mathbf{f} \cdot d\boldsymbol{\gamma} - \sum_{k=1}^n \mathbf{f}(\boldsymbol{\gamma}(t_{i-1})) \cdot (\boldsymbol{\gamma}(t_i) - \boldsymbol{\gamma}(t_{i-1})) \right| < \varepsilon$$

Therefore,

$$\left| \int_\gamma \mathbf{f} \cdot d\boldsymbol{\gamma} - \sum_{k=1}^n F(\boldsymbol{\gamma}(t_i)) - F(\boldsymbol{\gamma}(t_{i-1})) \right| \leq$$

$$\left| \int_\gamma \mathbf{f} \cdot d\boldsymbol{\gamma} - \sum_{k=1}^n \mathbf{f}(\boldsymbol{\gamma}(t_{i-1})) \cdot (\boldsymbol{\gamma}(t_i) - \boldsymbol{\gamma}(t_{i-1})) \right| +$$

$$\left| \sum_{k=1}^n \mathbf{f}(\boldsymbol{\gamma}(t_{i-1})) \cdot (\boldsymbol{\gamma}(t_i) - \boldsymbol{\gamma}(t_{i-1})) - (F(\boldsymbol{\gamma}(t_i)) - F(\boldsymbol{\gamma}(t_{i-1}))) \right|$$

$$\leq \varepsilon + \sum_{k=1}^n \varepsilon |\boldsymbol{\gamma}(t_i) - \boldsymbol{\gamma}(t_{i-1})| \leq \varepsilon (1 + V(\boldsymbol{\gamma}, [a,b]))$$

It follows that

$$\begin{aligned} & \left| \int_\gamma \mathbf{f} \cdot d\boldsymbol{\gamma} - (F(\boldsymbol{\gamma}(b)) - F(\boldsymbol{\gamma}(a))) \right| \\ = & \left| \int_\gamma \mathbf{f} \cdot d\boldsymbol{\gamma} - \sum_{k=1}^n F(\boldsymbol{\gamma}(t_i)) - F(\boldsymbol{\gamma}(t_{i-1})) \right| \leq \varepsilon (1 + V(\boldsymbol{\gamma})) \end{aligned}$$

Therefore, since ε is arbitrary,

$$\int_\gamma \mathbf{f} \cdot d\boldsymbol{\gamma} = F(\boldsymbol{\gamma}(b)) - F(\boldsymbol{\gamma}(a)) \ \blacksquare$$

Corollary 14.2.13 *If $\boldsymbol{\gamma} : [a,b] \to \mathbb{R}^n$ is continuous, has bounded variation, is a closed curve, $\boldsymbol{\gamma}(a) = \boldsymbol{\gamma}(b)$, and $\boldsymbol{\gamma}^* \subseteq \Omega$ where Ω is an open set on which $\nabla F = \mathbf{f}$, then*

$$\int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma} = 0.$$

Theorem 14.2.14 *Let Ω be a connected open set and let $\mathbf{f}:\Omega\to\mathbb{R}^n$ be continuous. Then $\mathbf{f}$ has a potential F if and only if*

$$\int_{\boldsymbol{\gamma}}\mathbf{f}\cdot d\boldsymbol{\gamma}$$

is path independent for all $\boldsymbol{\gamma}$ a bounded variation curve such that $\boldsymbol{\gamma}^*$ is contained in Ω. This means the above line integral depends only on $\boldsymbol{\gamma}(a)$ and $\boldsymbol{\gamma}(b)$.

Proof: The first part was proved in Theorem 14.2.11. It remains to verify the existence of a potential in the situation of path independence.

Let $x_0\in\Omega$ be fixed. Let S be the points $\mathbf{x}$ of Ω which have the property there is a bounded variation curve joining $\mathbf{x}_0$ to $\mathbf{x}$. Let $\boldsymbol{\gamma}_{\mathbf{x}_0\mathbf{x}}$ denote such a curve. Note first that S is nonempty. To see this, $B(\mathbf{x}_0,r)\subseteq\Omega$ for r small enough. Every $\mathbf{x}\in B(\mathbf{x}_0,r)$ is in S. Then S is open because if $\mathbf{x}\in S$, then $B(\mathbf{x},r)\subseteq\Omega$ for small enough r and if $\mathbf{y}\in B(\mathbf{x},r)$, you could go take $\boldsymbol{\gamma}_{\mathbf{x}_0\mathbf{x}}$ and from $\mathbf{x}$ follow the straight line segment joining $\mathbf{x}$ to $\mathbf{y}$. In addition to this, $\Omega\setminus S$ must also be open because if $\mathbf{x}\in\Omega\setminus S$, then choosing $B(\mathbf{x},r)\subseteq\Omega$, no point of $B(\mathbf{x},r)$ can be in S because then you could take the straight line segment from that point to $\mathbf{x}$ and conclude that $\mathbf{x}\in S$ after all. Therefore, since Ω is connected, it follows $\Omega\setminus S=\emptyset$. Thus for every $\mathbf{x}\in S$, there exists $\boldsymbol{\gamma}_{\mathbf{x}_0\mathbf{x}}$, a bounded variation curve from $\mathbf{x}_0$ to $\mathbf{x}$.

Define

$$F(\mathbf{x})\equiv\int_{\boldsymbol{\gamma}_{\mathbf{x}_0\mathbf{x}}}\mathbf{f}\cdot d\boldsymbol{\gamma}_{\mathbf{x}_0\mathbf{x}}$$

F is well defined by assumption. Now let $l_{\mathbf{x}(\mathbf{x}+t\mathbf{e}_k)}$ denote the linear segment from $\mathbf{x}$ to $\mathbf{x}+t\mathbf{e}_k$. Thus to get to $\mathbf{x}+t\mathbf{e}_k$ you could first follow $\boldsymbol{\gamma}_{\mathbf{x}_0\mathbf{x}}$ to $\mathbf{x}$ and from there follow $l_{\mathbf{x}(\mathbf{x}+t\mathbf{e}_k)}$ to $\mathbf{x}+t\mathbf{e}_k$. Hence

$$\frac{F(\mathbf{x}+t\mathbf{e}_k)-F(\mathbf{x})}{t}=\frac{1}{t}\int_{l_{\mathbf{x}(\mathbf{x}+t\mathbf{e}_k)}}\mathbf{f}\cdot dl_{\mathbf{x}(\mathbf{x}+t\mathbf{e}_k)}$$

$$=\frac{1}{t}\int_0^t\mathbf{f}(\mathbf{x}+s\mathbf{e}_k)\cdot\mathbf{e}_k ds\to f_k(\mathbf{x})$$

by continuity of $\mathbf{f}$. Thus $\nabla F=\mathbf{f}$ and This proves the theorem. ■

Corollary 14.2.15 *Let Ω be a connected open set and $\mathbf{f}:\Omega\to\mathbb{R}^n$. Then $\mathbf{f}$ has a potential if and only if every closed, $\boldsymbol{\gamma}(a)=\boldsymbol{\gamma}(b)$, bounded variation curve contained in Ω has the property that*

$$\int_{\boldsymbol{\gamma}}\mathbf{f}\cdot d\boldsymbol{\gamma}=0$$

Proof: Using Lemma 14.2.10, this condition about closed curves is equivalent to the condition that the line integrals of the above theorem are path independent. This proves the corollary. ■

Such a vector valued function is called conservative.

14.3 Simple Closed Rectifiable Curves

There are examples of space filling continuous curves. However, bounded variation curves are not like this. In fact, one can even say the two dimensional Lebesgue measure of a bounded variation curve is 0.

Theorem 14.3.1 *Let $\boldsymbol{\gamma} : [a,b] \to \boldsymbol{\gamma}^* \subseteq \mathbb{R}^n$ where $n \geq 2$ is a continuous bounded variation curve. Then*

$$m_n(\boldsymbol{\gamma}^*) = 0$$

where m_n denotes n dimensional Lebesgue measure.

Proof: Let $\varepsilon > 0$ be given. Let $t_0 \equiv a$ and if $t_0, \cdots, t_k$ have been chosen, let t_{k+1} be the first number larger than t_k such that

$$|\boldsymbol{\gamma}(t_{k+1}) - \boldsymbol{\gamma}(t_k)| = \varepsilon.$$

If the set of t such that $|\boldsymbol{\gamma}(t) - \boldsymbol{\gamma}(t_k)| = \varepsilon$ is nonempty, then this set is clearly closed and so such a t_{k+1} exists until k is such that

$$\boldsymbol{\gamma}^* \subseteq \cup_{j=0}^{k} B(\boldsymbol{\gamma}(t_j), \varepsilon)$$

Let m be the last index of this process where t_{m+1} does not exist. How large is m? This can be estimated because

$$V(\boldsymbol{\gamma}, [a,b]) \geq \sum_{k=0}^{m} |\boldsymbol{\gamma}(t_{k+1}) - \boldsymbol{\gamma}(t_k)| = m\varepsilon$$

and so $m \leq V(\boldsymbol{\gamma}, [a,b])/\varepsilon$. Since $\boldsymbol{\gamma}^* \subseteq \cup_{j=0}^{m} B(\boldsymbol{\gamma}(t_j), \varepsilon)$,

$$\begin{aligned} m_n(\boldsymbol{\gamma}^*) &\leq \sum_{j=0}^{m} m_n(B(\boldsymbol{\gamma}(t_j), \varepsilon)) \\ &\leq \frac{V(\boldsymbol{\gamma}, [a,b])}{\varepsilon} c_n \varepsilon^n = c_n V(\boldsymbol{\gamma}, [a,b]) \varepsilon^{n-1} \end{aligned}$$

Since ε was arbitrary, This proves the theorem. ■

Since a ball has positive measure, this proves the following corollary.

Corollary 14.3.2 *Let $\boldsymbol{\gamma} : [a,b] \to \boldsymbol{\gamma}^* \subseteq \mathbb{R}^n$ where $n \geq 2$ is a continuous bounded variation curve. Then $\boldsymbol{\gamma}^*$ has empty interior.*

Lemma 14.3.3 *Let Γ be a simple closed curve. Then there exists a mapping $\boldsymbol{\theta} : S^1 \to \Gamma$ where S^1 is the unit circle*

$$\left\{(x,y) : x^2 + y^2 = 1\right\},$$

such that $\boldsymbol{\theta}$ is one to one and continuous.

Proof: Since Γ is a simple closed curve, there is a parameterization $\boldsymbol{\gamma}$ and an interval $[a,b]$ such that $\boldsymbol{\gamma}$ is continuous and one to one on $[a,b)$ and $(a,b]$ with $\boldsymbol{\gamma}(a) = \boldsymbol{\gamma}(b)$. Define $\boldsymbol{\theta}^{-1} : \Gamma \to S^1$ by

$$\boldsymbol{\theta}^{-1}(\mathbf{x}) \equiv \left(\cos\left(\frac{2\pi}{b-a}\left(\boldsymbol{\gamma}^{-1}(\mathbf{x}) - a\right)\right), \sin\left(\frac{2\pi}{b-a}\left(\boldsymbol{\gamma}^{-1}(\mathbf{x}) - a\right)\right)\right)$$

Note that $\boldsymbol{\theta}^{-1}$ is onto S^1. The function is well defined because it sends the point $\boldsymbol{\gamma}(a) = \boldsymbol{\gamma}(b)$ to the same point, $(1,0)$. It is also one to one. To see this note $\boldsymbol{\gamma}^{-1}$ is one to one on $\Gamma \setminus \{\boldsymbol{\gamma}(a), \boldsymbol{\gamma}(b)\}$. What about the case where $\mathbf{x} \neq \boldsymbol{\gamma}(a) = \boldsymbol{\gamma}(b)$? Could $\boldsymbol{\theta}^{-1}(\mathbf{x}) = \boldsymbol{\theta}^{-1}(\boldsymbol{\gamma}(a))$? In this case, $\boldsymbol{\gamma}^{-1}(\mathbf{x})$ is in (a,b) while $\boldsymbol{\gamma}^{-1}(\boldsymbol{\gamma}(a)) = a$ so

$$\boldsymbol{\theta}^{-1}(\mathbf{x}) \neq \boldsymbol{\theta}^{-1}(\boldsymbol{\gamma}(a)) = (1,0).$$

Thus $\boldsymbol{\theta}^{-1}$ is one to one on Γ.

Why is $\boldsymbol{\theta}^{-1}$ continuous? Suppose $\mathbf{x}_n \to \boldsymbol{\gamma}(a) = \boldsymbol{\gamma}(b)$ first. Why does $\boldsymbol{\theta}^{-1}(\mathbf{x}_n) \to (1,0) = \boldsymbol{\theta}^{-1}(\boldsymbol{\gamma}(a))$? Let $\{\mathbf{x}_n\}$ denote any subsequence of the given sequence. Then by compactness of $[a,b]$ there exists a further subsequence, still denoted by $\mathbf{x}_n$ such that

$$\boldsymbol{\gamma}^{-1}(\mathbf{x}_n) \to t \in [a,b]$$

Hence by continuity of $\boldsymbol{\gamma}$, $\mathbf{x}_n \to \boldsymbol{\gamma}(t)$ and so $\boldsymbol{\gamma}(t)$ must equal $\boldsymbol{\gamma}(a) = \boldsymbol{\gamma}(b)$. It follows from the assumption of what a simple curve is that $t \in \{a,b\}$. Hence $\boldsymbol{\theta}^{-1}(\mathbf{x}_n)$ converges to either

$$\left(\cos\left(\frac{2\pi}{b-a}(a-a)\right), \sin\left(\frac{2\pi}{b-a}(a-a)\right)\right)$$

or

$$\left(\cos\left(\frac{2\pi}{b-a}(b-a)\right), \sin\left(\frac{2\pi}{b-a}(b-a)\right)\right)$$

but these are the same point. This has shown that if $\mathbf{x}_n \to \boldsymbol{\gamma}(a) = \boldsymbol{\gamma}(b)$, there is a subsequence such that $\boldsymbol{\theta}^{-1}(\mathbf{x}_n) \to \boldsymbol{\theta}^{-1}(\boldsymbol{\gamma}(a))$. Thus $\boldsymbol{\theta}^{-1}$ is continuous at $\boldsymbol{\gamma}(a) = \boldsymbol{\gamma}(b)$. Next suppose $\mathbf{x}_n \to \mathbf{x} \neq \boldsymbol{\gamma}(a) \equiv \mathbf{p}$. Then there exists $B(\mathbf{p}, r)$ such that for all n large enough, $\mathbf{x}_n$ and $\mathbf{x}$ are contained in the compact set $\Gamma \setminus B(\mathbf{p}, r) \equiv K$. Then $\boldsymbol{\gamma}$ is continuous and one to one on the compact set $\boldsymbol{\gamma}^{-1}(K) \subseteq (a,b)$ and so by Theorem 5.1.3 $\boldsymbol{\gamma}^{-1}$ is continuous on K. In particular it is continuous at $\mathbf{x}$ so $\boldsymbol{\theta}^{-1}(\mathbf{x}_n) \to \boldsymbol{\theta}^{-1}(\mathbf{x})$. This proves the lemma. ■

14.3.1 The Jordan Curve Theorem

Recall the Jordan curve theorem, Theorem 13.0.28. Here it is stated for convenience. Recall that for O an open set in $\mathbb{R}^2$, $\partial O = \bar{O} \setminus O$.

Theorem 14.3.4 *Let C denote the unit circle, $\{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$. Suppose $\boldsymbol{\gamma} : C \to \Gamma \subseteq \mathbb{R}^2$ is one to one onto and continuous. Then $\mathbb{R}^2 \setminus \Gamma$ consists of two components, a bounded component (called the inside) U_1 and an unbounded component (called the outside), U_2. Also the boundary of each of these two components of $\mathbb{R}^2 \setminus \Gamma$ is Γ and Γ has empty interior.*

Proof: Here is a sketch of an alternative proof based on degree theory. The main thing is to show that Γ is the same as ∂U_i. By the Jordan separation theorem, $\mathbb{R}^2 = U_1 \cup \Gamma \cup U_2$ where U_i is open and connected, one is bounded and the other is unbounded. Then by this theorem again, Γ has empty interior. This is because $\boldsymbol{\gamma}^{-1}$ is continuous and C has empty interior. If Γ contained a ball, then invariance of domain would imply that this ball would map to an open set of $\mathbb{R}^2$ contained in C. Thus each point of Γ must be in one ∂U_i. Otherwise Γ would have non-empty interior. The idea is to show that $\partial U_1 = \partial U_2$. Suppose $\partial U_1 \setminus \partial U_2 \neq \emptyset$. Then let p be a point in this set. Then p is not in U_2 because if it were, there would be points of U_1 in U_2. There exists $B(p,r)$ which has empty intersection with ∂U_2. If for every r this ball has points of U_2 then $p \in \partial U_2$ because it is not in U_2. However, it was assumed this is not the case. Hence $B(p,r) \cap \overline{U_2} = \emptyset$ for suitable positive r. Therefore, $B(p,r) \subseteq \overline{U_1}$ since, as noted above, $\overline{U_1} \cup \overline{U_2} = \mathbb{R}^2$. Now here is a claim.

Claim: Let $S \equiv \{p \in \Gamma : B(p,r) \subseteq \overline{U_1} \text{ for some } r > 0\}$. Then $S = \emptyset$.

Proof of claim: It is clear that S is the intersection of an open set with Γ. Let $\hat{\Gamma} \equiv \Gamma \setminus S$. Thus $\hat{\Gamma}$ is a compact set and it is homeomorphic to $\hat{C}$, a compact proper subset of C. Since an open set is missing, $\hat{C}^C$ consists of only one connected component. Hence the same is true of $\hat{\Gamma}^C$. Let $x \in U_2$ and $y \in U_1$. Then there is a continuous curve which goes from x to y which misses $\hat{\Gamma}$. Therefore, it must contain some point of Γ. However,

since it misses $\hat{\Gamma}$, this point must be in S. Let p be the first point of Γ encountered by this curve in going from x to y, which must be in S. Therefore, p is in ∂U_2. Thus for some $r > 0$, $B(p,r) \subseteq \overline{U_1}$ which implies that there are no points of U_2 in this ball. But this contradicts $p \in \partial U_2$ because if $p \in \partial U_2$, then in the given ball there must be points of U_2.

This has proved that $\partial U_1 \setminus \partial U_2 = \emptyset$. Hence $\partial U_1 \subseteq \partial U_2$. Similarly $\partial U_2 \subseteq \partial U_1$. Hence $\Gamma = \partial U_1 \cup \partial U_2 = \partial U_i$. ■

The following lemma will be of importance in what follows. To say two sets are homeomorphic is to say there is a one to one continuous onto mapping from one to the other which also has continuous inverse. Clearly the statement that two sets are homeomorphic defines an equivalence relation. Then from Lemma 13.0.26 a Jordan curve is just a curve which is homeomorphic to a circle.

From now on, we will refer to U_o as the outside component which is unbounded and U_i as the inside component. Also I will often write $\boldsymbol{\gamma}^*$ to emphasize that the set of points in the curve is being considered rather than some parametrization.

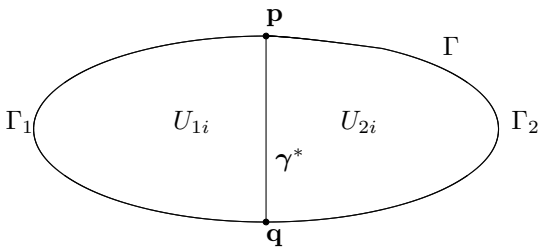
Lemma 14.3.5 *In the situation of Theorem 13.0.28, let Γ be a simple closed curve and let $\boldsymbol{\gamma}^*$ be a straight line segment such that the open segment, $\boldsymbol{\gamma}_o^*, \boldsymbol{\gamma}^*$ without its endpoints, is contained in U_i such that the intersection of $\boldsymbol{\gamma}^*$ with Γ equals $\{\mathbf{p}, \mathbf{q}\}$. Then this line segment divides U_i into two connected open sets U_{1i}, U_{2i} which are the insides of two simple closed curves such that*

$$U_i = U_{1i} \cup \boldsymbol{\gamma}_o^* \cup U_{2i}$$

Proof: Denote by C the unit circle and let $\boldsymbol{\theta} : C \to \Gamma$ be continuous one to one and onto. Say $\boldsymbol{\theta}(\mathbf{a}) = \mathbf{p}$ and $\boldsymbol{\theta}(\mathbf{b}) = \mathbf{q}$. Let $C_j, j = 1, 2$ denote the two circular arcs joining $\mathbf{a}$ and $\mathbf{b}$. Thus letting $\Gamma_j \equiv \boldsymbol{\theta}(C_j)$ it follows Γ_1, Γ_2 are simple curves whose union is Γ which intersect at the points $\mathbf{p}$ and $\mathbf{q}$. Letting $\Gamma_j \cup \boldsymbol{\gamma}^* \equiv J_j$ it follows J_j is a simple closed curve. Here is why. Define

$$\mathbf{h}_1(\mathbf{x}) \equiv \begin{cases} \boldsymbol{\theta}(\mathbf{x}) & \text{if } \mathbf{x} \in C_1 \\ \mathbf{f}_2(\mathbf{x}) & \text{if } \mathbf{x} \in C_2 \end{cases}$$

where $\mathbf{f}_j$ is a continuous one to one onto mapping from C_j to $\boldsymbol{\gamma}^*$. Then $\mathbf{h}_1$ is continuous and maps C one to one and onto J_1. Define $\mathbf{h}_2$ similarly. Denote by U_{ji} the inside and U_{jo} the outside of J_j.



Claim 1: $U_{1i}, U_{2i} \subseteq U_i$, U_{1i}, U_{2i} contain no points of $J_2 \cup J_1$.

Proof: First consider the claim that U_{1i}, U_{2i} contain no points of $J_2 \cup J_1$. If $\mathbf{x} \in \Gamma_2 \setminus \{\mathbf{p}, \mathbf{q}\}$, then near $\mathbf{x}$ there are points of U_i and U_o. Therefore, if $\mathbf{x} \in U_{1i}$, there would be points of both U_i and U_o in U_{1i}. Now U_o contains no points of $\boldsymbol{\gamma}^*$ by assumption and it contains no points of Γ by definition. Therefore, since U_o is connected, it must be contained in U_{1i} but this is impossible because U_o is unbounded and U_{1i} is bounded. Thus $\mathbf{x} \notin U_{1i}$. Similarly $\mathbf{x} \notin U_{2i}$. Similar reasoning applied to $\mathbf{x} \in \Gamma_1 \setminus \{\mathbf{p}, \mathbf{q}\}$ implies no point of $\Gamma \setminus \{\mathbf{p}, \mathbf{q}\}$ can be in either U_{1i} or U_{2i}. If $\mathbf{x} \in \boldsymbol{\gamma}^*$ then by definition it

is not in either U_{1i} or U_{2i} and so neither U_{1i} not U_{2i} contain any points of $J_2 \cup J_1$. If $U_{1i} \cap U_o \neq \emptyset$, then the whole connected set U_{1i} is contained in U_o since otherwise U_{1i} would contain points of Γ which, by what was just shown, is not the case. But now this is a contradiction to the open segment of $\boldsymbol{\gamma}^*$ being contained in U_i. A point $\mathbf{x}$ on this open segment must have points of U_{1i} near it by the Jordan curve theorem but if $U_{1i} \subseteq U_o$, then this cannot happen because a small ball centered at $\mathbf{x}$ contains only points of U_i and none of U_o.

Similarly $U_{2i} \subseteq U_i$. Letting $\boldsymbol{\gamma}_o^*$ denote the open segment of $\boldsymbol{\gamma}^*$, it follows

$$U_i \supseteq U_{1i} \cup \boldsymbol{\gamma}_o^* \cup U_{2i} \tag{14.19}$$

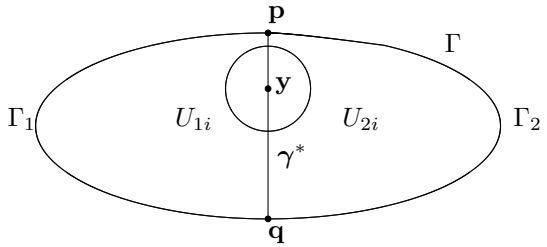
Claim 2: $U_{1i} \cap U_{2i} = \emptyset, U_{2i} \subseteq U_{1o}, U_{1i} \subseteq U_{2o}$.

Proof: If $\mathbf{x}$ is a point of U_{1i} which is in U_{2i} then U_{1i} must be contained in U_{2i}. This is because U_{1i} is connected and, as noted above, U_{1i} contains no boundary points of U_{2i} because these points are all contained in $J_2 \cup J_2$. Similarly U_{2i} would need to be contained in U_{1i} and so these two interior components would need to coincide. But now consider $\mathbf{x} \in \Gamma_2 \setminus \{\mathbf{p}, \mathbf{q}\}$. Let r be small enough that $B(\mathbf{x}, r) \cap J_1 = \emptyset$. Then this ball contains points of $U_{2i} = U_{1i}$ and points of U_{2o}. However, this ball is contained in one of the complementary components of J_1 and since it contains points of U_{1i}, this forces $B(\mathbf{x}, r)$ to be contained in U_{1i}. But now the same contradiction as above holds namely: There exists a point of U_o in U_{1i} which forces U_o to be contained in U_{1i}. Therefore, $U_{1i} \cap U_{2i} = \emptyset$. Also $U_{2i} \subseteq U_{1o}$ because if not, there would be points of U_{1i} in U_{2i} and it was just shown this doesn't happen. Similarly $U_{1i} \subseteq U_{2o}$. This shows Claim 2.

Next I need to verify that equality holds in 14.19. First I will argue

$$U_{1i} \cup \boldsymbol{\gamma}_o^* \cup U_{2i}$$

is an open set. Let $\mathbf{y} \in \boldsymbol{\gamma}_o^*$ and let $B(\mathbf{y}, r) \cap \Gamma = \emptyset$.



Thus $\boldsymbol{\gamma}_o^*$ divides $B(\mathbf{y}, r)$ into two halves, H_1, H_2. The ball contains points of U_{2i} by the Jordan curve theorem. Say H_2 contains some of these points. Then I claim H_2 cannot contain any points of U_{1i}. This is because if it did, there would be a segment joining a point of U_{1i} with a point of U_{2i} which is contained in H_2 which is a connected open set which is therefore contained in a single component of J_1^C. This is a contradiction because as shown above, $U_{2i} \subseteq U_{1o}$. Could H_2 contain any points of U_{2o}? No because then there would be a segment joining a point of U_{2o} to a point of U_{2i} which is contained in the same component of J_2^C. Therefore, H_2 consists entirely of points of U_{2i}. Similarly H_1 consists entirely of points of U_{1i}. Therefore, $U_{1i} \cup \boldsymbol{\gamma}_o^* \cup U_{2i}$ is an open set because the only points which could possibly fail to be interior points, those on $\boldsymbol{\gamma}_0^*$ are interior points of $U_{1i} \cup \boldsymbol{\gamma}_o^* \cup U_{2i}$.

Suppose equality does not hold in 14.19. Then there exists $\mathbf{w} \in U_i \setminus (U_{1i} \cup \boldsymbol{\gamma}_o^* \cup U_{2i})$. Let $\mathbf{x} \in U_{1i} \cup \boldsymbol{\gamma}_o^* \cup U_{2i}$. Then since U_i is connected and open, there exists a continuous mapping $\mathbf{r} : [0,1] \to U_i$ such that $\mathbf{r}(0) = \mathbf{x}$ and $\mathbf{r}(1) = \mathbf{w}$. Since $U_{1i} \cup \boldsymbol{\gamma}_o^* \cup U_{2i}$ is open, there exists a first point in the closed set $\mathbf{r}^{-1}\left((U_{1i} \cup \boldsymbol{\gamma}_o^* \cup U_{2i})^C\right)$, s. Thus $\mathbf{r}(s)$ is a

limit point of $U_{1i} \cup \boldsymbol{\gamma}_o^* \cup U_{2i}$ but is not in this set which implies it is in $U_{1o} \cap U_{2o}$. It follows $\mathbf{r}(s)$ is a limit point of either U_{1i} or U_{2i} because each point of $\boldsymbol{\gamma}_o^*$ is a limit point of U_{1i} and U_{2i}. Also, $\mathbf{r}(s)$ cannot be in $\boldsymbol{\gamma}_0^*$ because it is not in $U_{1i} \cup \boldsymbol{\gamma}_o^* \cup U_{2i}$. Suppose without loss of generality it is a limit point of U_{1i}. Then every ball containing $\mathbf{r}(s)$ must contain points of $U_{1o} \cap U_{2o} \subseteq U_{1o}$ as well as points U_{1i}. But by the Jordan curve theorem, this implies $\mathbf{r}(s)$ is in J_1 but is not in $\boldsymbol{\gamma}_o^*$. Therefore, $\mathbf{r}(s)$ is a point of Γ and this contradicts $\mathbf{r}(s) \in U_i$. Therefore, equality must hold in 14.19 after all. This proves the lemma. ■

The following lemma has to do with decomposing the inside and boundary of a simple closed rectifiable curve into small pieces. The argument is like one given in Apostol [3]. In doing this I will refer to a region as the union of a connected open set with its boundary. Also, two regions will be said to be non overlapping if they either have empty intersection or the intersection is contained in the intersection of their boundaries.The height of a set A equals $\sup\{|y_1 - y_2| : (x_1, y_1), (x_2, y_2) \in A\}$. The width of A will be defined similarly.

Lemma 14.3.6 *Let Γ be a simple closed rectifiable curve. Also let $\delta > 0$ be given such that 2δ is smaller than both the height and width of Γ. Then there exist finitely many non overlapping regions $\{R_k\}_{k=1}^n$ consisting of simple closed rectifiable curves along with their interiors whose union equals $U_i \cup \Gamma$. These regions consist of two kinds, those contained in U_i and those with nonempty intersection with Γ. These latter regions are called "border" regions. The boundary of a border region consists of straight line segments parallel to the coordinate axes of the form $x = m\delta$ or $y = k\delta$ for m, k integers along with arcs from Γ. The regions contained in U_i consist of rectangles. Thus all of these regions have boundaries which are rectifiable simple closed curves. Also each region is contained in a square having sides of length no more than 2δ. There are at most*

$$4\left(\frac{V(\Gamma)}{\delta} + 1\right)$$

border regions. The construction also yields an orientation for Γ and for all these regions, and the orientations for any segment shared by two regions are opposite.

Proof: Let $\Gamma = \boldsymbol{\gamma}([a,b])$ where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$. Let

$$y_1 \equiv \max\{\gamma_2(t) : t \in [a,b]\}$$

and let

$$y_2 \equiv \min\{\gamma_2(t) : t \in [a,b]\}.$$

Thus $(x_1, y_1), x_1 \equiv \gamma_1\left(\gamma_2^{-1}(y_1)\right)$ is the "top" point of Γ while (x_2, y_2) is the "bottom" point of Γ. Consider the lines $y = y_1$ and $y = y_2$. By assumption $|y_1 - y_2| > 2\delta$. Consider the line l given by $y = m\delta$ where m is chosen to make $m\delta$ as close as possible to $(y_1 + y_2)/2$. Thus $y_1 > m\delta > y_2$. By Theorem 13.0.28 (x_j, y_j) $j = 1, 2$, being on Γ are both limit points of U_i so there exist points $\mathbf{p}_j \in U_i$ such that $\mathbf{p}_1$ is above l and $\mathbf{p}_2$ is below l. (Simply pick $\mathbf{p}_j$ very close to (x_j, y_j) and yet in U_i and this will take place.) The horizontal line l must have nonempty intersection with U_i because U_i is connected. If it had empty intersection it would be possible to separate U_i into two nonempty open sets, one containing $\mathbf{p}_1$ and the other containing $\mathbf{p}_2$.

Let $\mathbf{q}$ be a point of U_i which is also in l. Then there exists a maximal segment of the line l containing $\mathbf{q}$ which is contained in $U_i \cup \Gamma$. This segment, $\boldsymbol{\gamma}^*$ satisfies the conditions of Lemma 14.3.5 and so it divides U_i into disjoint open connected sets whose boundaries are simple rectifiable closed curves. Note the line segment has finite length. Letting Γ_j be the simple closed curve which contains $\mathbf{p}_j$, orient $\boldsymbol{\gamma}^*$ as part of Γ_2 such that motion is from right to left. As part of Γ_1 the motion along $\boldsymbol{\gamma}^*$ is from left to right.

By Proposition 14.1.7 this provides an orientation to each Γ_j. By Proposition 14.1.8 there exists an orientation for Γ which is consistent with these two orientations on the Γ_j.

Now do the same process to the two simple closed curves just obtained and continue till all regions have height less than 2δ. Each application of the process yields two new non overlapping regions of the desired sort in place of an earlier region of the desired sort except possibly the regions might have excessive height. The orientation of a new line segment in the construction is determined from the orientations of the simple closed curves obtained earlier. By Proposition 14.1.7 the orientations of the segments shared by two regions are opposite so eventually the line integrals over these segments cancel. Eventually this process ends because all regions have height less than 2δ. The reason for this is that if it did not end, the curve Γ could not have finite total variation because there would exist an arbitrarily large number of non overlapping regions each of which have a pair of points which are farther apart than 2δ. This takes care of finding the subregions so far as height is concerned.

Now follow the same process just described on each of the non overlapping "short" regions just obtained using vertical rather than horizontal lines, letting the orientation of the vertical edges be determined from the orientation already obtained, but this time feature width instead of height and let the lines be vertical of the form $x = k\delta$ where k is an integer.

How many border regions are there? Denote by $V(\Gamma)$ the length of Γ. Now decompose Γ into N arcs of length δ with maybe one having length less than δ. Thus $N - 1 \leq \frac{V(\Gamma)}{\delta}$ and so

$$N \leq \frac{V(\Gamma)}{\delta} + 1$$

The resulting regions are each contained in a box having sides of length no more than 2δ in length. Each of these N arcs can't intersect any more than four of these boxes because of their short length. Therefore, at most $4N$ boxes of the construction can intersect Γ. Thus there are no more than

$$4\left(\frac{V(\Gamma)}{\delta} + 1\right)$$

border regions. This proves the lemma. ■

Note that this shows that the region of U_i which is included by the boundary boxes has area at most equal to

$$4\left(\frac{V(\Gamma)}{\delta} + 1\right)\left(16\delta^2\right)$$

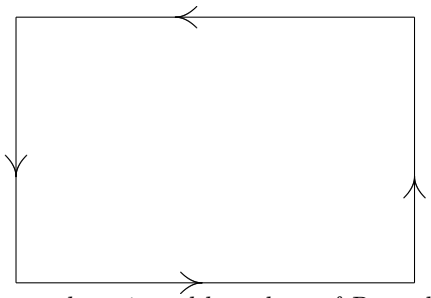
which converges to 0 as $\delta \to 0$.

14.3.2 Orientation And Green's Formula

How do you describe the orientation of a simple closed rectifiable curve analytically? The above process did it but I want another way to identify this which is more geometrically appealing. For simple examples, this is not too hard but it becomes less obvious when you consider the general case. The problem is the simple closed curve could be very wriggly.

The orientation of a rectifiable simple closed curve will be defined in terms of a very important formula known as Green's formula. First I will present Green's formula for a rectangle. In this lemma, it is very easy to understand the orientation of the bounding curve. The direction of motion is counter clockwise. As described in Proposition 14.1.7 it suffices to describe a direction of motion along the curve using any two points.

Lemma 14.3.7 *Let $R = [a,b] \times [c,d]$ be a rectangle and let P, Q be functions which are C^1 in some open set containing R. Orient the boundary of R as shown in the following picture. This is called the counter clockwise direction or the positive orientation*



Then letting $\boldsymbol{\gamma}$ denote the oriented boundary of R as shown,

$$\int_R \left(Q_x(x,y) - P_y(x,y)\right) dm_2 = \int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma}$$

where

$$\mathbf{f}(x,y) \equiv (P(x,y), Q(x,y)).$$

In this context the line integral is usually written using the notation

$$\int_{\partial R} P dx + Q dy.$$

Proof: This follows from direct computation. A parameterization for the bottom line of R is

$$\boldsymbol{\gamma}_B(t) = (a + t(b-a), c), \ t \in [0,1]$$

A parameterization for the top line of R with the given orientation is

$$\boldsymbol{\gamma}_T(t) = (b + t(a-b), d), \ t \in [0,1]$$

A parameterization for the line on the right side is

$$\boldsymbol{\gamma}_R(t) = (b, c + t(d-c)), \ t \in [0,1]$$

and a parameterization for the line on the left side is

$$\boldsymbol{\gamma}_L(t) = (a, d + t(c-d)), \ t \in [0,1]$$

Now it is time to do the computations using Theorem 14.2.8.

$$\begin{aligned} \int_{\boldsymbol{\gamma}} \mathbf{f} \cdot d\boldsymbol{\gamma} &= \int_0^1 P(a + t(b-a), c)(b-a)\, dt \\ &\quad + \int_0^1 P(b + t(a-b), d)(a-b)\, dt \end{aligned}$$

$$+ \int_0^1 Q(b, c + t(d-c))(d-c)\, dt + \int_0^1 Q(a, d + t(c-d))(c-d)\, dt$$

Changing the variables and combining the integrals, this equals

$$= \int_a^b P(x,c)\, dx - \int_a^b P(x,d)\, dx + \int_c^d Q(b,y)\, dy - \int_c^d Q(a,y)\, dy$$

$$\begin{aligned} &= -\int_a^b \int_c^d P_y(x,y)\,dydx + \int_c^d \int_a^b Q_x(x,y)\,dxdy \\ &= \int_R (Q_x - P_y)\,dm_2 \end{aligned}$$

By Fubini's theorem, Theorem 9.2.3 on Page 218. (To use this theorem you can extend the functions to equal 0 off R.) This proves the lemma. ■

Note that if the rectangle were oriented in the opposite way, you would get

$$\int_{\boldsymbol{\gamma}} \mathbf{f}\cdot d\boldsymbol{\gamma} = \int_R (P_y - Q_x)\,dm_2$$

With this lemma, it is possible to prove Green's theorem and also give an analytic criterion which will distinguish between different orientations of a simple closed rectifiable curve. First here is a discussion which amounts to a computation.

Let Γ be a rectifiable simple closed curve with inside U_i and outside U_o. Let $\{R_k\}_{k=1}^{n_\delta}$ denote the non overlapping regions of Lemma 14.3.6 all oriented as explained there and let Γ also be oriented as explained there. It could be shown that all the regions contained in U_i have positive orientation but this will not be fussed over here. What can be said with no fussing is that since the shared edges have opposite orientations, all these interior regions are either oriented positively or they are all oriented negatively.

Let $\mathcal{B}_\delta$ be the set of border regions and let $\mathcal{I}_\delta$ be the rectangles contained in U_i. Thus in taking the sum of the line integrals over the boundaries of the interior rectangles, the integrals over the "interior edges" cancel out and you are left with a line integral over the exterior edges of a polygon which is composed of the union of the squares in $\mathcal{I}_\delta$.

Now let $\mathbf{f}(x,y) = (P(x,y), Q(x,y))$ be a vector field which is C^1 on U_i, and suppose also that both P_y and Q_x are in $L^1(U_i)$ (Absolutely integrable) and that P, Q are continuous on $U_i \cup \Gamma$. (An easy way to get all this to happen is to let P, Q be restrictions to $U_i \cup \Gamma$ of functions which are C^1 on some open set containing $U_i \cup \Gamma$.) Note that

$$\cup_{\delta>0}\{R : R \in \mathcal{I}_\delta\} = U_i$$

and that for

$$I_\delta \equiv \cup\{R : R \in \mathcal{I}_\delta\},$$

the following pointwise convergence holds.

$$\lim_{\delta\to 0} \mathcal{X}_{I_\delta}(\mathbf{x}) = \mathcal{X}_{U_i}(\mathbf{x}).$$

By the dominated convergence theorem,

$$\begin{aligned} \lim_{\delta\to 0}\int_{I_\delta} (Q_x - P_y)\,dm_2 &= \int_{U_i} (Q_x - P_y)\,dm_2 \\ \lim_{\delta\to 0}\int_{I_\delta} (P_y - Q_x)\,dm_2 &= \int_{U_i} (P_y - Q_x)\,dm_2 \end{aligned}$$

Let ∂R denote the boundary of R for R one of these regions of Lemma 14.3.6 oriented as described. Let $w_\delta(R)^2$ denote

$$\begin{aligned} &(\max\{Q(\mathbf{x}) : \mathbf{x} \in \partial R\} - \min\{Q(\mathbf{x}) : \mathbf{x} \in \partial R\})^2 \\ &+ (\max\{P(\mathbf{x}) : \mathbf{x} \in \partial R\} - \min\{P(\mathbf{x}) : \mathbf{x} \in \partial R\})^2 \end{aligned}$$

By uniform continuity of P, Q on the compact set $U_i \cup \Gamma$, if δ is small enough, $w_\delta(R) < \varepsilon$ for all $R \in \mathcal{B}_\delta$. Then for $R \in \mathcal{B}_\delta$, it follows from Theorem 14.2.4

$$\left|\int_{\partial R} \mathbf{f}\cdot d\boldsymbol{\gamma}\right| \le \frac{1}{2} w_\delta(R)\,(V(\partial R)) < \varepsilon\,(V(\partial R)) \tag{14.20}$$

whenever δ is small enough. Always let δ be this small.

Also since the line integrals cancel on shared edges

$$\sum_{R\in\mathcal{I}_\delta}\int_{\partial R}\mathbf{f}\cdot d\boldsymbol{\gamma}+\sum_{R\in\mathcal{B}_\delta}\int_{\partial R}\mathbf{f}\cdot d\boldsymbol{\gamma}=\int_\Gamma \mathbf{f}\cdot d\boldsymbol{\gamma} \tag{14.21}$$

Consider the second sum on the left. From 14.20

$$\left|\sum_{R\in\mathcal{B}_\delta}\int_{\partial R}\mathbf{f}\cdot d\boldsymbol{\gamma}\right|\leq\sum_{R\in\mathcal{B}_\delta}\left|\int_{\partial R}\mathbf{f}\cdot d\boldsymbol{\gamma}\right|\leq\varepsilon\sum_{R\in\mathcal{B}_\delta}\left(V\left(\partial R\right)\right)$$

Denote by Γ_R the part of Γ which is contained in $R\in\mathcal{B}_\delta$ and $V\left(\Gamma_R\right)$ is its length. Then the above sum equals

$$\varepsilon\left(\sum_{R\in\mathcal{B}_\delta}V\left(\Gamma_R\right)+B_\delta\right)=\varepsilon\left(V\left(\Gamma\right)+B_\delta\right)$$

where B_δ is the sum of the lengths of the straight edges. This is easy to estimate. Recall from 14.3.6 there are no more than

$$4\left(\frac{V\left(\Gamma\right)}{\delta}+1\right)$$

of these border regions. Furthermore, the sum of the lengths of all four edges of one of these is no more than 8δ and so

$$B_\delta\leq 4\left(\frac{V\left(\Gamma\right)}{\delta}+1\right)8\delta=32V\left(\Gamma\right)+32\delta.$$

Thus the absolute value of the second sum on the right in 14.21 is dominated by

$$\varepsilon\left(33V\left(\Gamma\right)+32\delta\right)$$

Since ε was arbitrary, this formula implies with Green's theorem proved above for squares,

$$\int_\Gamma\mathbf{f}\cdot d\boldsymbol{\gamma}=\lim_{\delta\to0}\sum_{R\in\mathcal{I}_\delta}\int_{\partial R}\mathbf{f}\cdot d\boldsymbol{\gamma}+\lim_{\delta\to0}\sum_{R\in\mathcal{B}_\delta}\int_{\partial R}\mathbf{f}\cdot d\boldsymbol{\gamma}$$

$$=\lim_{\delta\to0}\sum_{R\in\mathcal{I}_\delta}\int_{\partial R}\mathbf{f}\cdot d\boldsymbol{\gamma}=\lim_{\delta\to0}\int_{I_\delta}\pm\left(Q_x-P_y\right)dm_2=\int_{U_i}\pm\left(Q_x-P_y\right)dm_2$$

where the $\pm$ adusts for whether the interior rectangles are all oriented positively or all oriented negatively. ■

This has proved the general form of Green's theorem which is stated in the following theorem.

Theorem 14.3.8 *Let Γ be a rectifiable simple closed curve in $\mathbb{R}^2$ having inside U_i and outside U_o. Let P,Q be functions with the property that*

$$Q_x,P_y\in L^1\left(U_i\right)$$

and P,Q are C^1 on U_i. Assume also P,Q are continuous on $\Gamma\cup U_i$. Then there exists an orientation for Γ (Remember there are only two.) such that for

$$\mathbf{f}\left(x,y\right)=\left(P\left(x,y\right),Q\left(x,y\right)\right),$$

$$\int_\Gamma\mathbf{f}\cdot d\boldsymbol{\gamma}=\int_{U_i}\left(Q_x-P_y\right)dm_2.$$

Proof: In the construction of the regions, an orientation was imparted to Γ. The above computation shows

$$\int_\Gamma \mathbf{f}\cdot d\boldsymbol{\gamma} = \int_{U_i} \pm\left(Q_x - P_y\right) dm_2$$

If the area integral equals

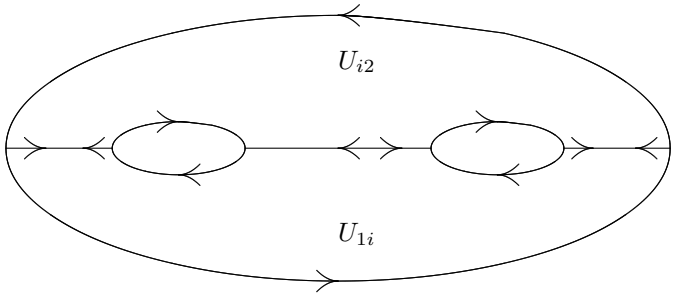
$$\int_{U_i} -\left(Q_x - P_y\right) dm_2,$$

just take the other orientation for Γ. This proves the theorem. ■

With this wonderful theorem, it is possible to give an analytic description of the two different orientations of a rectifiable simple closed curve. The positive orientation is the one for which Greens theorem holds and the other one, called the negative orientation is the one for which

$$\int_\Gamma \mathbf{f}\cdot d\boldsymbol{\gamma} = \int_{U_i} \left(P_y - Q_x\right) dm_2.$$

There are other regions for which Green's theorem holds besides just the inside and boundary of a simple closed curve. For Γ a simple closed curve and U_i its inside, lets refer to $U_i \cup \Gamma$ as a Jordan region. When you have two non overlapping Jordan regions which intersect in a finite number of simple curves, you can delete the interiors of these simple curves and what results will also be a region for which Green's theorem holds. This is illustrated in the following picture.

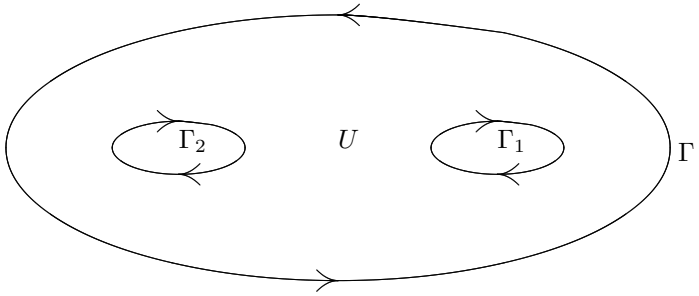


There are two Jordan regions here with insides U_{1i} and U_{2i} and these regions intersect in three simple curves. As indicated in the picture, opposite orientations are given to each of these three simple curves. Then the line integrals over these cancel. The area integrals add. Recall the two dimensional area of a bounded variation curve equals 0.

Denote by Γ the curve on the outside of the whole thing and Γ_1 and Γ_2 the oriented boundaries of the two holes which result when the curves of intersection are removed, the orientations as shown. Then letting $\mathbf{f}\left(x,y\right) = \left(P\left(x,y\right), Q\left(x,y\right)\right)$, and

$$U = U_{1i} \cup U_{2i} \cup \{\text{Open segments of intersection}\}$$

as shown in the following picture,



it follows from applying Green's theorem to both of the Jordan regions,

$$\begin{aligned}\int_{\Gamma}\mathbf{f}\cdot d\boldsymbol{\gamma}+\int_{\Gamma_1}\mathbf{f}\cdot d\boldsymbol{\gamma}_1+\int_{\Gamma_2}\mathbf{f}\cdot d\boldsymbol{\gamma}_2 &= \int_{U_{1i}\cup U_{2i}}(Q_x-P_y)\,dm_2\\ &= \int_{U}(Q_x-P_y)\,dm_2\end{aligned}$$

To make this simpler, just write it in the form

$$\int_{\partial U}\mathbf{f}\cdot d\boldsymbol{\gamma}=\int_{U}(Q_x-P_y)\,dm_2$$

where ∂U is oriented as indicated in the picture and involves the three oriented curves Γ,Γ_1,Γ_2.

14.4 Stoke's Theorem

Stokes theorem is usually presented in calculus courses under far more restrictive assumptions than will be used here. It turns out that all the hard questions are related to Green's theorem and that when you have the general version of Green's theorem this can be used to obtain a general version of Stoke's theorem using a simple identity. This is because Stoke's theorem is really just a three dimensional version of the two dimensional Green's theorem. This will be made more precise below.

To begin with suppose Γ is a rectifiable curve in $\mathbb{R}^2$ having parameterization $\boldsymbol{\alpha}:[a,b]\to\Gamma$ for $\boldsymbol{\alpha}$ a continuous function. Let $\mathbf{R}:U\to\mathbb{R}^n$ be a C^1 function where U contains $\boldsymbol{\alpha}^*$. Then one could define a curve

$$\boldsymbol{\gamma}(t)\equiv\mathbf{R}(\boldsymbol{\alpha}(t)),\ t\in[a,b].$$

Lemma 14.4.1 *The curve $\boldsymbol{\gamma}^*$ where $\boldsymbol{\gamma}$ is as just described is a rectifiable curve. If $\mathbf{F}$ is defined and continuous on $\boldsymbol{\gamma}^*$ then*

$$\int_{\boldsymbol{\gamma}}\mathbf{F}\cdot d\boldsymbol{\gamma}=\int_{\boldsymbol{\alpha}}((\mathbf{F}\circ\mathbf{R})\cdot\mathbf{R}_u,(\mathbf{F}\circ\mathbf{R})\cdot\mathbf{R}_v)\cdot d\boldsymbol{\alpha}$$

where $\mathbf{R}_u$ signifies the partial derivative of $\mathbf{R}$ with respect to the variable u.

Proof: Let

$$K\equiv\left\{\mathbf{y}\in\mathbb{R}^2:\operatorname{dist}(\mathbf{y},\boldsymbol{\alpha}^*)\le r\right\}$$

where r is small enough that $K\subseteq U$. This is easily done because $\boldsymbol{\alpha}^*$ is compact. Let

$$C_K\equiv\max\{||D\mathbf{R}(\mathbf{x})||:\mathbf{x}\in K\}$$

Consider

$$\sum_{j=0}^{n-1}|\mathbf{R}(\boldsymbol{\alpha}(t_{j+1}))-\mathbf{R}(\boldsymbol{\alpha}(t_j))| \tag{14.22}$$

where $\{t_0,\cdots,t_n\}$ is a partition of $[a,b]$. Since $\boldsymbol{\alpha}$ is continuous, there exists a δ such that if $||\mathcal{P}||<\delta$, then the segment

$$\{\boldsymbol{\alpha}(t_j)+s(\boldsymbol{\alpha}(t_{j+1})-\boldsymbol{\alpha}(t_j)):s\in[0,1]\}$$

is contained in K. Therefore, by the mean value inequality, Theorem 6.5.2,

$$\sum_{j=0}^{n-1}|\mathbf{R}(\boldsymbol{\alpha}(t_{j+1}))-\mathbf{R}(\boldsymbol{\alpha}(t_j))|\le\sum_{j=0}^{n-1}C_K|\boldsymbol{\alpha}(t_{j+1})-\boldsymbol{\alpha}(t_j)|$$

Now if $\mathcal{P}$ is any partition, 14.22 can always be made larger by adding in points to $\mathcal{P}$ till $||\mathcal{P}|| < \delta$ and so this shows

$$V(\boldsymbol{\gamma}, [a,b]) \leq C_K V(\boldsymbol{\alpha}, [a,b]).$$

This proves the first part.

Next consider the claim about the integral. Let

$$G(\mathbf{v}, \mathbf{x}) \equiv \mathbf{R}(\mathbf{x}+\mathbf{v}) - \mathbf{R}(\mathbf{x}) - D\mathbf{R}(\mathbf{x})(\mathbf{v}).$$

Then

$$D_1 G(\mathbf{v}, \mathbf{x}) = D\mathbf{R}(\mathbf{x}+\mathbf{v}) - D\mathbf{R}(\mathbf{x})$$

and so by uniform continuity of $D\mathbf{R}$ on the compact set K, it follows there exists $\delta > 0$ such that if $|\mathbf{v}| < \delta$, then for all $\mathbf{x} \in \boldsymbol{\alpha}^*$,

$$||D\mathbf{R}(\mathbf{x}+\mathbf{v}) - D\mathbf{R}(\mathbf{x})|| = ||D_1 G(\mathbf{v}, \mathbf{x})|| < \varepsilon.$$

By Theorem 6.5.2 again it follows that for all $\mathbf{x} \in \boldsymbol{\alpha}^*$ and $|\mathbf{v}| < \delta$,

$$|G(\mathbf{v}, \mathbf{x})| = |\mathbf{R}(\mathbf{x}+\mathbf{v}) - \mathbf{R}(\mathbf{x}) - D\mathbf{R}(\mathbf{x})(\mathbf{v})| \leq \varepsilon |\mathbf{v}| \tag{14.23}$$

Letting $||\mathcal{P}||$ be small enough, it follows from the continuity of $\boldsymbol{\alpha}$ that

$$|\boldsymbol{\alpha}(t_{j+1}) - \boldsymbol{\alpha}(t_j)| < \delta$$

Therefore for such $\mathcal{P}$,

$$\begin{aligned}
&\sum_{j=0}^{n-1} \mathbf{F}(\boldsymbol{\gamma}(t_j)) \cdot (\boldsymbol{\gamma}(t_{j+1}) - \boldsymbol{\gamma}(t_j)) \\
&= \sum_{j=0}^{n-1} \mathbf{F}(\mathbf{R}(\boldsymbol{\alpha}(t_j))) \cdot (\mathbf{R}(\boldsymbol{\alpha}(t_{j+1})) - \mathbf{R}(\boldsymbol{\alpha}(t_j))) \\
&= \sum_{j=0}^{n-1} \mathbf{F}(\mathbf{R}(\boldsymbol{\alpha}(t_j))) \cdot [D\mathbf{R}(\boldsymbol{\alpha}(t_j))(\boldsymbol{\alpha}(t_{j+1}) - \boldsymbol{\alpha}(t_j)) + \mathbf{o}(\boldsymbol{\alpha}(t_{j+1}) - \boldsymbol{\alpha}(t_j))]
\end{aligned}$$

where

$$\mathbf{o}(\boldsymbol{\alpha}(t_{j+1}) - \boldsymbol{\alpha}(t_j)) = \mathbf{R}(\boldsymbol{\alpha}(t_{j+1})) - \mathbf{R}(\boldsymbol{\alpha}(t_j)) - D\mathbf{R}(\boldsymbol{\alpha}(t_j))(\boldsymbol{\alpha}(t_{j+1}) - \boldsymbol{\alpha}(t_j))$$

and by 14.23,

$$|\mathbf{o}(\boldsymbol{\alpha}(t_{j+1}) - \boldsymbol{\alpha}(t_j))| < \varepsilon |\boldsymbol{\alpha}(t_{j+1}) - \boldsymbol{\alpha}(t_j)|$$

It follows

$$\left| \sum_{j=0}^{n-1} \mathbf{F}(\boldsymbol{\gamma}(t_j)) \cdot (\boldsymbol{\gamma}(t_{j+1}) - \boldsymbol{\gamma}(t_j)) - \sum_{j=0}^{n-1} \mathbf{F}(\mathbf{R}(\boldsymbol{\alpha}(t_j))) \cdot D\mathbf{R}(\boldsymbol{\alpha}(t_j))(\boldsymbol{\alpha}(t_{j+1}) - \boldsymbol{\alpha}(t_j)) \right| \tag{14.24}$$

$$\leq \sum_{j=0}^{n-1} |\mathbf{o}(\boldsymbol{\alpha}(t_{j+1}) - \boldsymbol{\alpha}(t_j))| \leq \sum_{j=0}^{n-1} \varepsilon |\boldsymbol{\alpha}(t_{j+1}) - \boldsymbol{\alpha}(t_j)| \leq \varepsilon V(\boldsymbol{\alpha}, [a,b])$$

Consider the second sum in 14.24. A term in the sum equals

$$\begin{aligned}
&\mathbf{F}(\mathbf{R}(\boldsymbol{\alpha}(t_j))) \cdot (\mathbf{R}_u(\boldsymbol{\alpha}(t_j))(\alpha_1(t_{j+1}) - \alpha_1(t_j)) + \mathbf{R}_v(\boldsymbol{\alpha}(t_j))(\alpha_2(t_{j+1}) - \alpha_2(t_j))) \\
&= (\mathbf{F}(\mathbf{R}(\boldsymbol{\alpha}(t_j))) \cdot \mathbf{R}_u(\boldsymbol{\alpha}(t_j)), \mathbf{F}(\mathbf{R}(\boldsymbol{\alpha}(t_j))) \cdot \mathbf{R}_v(\boldsymbol{\alpha}(t_j))) \cdot (\boldsymbol{\alpha}(t_{j+1}) - \boldsymbol{\alpha}(t_j))
\end{aligned}$$

By continuity of $\mathbf{F}$, $\mathbf{R}_u$ and $\mathbf{R}_v$, it follows that sum converges as $||\mathcal{P}|| \to 0$ to

$$\int_{\boldsymbol{\alpha}} ((\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_u, (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_v) \cdot d\boldsymbol{\alpha}$$

Therefore, taking the limit as $||\mathcal{P}|| \to 0$ in 14.24

$$\left| \int_{\boldsymbol{\gamma}} \mathbf{F} \cdot d\boldsymbol{\gamma} - \int_{\boldsymbol{\alpha}} ((\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_u, (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_v) \cdot d\boldsymbol{\alpha} \right| < \varepsilon V(\boldsymbol{\alpha}, [a,b]).$$

Since $\varepsilon > 0$ is arbitrary, This proves the lemma. ■

The following is a little identity which will allow a proof of Stoke's theorem to follow from Green's theorem. First recall the following definition from calculus of the curl of a vector field and the cross product of two vectors from calculus.

Definition 14.4.2 *Let $\mathbf{u} \equiv (a,b,c)$ and $\mathbf{v} \equiv (d,e,f)$ be two vectors in $\mathbb{R}^3$. Then*

$$\mathbf{u} \times \mathbf{v} \equiv \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a & b & c \\ d & e & f \end{vmatrix}$$

*where the determinant is expanded formally along the top row. Let $\mathbf{f} : U \to \mathbb{R}^3$ for $U \subseteq \mathbb{R}^3$ denote a vector field. The **curl** of the vector field yields another vector field and it is defined as follows.*

$$(\operatorname{curl}(\mathbf{f})(\mathbf{x}))_i \equiv (\nabla \times \mathbf{f}(\mathbf{x}))_i$$

where here ∂_j means the partial derivative with respect to x_j and the subscript of i in $(\operatorname{curl}(\mathbf{f})(\mathbf{x}))_i$ means the i^{th} Cartesian component of the vector, $\operatorname{curl}(\mathbf{f})(\mathbf{x})$. Thus the curl is evaluated by expanding the following determinant along the top row.

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ f_1(x,y,z) & f_2(x,y,z) & f_3(x,y,z) \end{vmatrix}.$$

Note the similarity with the cross product. More precisely and less evocatively,

$$\nabla \times \mathbf{f}(x,y,z) = \equiv \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z} \right) \mathbf{i} + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x} \right) \mathbf{j} + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) \mathbf{k}.$$

In the above, $\mathbf{i} = \mathbf{e}_1, \mathbf{j} = \mathbf{e}_2$, and $\mathbf{k} = \mathbf{e}_3$ the standard unit basis vectors for $\mathbb{R}^3$.

With this definition, here is the identity.

Lemma 14.4.3 *Let $\mathbf{R} : U \to V \subseteq \mathbb{R}^3$ where U is an open subset of $\mathbb{R}^2$ and V is an open subset of $\mathbb{R}^3$. Suppose $\mathbf{R}$ is C^2 and let $\mathbf{F}$ be a C^1 vector field defined in V.*

$$(\mathbf{R}_u \times \mathbf{R}_v) \cdot (\nabla \times \mathbf{F})(\mathbf{R}(u,v)) = ((\mathbf{F} \circ \mathbf{R})_u \cdot \mathbf{R}_v - (\mathbf{F} \circ \mathbf{R})_v \cdot \mathbf{R}_u)(u,v). \tag{14.25}$$

Proof: Letting x, y, z denote the components of $\mathbf{R}(\mathbf{u})$ and f_1, f_2, f_3 denote the components of $\mathbf{F}$, and letting a subscripted variable denote the partial derivative with respect to that variable, the left side of 14.25 equals

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x_u & y_u & z_u \\ x_v & y_v & z_v \end{vmatrix} \cdot \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \partial_x & \partial_y & \partial_z \\ f_1 & f_2 & f_3 \end{vmatrix}$$

$$= (f_{3y} - f_{2z})(y_u z_v - z_u y_v) + (f_{1z} - f_{3x})(z_u x_v - x_u z_v) + (f_{2x} - f_{1y})(x_u y_v - y_u x_v)$$

$$\begin{array}{c} = f_{3y} y_u z_v + f_{2z} z_u y_v + f_{1z} z_u x_v + f_{3x} x_u z_v + f_{2x} x_u y_v + f_{1y} y_u x_v \\ - (f_{2z} y_u z_v + f_{3y} z_u y_v + f_{1z} x_u z_v + f_{3x} z_u x_v + f_{2x} y_u x_v + f_{1y} x_u y_v) \end{array}$$

$$\begin{array}{c} = f_{1y} y_u x_v + f_{1z} z_u x_v + f_{2x} x_u y_v + f_{2z} z_u y_v + f_{3x} x_u z_v + f_{3y} y_u z_v \\ - (f_{1y} y_v x_u + f_{1z} z_v x_u + f_{2x} x_v y_u + f_{2z} z_v y_u + f_{3x} x_v z_u + f_{3y} y_v z_u) \end{array}$$

At this point add in and subtract off certain terms. Then the above equals

$$\begin{array}{c} = f_{1x} x_u x_v + f_{1y} y_u x_v + f_{1z} z_u x_v + f_{2x} x_u y_v + f_{2x} y_u y_v \\ + f_{2z} z_u y_v + f_{3x} x_u z_v + f_{3y} y_u z_v + f_{3z} z_u z_v \\ - \left(\begin{array}{c} f_{1x} x_v x_u + f_{1y} y_v x_u + f_{1z} z_v x_u + f_{2x} x_v y_u + f_{2x} y_v y_u \\ + f_{2z} z_v y_u + f_{3x} x_v z_u + f_{3y} y_v z_u + f_{3z} z_v z_u \end{array} \right) \end{array}$$

$$\begin{array}{c} = \dfrac{\partial f_1 \circ \mathbf{R}(u,v)}{\partial u} x_v + \dfrac{\partial f_2 \circ \mathbf{R}(u,v)}{\partial u} y_v + \dfrac{\partial f_3 \circ \mathbf{R}(u,v)}{\partial u} z_v \\ - \left(\dfrac{\partial f_1 \circ \mathbf{R}(u,v)}{\partial v} x_u + \dfrac{\partial f_2 \circ \mathbf{R}(u,v)}{\partial v} y_u + \dfrac{\partial f_3 \circ \mathbf{R}(u,v)}{\partial v} z_u \right) \end{array}$$

$$= ((\mathbf{F} \circ \mathbf{R})_u \cdot \mathbf{R}_v - (\mathbf{F} \circ \mathbf{R})_v \cdot \mathbf{R}_u)(u,v).$$

This proves the lemma. ■

Let U be a region in $\mathbb{R}^2$ for which Green's theorem holds. Thus Green's theorem says that for P, Q continuous on $U_i \cup \Gamma, P_v, Q_u \in L^1(U_i \cup \Gamma), P, Q$ being C^1 on U_i,

$$\int_U (Q_u - P_v)\, dm_2 = \int_{\partial U} \mathbf{f} \cdot d\boldsymbol{\alpha}$$

where ∂U consists of some simple closed rectifiable oriented curves as explained above. Here the u and v axes are in the same relation as the x and y axes.

Theorem 14.4.4 *(Stoke's Theorem) Let U be any region in $\mathbb{R}^2$ for which the conclusion of Green's theorem holds. Let $\mathbf{R} \in C^2(\overline{U}, \mathbb{R}^3)$ be a one to one function. Let*

$$\boldsymbol{\gamma}_j = \mathbf{R} \circ \boldsymbol{\alpha}_j,$$

where the $\boldsymbol{\alpha}_j$ are parameterizations for the oriented curves making up the boundary of U such that the conclusion of Green's theorem holds. Let S denote the surface,

$$S \equiv \{\mathbf{R}(u,v) : (u,v) \in U\},$$

Then for $\mathbf{F}$ a C^1 vector field defined near S,

$$\sum_{i=1}^{n} \int_{\boldsymbol{\gamma}_i} \mathbf{F} \cdot d\boldsymbol{\gamma}_i = \int_U (\mathbf{R}_u(u,v) \times \mathbf{R}_v(u,v)) \cdot (\nabla \times \mathbf{F}(\mathbf{R}(u,v)))\, dm_2$$

Proof: By Lemma 14.4.1,

$$\sum_{j=1}^{n} \int_{\boldsymbol{\gamma}_j} \mathbf{F} \cdot d\boldsymbol{\gamma}_j =$$

$$\sum_{j=1}^{n} \int_{\boldsymbol{\alpha}_j} ((\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_u, (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_v) \cdot d\boldsymbol{\alpha}_j$$

By the assumption that the conclusion of Green's theorem holds for U, this equals

$$\begin{aligned}
&\int_U \left[\left(\left(\mathbf{F}\circ\mathbf{R}\right)\cdot\mathbf{R}_v\right)_u - \left(\left(\mathbf{F}\circ\mathbf{R}\right)\cdot\mathbf{R}_u\right)_v\right] dm_2 \\
= &\int_U \left[\left(\mathbf{F}\circ\mathbf{R}\right)_u\cdot\mathbf{R}_v + \left(\mathbf{F}\circ\mathbf{R}\right)\cdot\mathbf{R}_{vu} - \left(\mathbf{F}\circ\mathbf{R}\right)\cdot\mathbf{R}_{uv} - \left(\mathbf{F}\circ\mathbf{R}\right)_v\cdot\mathbf{R}_u\right] dm_2 \\
= &\int_U \left[\left(\mathbf{F}\circ\mathbf{R}\right)_u\cdot\mathbf{R}_v - \left(\mathbf{F}\circ\mathbf{R}\right)_v\cdot\mathbf{R}_u\right] dm_2
\end{aligned}$$

the last step holding by equality of mixed partial derivatives, a result of the assumption that $\mathbf{R}$ is C^2. Now by Lemma 14.4.3, this equals

$$\int_U \left(\mathbf{R}_u\left(u,v\right)\times\mathbf{R}_v\left(u,v\right)\right)\cdot\left(\nabla\times\mathbf{F}\left(\mathbf{R}\left(u,v\right)\right)\right) dm_2$$

This proves Stoke's theorem. ■

With approximation arguments one can remove the assumption that $\mathbf{R}$ is C^2 and replace this condition with weaker conditions. This is not surprising because in the final result, only first derivatives of $\mathbf{R}$ occur.

14.5 Interpretation And Review

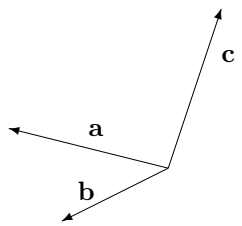
To understand the interpretation of Stoke's theorem in terms of an integral over the surface S, it is necessary to either do more theoretical development or to review some beginning calculus. I will do the latter here. First of all, it is important to understand the geometrical properties of the cross product. Those who have had a typical calculus course will probably not have seen this so I will present it here. It is elementary material which is a little out of place in an advanced calculus book but it is nevertheless useful and important and if you have not seen it, you should.

14.5.1 The Geometric Description Of The Cross Product

The cross product is a way of multiplying two vectors in $\mathbb{R}^3$. It is very different from the dot product in many ways. First the geometric meaning is discussed and then a description in terms of coordinates is given. Both descriptions of the cross product are important. The geometric description is essential in order to understand the applications to physics and geometry while the coordinate description is the only way to practically compute the cross product. In this presentation a vector is something which is characterized by direction and magnitude.

Definition 14.5.1 *Three vectors,* $\mathbf{a},\mathbf{b},\mathbf{c}$ *form a right handed system if when you extend the fingers of your right hand along the vector,* $\mathbf{a}$ *and close them in the direction of* $\mathbf{b}$*, the thumb points roughly in the direction of* $\mathbf{c}$.

For an example of a right handed system of vectors, see the following picture.



In this picture the vector **c** points upwards from the plane determined by the other two vectors. You should consider how a right hand system would differ from a left hand system. Try using your left hand and you will see that the vector, **c** would need to point in the opposite direction as it would for a right hand system.

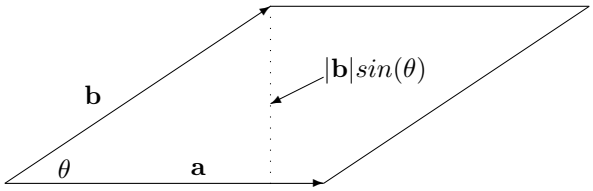
From now on, the vectors, $\mathbf{i}, \mathbf{j}, \mathbf{k}$ will **always** form a right handed system. To repeat, if you extend the fingers of your right hand along **i** and close them in the direction **j**, the thumb points in the direction of **k**. Recall these are the basis vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$.

The following is the geometric description of the cross product. It gives both the direction and the magnitude and therefore specifies the vector.

Definition 14.5.2 *Let* $\mathbf{a}$ *and* $\mathbf{b}$ *be two vectors in* $\mathbb{R}^3$. *Then* $\mathbf{a} \times \mathbf{b}$ *is defined by the following two rules.*

1. $|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}|\,|\mathbf{b}| \sin\theta$ *where* θ *is the included angle.*
2. $\mathbf{a} \times \mathbf{b} \cdot \mathbf{a} = 0$, $\mathbf{a} \times \mathbf{b} \cdot \mathbf{b} = 0$, *and* $\mathbf{a}, \mathbf{b}, \mathbf{a} \times \mathbf{b}$ *forms a right hand system.*

Note that $|\mathbf{a} \times \mathbf{b}|$ is the **area of the parallelogram** spanned by **a** and **b**.



The cross product satisfies the following properties.

$$\mathbf{a} \times \mathbf{b} = -(\mathbf{b} \times \mathbf{a}) \ , \ \mathbf{a} \times \mathbf{a} = \mathbf{0}, \tag{14.26}$$

For α a scalar,

$$(\alpha\mathbf{a}) \times \mathbf{b} = \alpha(\mathbf{a} \times \mathbf{b}) = \mathbf{a} \times (\alpha\mathbf{b}), \tag{14.27}$$

For $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$ vectors, one obtains the distributive laws,

$$\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}, \tag{14.28}$$

$$(\mathbf{b} + \mathbf{c}) \times \mathbf{a} = \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}. \tag{14.29}$$

Formula 14.26 follows immediately from the definition. The vectors $\mathbf{a} \times \mathbf{b}$ and $\mathbf{b} \times \mathbf{a}$ have the same magnitude, $|\mathbf{a}|\,|\mathbf{b}| \sin\theta$, and an application of the right hand rule shows they have opposite direction. Formula 14.27 is also fairly clear. If α is a nonnegative scalar, the direction of $(\alpha\mathbf{a}) \times \mathbf{b}$ is the same as the direction of $\mathbf{a} \times \mathbf{b}$,$\alpha(\mathbf{a} \times \mathbf{b})$ and $\mathbf{a} \times (\alpha\mathbf{b})$ while the magnitude is just α times the magnitude of $\mathbf{a} \times \mathbf{b}$ which is the same as the magnitude of $\alpha(\mathbf{a} \times \mathbf{b})$ and $\mathbf{a} \times (\alpha\mathbf{b})$. Using this yields equality in 14.27. In the case where $\alpha < 0$, everything works the same way except the vectors are all pointing in the opposite direction and you must multiply by $|\alpha|$ when comparing their magnitudes. The distributive laws are much harder to establish but the second follows from the first quite easily. Thus, assuming the first, and using 14.26,

$$\begin{aligned}(\mathbf{b} + \mathbf{c}) \times \mathbf{a} &= -\mathbf{a} \times (\mathbf{b} + \mathbf{c}) \\ &= -(\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}) \\ &= \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}.\end{aligned}$$

To verify the distributive law one can consider something called the box product.

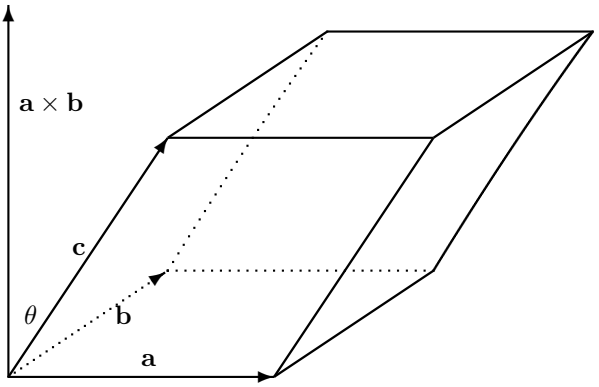
14.5.2 The Box Product, Triple Product

Definition 14.5.3 *A parallelepiped determined by the three vectors,* $\mathbf{a}, \mathbf{b}$, *and* $\mathbf{c}$ *consists of*

$$\{r\mathbf{a}+s\mathbf{b}+t\mathbf{c} : r, s, t \in [0,1]\}.$$

That is, if you pick three numbers, r, s, *and* t *each in* $[0,1]$ *and form* $r\mathbf{a}+s\mathbf{b}+t\mathbf{c}$, *then the collection of all such points is what is meant by the parallelepiped determined by these three vectors.*

The following is a picture of such a thing.



You notice the area of the base of the parallelepiped, the parallelogram determined by the vectors, $\mathbf{a}$ and $\mathbf{b}$ has area equal to $|\mathbf{a}\times\mathbf{b}|$ while the altitude of the parallelepiped is $|\mathbf{c}|\cos\theta$ where θ is the angle shown in the picture between $\mathbf{c}$ and $\mathbf{a}\times\mathbf{b}$. Therefore, the volume of this parallelepiped is the area of the base times the altitude which is just

$$|\mathbf{a}\times\mathbf{b}|\,|\mathbf{c}|\cos\theta = \mathbf{a}\times\mathbf{b}\cdot\mathbf{c}.$$

This expression is known as the box product and is sometimes written as $[\mathbf{a},\mathbf{b},\mathbf{c}]$. You should consider what happens if you interchange the $\mathbf{b}$ with the $\mathbf{c}$ or the $\mathbf{a}$ with the $\mathbf{c}$. You can see geometrically from drawing pictures that this merely introduces a minus sign. In any case the box product of three vectors always equals either the volume of the parallelepiped determined by the three vectors or else minus this volume. From geometric reasoning like this you see that

$$\mathbf{a}\cdot\mathbf{b}\times\mathbf{c} = \mathbf{a}\times\mathbf{b}\cdot\mathbf{c}.$$

In other words, you can switch the $\times$ and the $\cdot$.

14.5.3 A Proof Of The Distributive Law For The Cross Product

Here is a proof of the distributive law for the cross product. Let $\mathbf{x}$ be a vector. From the above observation,

$$\begin{aligned}\mathbf{x}\cdot\mathbf{a}\times(\mathbf{b}+\mathbf{c}) &= (\mathbf{x}\times\mathbf{a})\cdot(\mathbf{b}+\mathbf{c})\\ &= (\mathbf{x}\times\mathbf{a})\cdot\mathbf{b}+(\mathbf{x}\times\mathbf{a})\cdot\mathbf{c}\\ &= \mathbf{x}\cdot\mathbf{a}\times\mathbf{b}+\mathbf{x}\cdot\mathbf{a}\times\mathbf{c}\\ &= \mathbf{x}\cdot(\mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c}).\end{aligned}$$

Therefore,

$$\mathbf{x}\cdot[\mathbf{a}\times(\mathbf{b}+\mathbf{c})-(\mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c})] = 0$$

for all $\mathbf{x}$. In particular, this holds for $\mathbf{x} = \mathbf{a}\times(\mathbf{b}+\mathbf{c})-(\mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c})$ showing that $\mathbf{a}\times(\mathbf{b}+\mathbf{c}) = \mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c}$ and this proves the distributive law for the cross product.

14.5.4 The Coordinate Description Of The Cross Product

Now from the properties of the cross product and its definition,

$$\begin{array}{ll} \mathbf{i}\times\mathbf{j}=\mathbf{k} & \mathbf{j}\times\mathbf{i}=-\mathbf{k} \\ \mathbf{k}\times\mathbf{i}=\mathbf{j} & \mathbf{i}\times\mathbf{k}=-\mathbf{j} \\ \mathbf{j}\times\mathbf{k}=\mathbf{i} & \mathbf{k}\times\mathbf{j}=-\mathbf{i} \end{array}$$

With this information, the following gives the coordinate description of the cross product.

Proposition 14.5.4 *Let* $\mathbf{a}=a_1\mathbf{i}+a_2\mathbf{j}+a_3\mathbf{k}$ *and* $\mathbf{b}=b_1\mathbf{i}+b_2\mathbf{j}+b_3\mathbf{k}$ *be two vectors. Then*

$$\begin{aligned}\mathbf{a}\times\mathbf{b} &= (a_2b_3-a_3b_2)\,\mathbf{i}+(a_3b_1-a_1b_3)\,\mathbf{j}+ \\ &\quad +(a_1b_2-a_2b_1)\,\mathbf{k}. \end{aligned} \tag{14.30}$$

Proof: From the above table and the properties of the cross product listed,

$$\begin{aligned}&(a_1\mathbf{i}+a_2\mathbf{j}+a_3\mathbf{k})\times(b_1\mathbf{i}+b_2\mathbf{j}+b_3\mathbf{k})= \\ &a_1b_2\mathbf{i}\times\mathbf{j}+a_1b_3\mathbf{i}\times\mathbf{k}+a_2b_1\mathbf{j}\times\mathbf{i}+a_2b_3\mathbf{j}\times\mathbf{k}+ \\ &\quad +a_3b_1\mathbf{k}\times\mathbf{i}+a_3b_2\mathbf{k}\times\mathbf{j} \\ &=a_1b_2\mathbf{k}-a_1b_3\mathbf{j}-a_2b_1\mathbf{k}+a_2b_3\mathbf{i}+a_3b_1\mathbf{j}-a_3b_2\mathbf{i} \\ &=(a_2b_3-a_3b_2)\,\mathbf{i}+(a_3b_1-a_1b_3)\,\mathbf{j}+(a_1b_2-a_2b_1)\,\mathbf{k} \end{aligned} \tag{14.31}$$

This proves the proposition. ■

The easy way to remember the above formula is to write it as follows.

$$\mathbf{a}\times\mathbf{b}=\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} \tag{14.32}$$

where you expand the determinant along the top row. This yields

$$(a_2b_3-a_3b_2)\,\mathbf{i}-(a_1b_3-a_3b_1)\,\mathbf{j}+(a_1b_2-a_2b_1)\,\mathbf{k} \tag{14.33}$$

which is the same as 14.31.

14.5.5 The Integral Over A Two Dimensional Surface

First it is good to define what is meant by a smooth surface.

Definition 14.5.5 *Let* S *be a subset of* $\mathbb{R}^3$. *Then* S *is a* ***smooth surface*** *if there exists an open set,* $U\subseteq\mathbb{R}^2$ *and a* C^1 *function,* $\mathbf{R}$ *defined on* U *such that* $\mathbf{R}(U)=S$, $\mathbf{R}$ *is one to one, and for all* $(u,v)\in U$,

$$\mathbf{R}_u\times\mathbf{R}_v\neq\mathbf{0}. \tag{14.34}$$

This last condition ensures that there is always a well defined normal on S. *This function,* $\mathbf{R}$ *is called a parameterization of the surface. It is just like a parameterization of a curve but here there are two parameters,* u,v.

One way to think of this is that there is a piece of rubber occupying U in the plane and then it is taken and stretched in three dimensions. This gives S.

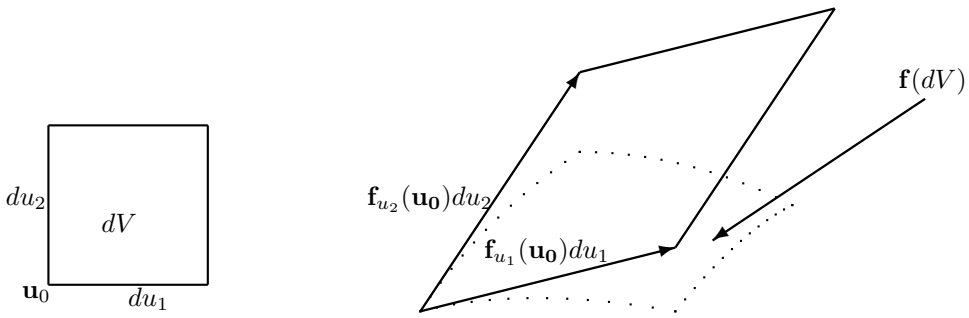
Definition 14.5.6 *Let $\mathbf{u}_1, \mathbf{u}_2$ be vectors in $\mathbb{R}^3$. The 2 dimensional parallelogram determined by these vectors will be denoted by $P(\mathbf{u}_1, \mathbf{u}_2)$ and it is defined as*

$$P(\mathbf{u}_1, \mathbf{u}_2) \equiv \left\{ \sum_{j=1}^{2} s_j \mathbf{u}_j : s_j \in [0,1] \right\}.$$

Then the area of this parallelogram is

$$\text{area } P(\mathbf{u}_1, \mathbf{u}_2) \equiv |\mathbf{u}_1 \times \mathbf{u}_2|.$$

Suppose then that $\mathbf{x} = \mathbf{R}(\mathbf{u})$ where $\mathbf{u} \in U$, a subset of $\mathbb{R}^2$ and $\mathbf{x}$ is a point in V, a subset of 3 dimensional space. Thus, letting the Cartesian coordinates of $\mathbf{x}$ be given by $\mathbf{x} = (x_1, x_2, x_3)^T$, each x_i being a function of $\mathbf{u}$, an infinitesimal rectangle located at $\mathbf{u}_0$ corresponds to an infinitesimal parallelogram located at $\mathbf{R}(\mathbf{u}_0)$ which is determined by the 2 vectors $\left\{ \frac{\partial \mathbf{R}(\mathbf{u}_0)}{\partial u}\, du, \frac{\partial \mathbf{R}(\mathbf{u}_0)}{\partial v}\, dv \right\}$, each of which is tangent to the surface defined by $\mathbf{x} = \mathbf{R}(\mathbf{u})$. This is a very vague and unacceptable description. What exactly is an infinitesimal rectangle? However, it can all be made precise later and this is good motivation for the real thing.



From Definition 14.5.6, the volume of this infinitesimal parallelepiped located at $\mathbf{R}(\mathbf{u}_0)$ is given by

$$\left| \frac{\partial \mathbf{R}(\mathbf{u}_0)}{\partial u}\, du \times \frac{\partial \mathbf{R}(\mathbf{u}_0)}{\partial v}\, dv \right| = \left| \frac{\partial \mathbf{R}(\mathbf{u}_0)}{\partial u} \times \frac{\partial \mathbf{R}(\mathbf{u}_0)}{\partial v} \right| du dv \tag{14.35}$$

$$= |\mathbf{R}_u \times \mathbf{R}_v|\, du dv \tag{14.36}$$

This motivates the following definition of what is meant by the integral over a parametrically defined surface in $\mathbb{R}^3$.

Definition 14.5.7 *Suppose U is a subset of $\mathbb{R}^2$ and suppose $\mathbf{R} : U \to \mathbf{R}(U) = S \subseteq \mathbb{R}^3$ is a one to one and C^1 function. Then if $h : \mathbf{R}(U) \to \mathbb{R}$, define the 2 dimensional surface integral, $\int_{\mathbf{R}(U)} h(\mathbf{x})\, dS$ according to the following formula.*

$$\int_{\mathbf{S}} h(\mathbf{x})\, dS \equiv \int_U h(\mathbf{R}(\mathbf{u}))\, |\mathbf{R}_u(\mathbf{u}) \times \mathbf{R}_v(\mathbf{u})|\, du dv.$$

With this understanding, it becomes possible to interpret the meaning of Stoke's theorem. This is stated in the following theorem. Note that slightly more is assumed here than earlier. In particular, it is assumed that $\mathbf{R}_u \times \mathbf{R}_v \neq \mathbf{0}$. This allows the definition of a well defined normal vector which varies continuously over the surface, S.

Theorem 14.5.8 *(Stoke's Theorem) Let U be any region in $\mathbb{R}^2$ for which the conclusion of Green's theorem holds. Let $\mathbf{R} \in C^2\left(\overline{U_i}, \mathbb{R}^3\right)$ be a one to one function such that $\mathbf{R}_u \times \mathbf{R}_v \neq \mathbf{0}$ on U. Let*

$$\boldsymbol{\gamma}_j = \mathbf{R} \circ \boldsymbol{\alpha}_j,$$

where the $\boldsymbol{\alpha}_j$ are parameterizations for the oriented bounded variation curves bounding the region U oriented such that the conclusion of Green's theorem holds. Let S denote the surface,

$$S \equiv \{\mathbf{R}(u,v) : (u,v) \in U\},$$

Then for $\mathbf{F}$ a C^1 vector field defined near S,

$$\sum_{j=1}^{n} \int_{\boldsymbol{\gamma}_j} \mathbf{F} \cdot d\boldsymbol{\gamma}_i = \int_U (\mathbf{R}_u \times \mathbf{R}_v) \cdot (\nabla \times \mathbf{F})(\mathbf{R}(u,v))\, dm_2 \tag{14.37}$$

$$= \int_S (\nabla \times \mathbf{F}) \cdot \mathbf{n} dS \tag{14.38}$$

Proof: Formula 14.37 was established in Theorem 14.4.4. The unit normal of the point $\mathbf{R}(u,v)$ of S is $(\mathbf{R}_u \times \mathbf{R}_v) / |\mathbf{R}_u \times \mathbf{R}_v|$ and from the definition of the integral over the surface, Definition 14.5.7, Formula 14.38 follows. ■

14.6 Introduction To Complex Analysis

14.6.1 Basic Theorems, The Cauchy Riemann Equations

With Green's theorem and the technique of proof used in proving it, it is possible to present the most important parts of complex analysis almost effortlessly. I will do this here and leave some of the other parts for the exercises. Recall the complex numbers should be considered as points in the plane. Thus a complex number is of the form $x + iy$ where $i^2 = -1$. The complex conjugate is defined by

$$\overline{x + iy} \equiv x - iy$$

and for z a complex number,

$$|z| \equiv (z\overline{z})^{1/2} = \sqrt{x^2 + y^2}.$$

Thus when $x + iy$ is considered an ordered pair $(x, y) \in \mathbb{R}^2$ the magnitude of a complex number is nothing more than the usual norm of the ordered pair. Also for $z = x+iy, w = u + iv$,

$$|z - w| = \sqrt{(x-u)^2 + (y-v)^2}$$

so in terms of all topological considerations, $\mathbb{R}^2$ is the same as $\mathbb{C}$. Thus to say $z \to f(z)$ is continuous, is the same as saying

$$(x,y) \to u(x,y),\ (x,y) \to v(x,y)$$

are continuous where $f(z) \equiv u(x,y) + iv(x,y)$ with u and v being called the real and imaginary parts of f. The only new thing is that writing an ordered pair (x, y) as $x + iy$ with the convention $i^2 = -1$ makes $\mathbb{C}$ into a field. Now here is the definition of what it means for a function to be analytic.

Definition 14.6.1 *Let U be an open subset of $\mathbb{C}$ $(\mathbb{R}^2)$ and let $f : U \to \mathbb{C}$ be a function. Then f is said to be analytic on U if for every $z \in U$,*

$$\lim_{\Delta z \to 0} \frac{f(z+\Delta z) - f(z)}{\Delta z} \equiv f'(z)$$

exists and is a continuous function of $z \in U$. For a function having values in $\mathbb{C}$ denote by $u(x,y)$ the real part of f and $v(x,y)$ the imaginary part. Both u and v have real values and

$$f(x+iy) \equiv f(z) \equiv u(x,y) + iv(x,y)$$

Proposition 14.6.2 *Let U be an open subset of $\mathbb{C}$. Then $f : U \to \mathbb{C}$ is analytic if and only if for*

$$f(x+iy) \equiv u(x,y) + iv(x,y)$$

$u(x,y), v(x,y)$ being the real and imaginary parts of f, it follows

$$u_x(x,y) = v_y(x,y),\ u_y(x,y) = -v_x(x,y)$$

and all these partial derivatives, u_x, u_y, v_x, v_y are continuous on U. (The above equations are called the Cauchy Riemann equations.)

Proof: First suppose f is analytic. First let $\Delta z = ih$ and take the limit of the difference quotient as $h \to 0$ in the definition. Thus from the definition,

$$\begin{aligned} f'(z) &\equiv \lim_{h\to 0} \frac{f(z+ih) - f(z)}{ih} \\ &= \lim_{h\to 0} \frac{u(x,y+h) + iv(x,y+h) - (u(x,y) + iv(x,y))}{ih} \\ &= \lim_{h\to 0} \frac{1}{i}(u_y(x,y) + iv_y(x,y)) = -iu_y(x,y) + v_y(x,y) \end{aligned}$$

Next let $\Delta z = h$ and take the limit of the difference quotient as $h \to 0$.

$$\begin{aligned} f'(z) &\equiv \lim_{h\to 0} \frac{f(z+h) - f(z)}{h} \\ &= \lim_{h\to 0} \frac{u(x+h,y) + iv(x+h,y) - (u(x,y) + iv(x,y))}{h} \\ &= u_x(x,y) + iv_x(x,y). \end{aligned}$$

Therefore, equating real and imaginary parts,

$$u_x = v_y,\ v_x = -u_y$$

and this yields the Cauchy Riemann equations. Since $z \to f'(z)$ is continuous, it follows the real and imaginary parts of this function must also be continuous. Thus from the above formulas for $f'(z)$, it follows from the continuity of $z \to f'(z)$ all the partial derivatives of the real and imaginary parts are continuous.

Next suppose the Cauchy Riemann equations hold and these partial derivatives are all continuous. For $\Delta z = h + ik$,

$$f(z+\Delta z) - f(z) = u(x+h,y+k) + iv(x+h,y+k) - (u(x,y) + iv(x,y))$$

$$= u_x(x,y)h + u_y(x,y)k + i(v_x(x,y)h + v_y(x,y)k) + o((h,k))$$

$$= u_x(x,y)h + u_y(x,y)k + i(v_x(x,y)h + v_y(x,y)k) + o(\Delta z)$$

This follows from Theorem 6.6.1 which says that C^1 implies differentiable along with the definition of the norm (absolute value) in $\mathbb{C}$. By the Cauchy Riemann equations this equals

$$\begin{aligned} &= u_x(x,y)\,h - v_x(x,y)\,k + i\left(v_x(x,y)\,h + u_x(x,y)\,k\right) + o(\Delta z) \\ &= u_x(x,y)\,(h+ik) + iv_x(x,y)\,(h+ik) + o(\Delta z) \\ &= u_x(x,y)\,\Delta z + iv_x(x,y)\,\Delta z + o(\Delta z) \end{aligned}$$

Dividing by Δz and taking a limit yields $f'(z)$ exists and equals $u_x(x,y) + iv_x(x,y)$ which are assumed to be continuous. This proves the proposition. ■

14.6.2 Contour Integrals

The most important tools in complex analysis are Cauchy's theorem in some form and Cauchy's formula for an analytic function. I will give one of the very best versions of these theorems. They all involve something called a contour integral. Now a contour integral is just a sort of line integral. Here is the definition.

Definition 14.6.3 *Let $\gamma : [a,b] \to \mathbb{C}$ be of bounded variation and let $f : \gamma^* \to \mathbb{C}$. Letting $\mathcal{P} \equiv \{t_0, \cdots, t_n\}$ where $a = t_0 < t_1 < \cdots < t_n = b$, define*

$$||\mathcal{P}|| \equiv \max\{|t_j - t_{j-1}| : j = 1, \cdots, n\}$$

and the Riemann Stieltjes sum by

$$S(\mathcal{P}) \equiv \sum_{j=1}^{n} f(\gamma(\tau_j))\,(\gamma(t_j) - \gamma(t_{j-1}))$$

where $\tau_j \in [t_{j-1}, t_j]$. (Note this notation is a little sloppy because it does not identify the specific point, τ_j used. It is understood that this point is arbitrary.) Define $\int_\gamma f(z)\,dz$ as the unique number which satisfies the following condition. For all $\varepsilon > 0$ there exists a $\delta > 0$ such that if $||\mathcal{P}|| \le \delta$, then

$$\left| \int_\gamma f(z)\,dz - S(\mathcal{P}) \right| < \varepsilon.$$

Sometimes this is written as

$$\int_\gamma f(z)\,dz \equiv \lim_{||\mathcal{P}|| \to 0} S(\mathcal{P}).$$

You note that this is essentially the same definition given earlier for the line integral only this time the function has values in $\mathbb{C}$ rather than $\mathbb{R}^n$ and there is no dot product involved. Instead, you multiply by the complex number $\gamma(t_j) - \gamma(t_{j-1})$ in the Riemann Stieltjes sum. To tie this in with the line integral even more, consider a typical term in the sum for $S(\mathcal{P})$. Let $\gamma(t) = \gamma_1(t) + i\gamma_2(t)$. Then letting u be the real part of f and v the imaginary part, $S(\mathcal{P})$ equals

$$\begin{aligned} &\sum_{j=1}^{n} \left(u(\gamma_1(\tau_j), \gamma_2(\tau_j)) + iv(\gamma_1(\tau_j), \gamma_2(\tau_j))\right) \\ &(\gamma_1(t_j) - \gamma_1(t_{j-1}) + i(\gamma_2(t_j) - \gamma_2(t_{j-1}))) \end{aligned}$$

$$\begin{aligned} = & \sum_{j=1}^{n} u(\gamma_1(\tau_j), \gamma_2(\tau_j))(\gamma_1(t_j) - \gamma_1(t_{j-1})) \\ & - \sum_{j=1}^{n} v(\gamma_1(\tau_j), \gamma_2(\tau_j))(\gamma_2(t_j) - \gamma_2(t_{j-1})) \\ & + i \sum_{j=1}^{n} v(\gamma_1(\tau_j), \gamma_2(\tau_j))(\gamma_1(t_j) - \gamma_1(t_{j-1})) \\ & + i \sum_{j=1}^{n} u(\gamma_1(\tau_j), \gamma_2(\tau_j))(\gamma_2(t_j) - \gamma_2(t_{j-1})) \end{aligned}$$

Combining these leads to

$$\begin{aligned} & \sum_{j=1}^{n} (u(\gamma(\tau_j)), -v(\gamma(\tau_j))) \cdot (\gamma(t_j) - \gamma(t_{j-1})) \\ & + i \sum_{j=1}^{n} (v(\gamma(\tau_j)), u(\gamma(\tau_j))) \cdot (\gamma(t_j) - \gamma(t_{j-1})) \end{aligned} \tag{14.39}$$

Since the functions u and v are continuous, the limit as $||\mathcal{P}|| \to 0$ of the above equals

$$\int_\gamma (u, -v) \cdot d\gamma + i \int_\gamma (v, u) \cdot d\gamma$$

■

This proves most of the following lemma.

Lemma 14.6.4 *Let Γ be a rectifiable curve in $\mathbb{C}$ having parameterization γ which is continuous with bounded variation. Also let $f : \Gamma \to \mathbb{C}$ be continuous. Then the contour integral $\int_\gamma f(z)\, dz$ exists and is given by the sum of the following line integrals.*

$$\int_\gamma f(z)\, dz = \int_\gamma (u, -v) \cdot d\gamma + i \int_\gamma (v, u) \cdot d\gamma \tag{14.40}$$

Proof: The existence of the two line integrals as limits of $S(\mathcal{P})$ as $||\mathcal{P}|| \to 0$ follows from continuity of u, v and Theorem 14.2.3 along with the above discussion which decomposes the sum for the contour integral into the expression of 14.39 for which the two sums converge to the line integrals in the above formula. This proves the lemma. ■

The lemma implies all the algebraic properties for line integrals hold in the same way for contour integrals. In particular, if γ is C^1, then

$$\int_\gamma f(z)\, dz = \int_a^b f(\gamma(t))\, \gamma'(t)\, dt.$$

Another important observation is the following.

Proposition 14.6.5 *Suppose $F'(z) = f(z)$ for all $z \in \Omega$, an open set containing γ^* where $\gamma : [a, b] \to \mathbb{C}$ is a continuous bounded variation curve. Then*

$$\int_\gamma f(z)\, dz = F(\gamma(b)) - F(\gamma(a)).$$

Proof: Letting u and v be real and imaginary parts of f, it follows from Lemma 14.6.4

$$\int_\gamma f(z)\,dz = \int_\gamma (u,-v)\cdot d\gamma + i\int_\gamma (v,u)\cdot d\gamma \tag{14.41}$$

Consider the real valued function

$$G(x,y) \equiv \frac{1}{2}\left(F(x+iy) + \overline{F(x+iy)}\right) \equiv \mathrm{Re}\,F(x+iy)$$

By assumption,

$$F'(x+iy) = f(x+iy) = u(x,y) + iv(x,y).$$

Thus it is routine to verify $\nabla G = (u,-v)$. Next let the real valued function H be defined by

$$H(x,y) \equiv \frac{1}{2i}\left(F(x+iy) - \overline{F(x+iy)}\right) \equiv \mathrm{Im}\,F(x+iy)$$

Then $\nabla H = (v,u)$ and so from 14.41 and Theorem 14.2.11

$$\begin{aligned}\int_\gamma f(z)\,dz &= G(\gamma(b)) - G(\gamma(a)) + i\left(H(\gamma(b)) - H(\gamma(a))\right)\\ &= F(\gamma(b)) - F(\gamma(a)).\end{aligned}$$

This proves the proposition. ■

A function F such that $F' = f$ is called a **primitive** of f. See how it acts a lot like a potential, the difference being that a primitive has complex, not real values. In calculus, in the context of a function of one real variable, this is often called an antiderivative and every continuous function has one thanks to the fundamental theorem of calculus. However, it will be shown below that the situation is not at all the same for functions of a complex variable.

14.6.3 The Cauchy Integral

The following is the first form of the Cauchy integral theorem.

Lemma 14.6.6 *Let U be an open set in $\mathbb{C}$ and let Γ be a simple closed rectifiable curve contained in U having parameterization γ such that the inside of Γ is contained in U. Also let f be analytic in U. Then*

$$\int_\gamma f(z)\,dz = 0.$$

Proof: This follows right away from the Cauchy Riemann equations and the formula 14.40. Assume without loss of generality the orientation of Γ is the positive orientation. If not, the argument is the same. Then from formula 14.40,

$$\int_\gamma f(z)\,dz = \int_\gamma (u,-v)\cdot d\gamma + i\int_\gamma (v,u)\cdot d\gamma$$

and by Green's theorem and U_i the inside of Γ this equals

$$\int_{U_i} (-v_x - u_y)\,dm_2 + i\int_\gamma (u_x - v_y)\,dm_2 = 0$$

by the Cauchy Riemann equations. This proves the lemma. ■

It is easy to improve on this result using the argument for proving Green's theorem. You only need continuity on the bounding curve. You also don't need to make any assumption about the functions u_x, etc. being in $L^1(U)$. The following is a very general version of the Cauchy integral theorem.

Theorem 14.6.7 *Let U_i be the inside of Γ a simple closed rectifiable curve having parameterization γ. Also let f be analytic in U_i and continuous on $U_i \cup \Gamma$. Then*

$$\int_\gamma f(z)\,dz = 0.$$

Proof: Let $\mathcal{B}_\delta, \mathcal{I}_\delta$ be those regions of Lemma 14.3.6 where as earlier $\mathcal{I}_\delta$ are those which have empty intersection with Γ and $\mathcal{B}_\delta$ are the border regions. Without loss of generality, assume Γ is positively oriented. As in the proof of Green's theorem you can apply the same argument to the line integrals on the right of 14.40 to obtain, just as in the proof of Green's theorem

$$\sum_{R\in\mathcal{I}_\delta}\int_{\partial R} f(z)\,dz + \sum_{R\in\mathcal{B}_\delta}\int_{\partial R} f(z)\,dz = \int_\gamma f(z)\,dz$$

In this case the first sum on the left in the above formula equals 0 from Lemma 14.6.6 for any $\delta > 0$. Recall that there were at most

$$4\left(\frac{V(\Gamma)}{\delta}+1\right)$$

border regions where each of these border regions is contained in a box having sides of length no more than 2δ. Letting $\varepsilon > 0$ be given, suppose δ is so small that for $|z-w| < 8\delta$ with $z, w \in \overline{U_i}$,

$$|f(z) - f(w)| < \varepsilon$$

this by uniform continuity of f. Let R be a border region. Then picking z_1 a point in R,

$$\int_{\partial R} f(z)\,dz = \int_{\partial R} (f(z) - f(z_1))\,dz + \int_{\partial R} f(z_1)\,dz$$

The last contour integral equals 0 because $f(z_1)$ has a primitive, namely $F(z) = f(z_1)\,z$. If follows that

$$\left|\int_{\partial R} f(z)\,dz\right| = \left|\int_{\partial R} (f(z) - f(z_1))\,dz\right| \le \varepsilon V(\partial R) \le \varepsilon 8\delta + \varepsilon l_R$$

where l_R is the length of the part of Γ which is part of ∂R. It follows that

$$\begin{aligned}\left|\sum_{R\in\mathcal{B}_\delta}\int_{\partial R} f(z)\,dz\right| &\le \sum_{R\in\mathcal{B}_\delta}\left|\int_{\partial R} f(z)\,dz\right| \le \sum_{R\in\mathcal{B}_\delta}\varepsilon 8\delta + \sum_{R\in\mathcal{B}_\delta}\varepsilon l_R\\ &\le \varepsilon 8\delta\left(4\left(\frac{V(\Gamma)}{\delta}+1\right)\right) + \varepsilon V(\Gamma)\\ &\le \varepsilon\left(32V(\Gamma) + 32\delta\right) + \varepsilon V(\Gamma)\end{aligned}$$

So, since ε is arbitrary,

$$\lim_{\delta\to 0+}\left|\sum_{R\in\mathcal{B}_\delta}\int_{\partial R} f(z)\,dz\right| = 0$$

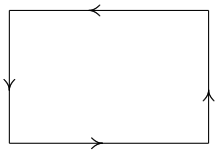
Hence

$$\int_\gamma f(z)\,dz = 0 + \lim_{\delta\to 0+}\sum_{R\in\mathcal{B}_\delta}\int_{\partial R} f(z)\,dz = 0$$

This proves the theorem.

With this really marvelous theorem it is time to consider the Cauchy integral formula which represents the value of an analytic function at a point on the inside in terms of its values on the boundary. First here are some lemmas.

Lemma 14.6.8 *Let R be a rectangle such that ∂R is positively oriented. Recall this means the direction of motion is counter clockwise.*

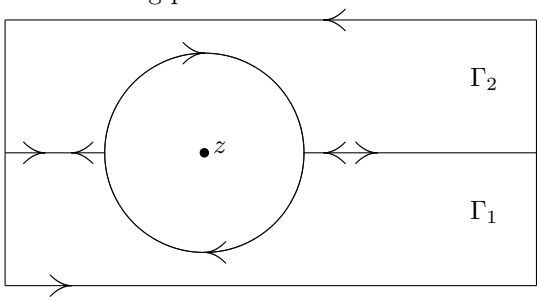


Then if z is on the inside of ∂R,

$$\frac{1}{2\pi i}\int_{\partial R}\frac{1}{w-z}dw = 1$$

while if z is on the outside of ∂R, the above integral equals 0.

Proof: Consider the following picture.



The idea is that you put a circle around z as shown. Then draw the two simple closed curves Γ_1, Γ_2 oriented as shown. Then the line integrals cancel on the two lines which involve two different directions. Therefore, with these orientations, it follows from the Cauchy integral theorem that

$$0 = \int_{\Gamma_1}\frac{1}{w-z}dw + \int_{\Gamma_2}\frac{1}{w-z}dw \tag{14.42}$$

Letting C denote the circle oriented as shown, this implies

$$\int_{\partial R}\frac{1}{w-z}dw - \int_{C}\frac{1}{w-z}dw = 0$$

However, it is easy to compute the integral over the circle. This equals

$$\int_{C}\frac{1}{w-z}dw = \int_0^{2\pi}\frac{1}{re^{it}}rie^{it}dt = 2\pi i$$

which shows the claimed formula in the case that z is on the inside of ∂R.

In the case where z is on the outside of ∂R, the conclusion follows from the Cauchy integral formula Theorem 14.6.7 as you can verify by noting that $f(w) \equiv 1/(w-z)$ is analytic on an open set containing R and that in fact its derivative equals what you would think,

$$-1/(w-z)^2 . \blacksquare$$

Now with this little lemma, here is the Cauchy integral formula.

Theorem 14.6.9 *Let Γ be a positively oriented simple closed rectifiable curve having parameterization γ and let $z \in U_i$, the inside of Γ. Also let f be analytic on U_i, and continuous on $U_i \cup \Gamma$. Then*

$$f(z) = \frac{1}{2\pi i}\int_{\gamma}\frac{f(w)}{w-z}dw.$$

In particular, letting $f(z) \equiv 1$,

$$\frac{1}{2\pi i}\int_\gamma \frac{1}{w-z}dw = 1.$$

Proof: In constructing the special regions in the proof of Green's theorem, always choose δ such that the point z is not on any of the lines $m\delta = y$ and $x = k\delta$. This makes it possible to avoid thinking about the case where z is not on the interior of any of the rectangles of $\mathcal{I}_\delta$. Pick δ small enough that $\mathcal{I}_\delta \neq \emptyset$ and z is contained in some $R_0 \in \mathcal{I}_\delta$. From Lemma 14.6.8 it follows for each $R \in \mathcal{I}_\delta$

$$\frac{1}{2\pi i}\int_R \frac{f(w)}{w-z}dw - f(z) = \frac{1}{2\pi i}\int_R \frac{f(w)-f(z)}{w-z}dw$$

Then as in the proof of Theorem 14.6.7

$$\begin{aligned} & \frac{1}{2\pi i}\sum_{R\in\mathcal{I}_\delta}\int_{\partial R}\frac{f(w)-f(z)}{w-z}dw + \frac{1}{2\pi i}\sum_{R\in\mathcal{B}_\delta}\int_{\partial R}\frac{f(w)-f(z)}{w-z}dw \\ = & \frac{1}{2\pi i}\int_\gamma \frac{f(w)-f(z)}{w-z}dw \end{aligned}$$

By Theorem 14.6.7, all these integrals on the left equal 0 except for R_0, the one which contains z on its interior.

Thus the above reduces to

$$\frac{1}{2\pi i}\int_{\partial R_0}\frac{f(w)-f(z)}{w-z}dw = \frac{1}{2\pi i}\int_\gamma \frac{f(w)-f(z)}{w-z}dw$$

The integrand of the left converges to $f'(z)$ as $\delta \to 0$ and the length of R_0 also converges to 0 so it follows from Theorem 14.2.4 that the limit as $\delta \to 0$ in the above exists and yields

$$0 = \frac{1}{2\pi i}\int_\gamma \frac{f(w)-f(z)}{w-z}dw = \frac{1}{2\pi i}\int_\gamma \frac{f(w)}{w-z}dw - f(z)\frac{1}{2\pi i}\int_\gamma \frac{1}{w-z}dw. \tag{14.43}$$

Consider the last integral above.

$$\sum_{R\in\mathcal{T}_\delta}\int_{\partial R}\frac{1}{w-z}dw + \sum_{R\in\mathcal{B}_\delta}\int_{\partial R}\frac{1}{w-z}dw = \int_\gamma \frac{1}{w-z}dw \tag{14.44}$$

As in the proof of Green's theorem, choosing δ small enough the second sum on the left in the above satisfies

$$\left|\sum_{R\in\mathcal{B}_\delta}\int_{\partial R}\frac{1}{w-z}dw\right| \leq \sum_{R\in\mathcal{B}_\delta}\left|\int_{\partial R}\frac{1}{w-z}dw\right| < \varepsilon.$$

By Lemma 14.6.8, the first sum on the left in 14.44 equals

$$\int_{\partial R_0}\frac{1}{w-z}dw$$

where R_0 is the rectangle for which z is on the inside of ∂R_0. Then by this lemma again, this equals $2\pi i$. Therefore for such small δ, 14.44 reduces to

$$\left|2\pi i - \int_\gamma \frac{1}{w-z}dw\right| < \varepsilon$$

Since ε is arbitrary, this shows

$$\frac{1}{2\pi i}\int_\gamma \frac{1}{w-z}dw = 1.$$

Now using this, 14.43 implies the claimed formula of the theorem. This proves the theorem. ■

Now here is an important lemma about the contour integral and limits. It says the integral of a limit is the limit of the integrals.

Lemma 14.6.10 *Let $\gamma : [a,b] \to \mathbb{C}$ be of bounded variation. Let f be continuous on γ^*. Also let $\{f_k\}$ be a sequence of continuous functions converging uniformly to f on γ^*. Then*

$$\int_\gamma f(z)\,dz = \lim_{k\to\infty}\int_\gamma f_k(z)\,dz$$

Proof: Let $\varepsilon > 0$ be given. Then there is $\delta > 0$ such that if $\|\mathcal{P}\| < \delta$, then for $\mathcal{P} \equiv \{t_0, \cdots, t_n\}$

$$\left|\int_\gamma f(z)\,dz - \sum_{k=1}^n f(\gamma(s_i))(\gamma(t_i)-\gamma(t_{i-1}))\right| < \varepsilon$$

whenever $s_i \in [t_{i-1}, t_i]$. Also let K be large enough that for $k \geq K$,

$$\max_{t\in[a,b]} |f(\gamma(t)) - f_k(\gamma(t))| < \varepsilon$$

Then pick such a k. Choose $\mathcal{P}$ such that $\|\mathcal{P}\| < \delta$ and also $\|\mathcal{P}\|$ is small enough that

$$\left|\int_\gamma f_k(z)\,dz - \sum_{k=1}^n f_k(\gamma(s_i))(\gamma(t_i)-\gamma(t_{i-1}))\right| < \varepsilon$$

for any choice of $s_i \in [t_{i-1}, t_i]$. Then

$$\left|\int_\gamma f(z)\,dz - \int_\gamma f_k(z)\,dz\right| \leq \left|\int_\gamma f(z)\,dz - \sum_{k=1}^n f(\gamma(s_i))(\gamma(t_i)-\gamma(t_{i-1}))\right| +$$

$$\left|\sum_{k=1}^n f(\gamma(s_i))(\gamma(t_i)-\gamma(t_{i-1})) - \sum_{k=1}^n f_k(\gamma(s_i))(\gamma(t_i)-\gamma(t_{i-1}))\right|$$

$$+\left|\int_\gamma f_k(z)\,dz - \sum_{k=1}^n f_k(\gamma(s_i))(\gamma(t_i)-\gamma(t_{i-1}))\right|$$

$$\leq 2\varepsilon + \varepsilon\sum_{k=1}^n |\gamma(t_i)-\gamma(t_{i-1})| \leq 2\varepsilon + \varepsilon V(\gamma,[a,b])$$

Since ε is arbitrary, this shows that

$$\lim_{k\to\infty}\int_\gamma f_k(z)\,dz = \int_\gamma f(z)\,dz$$

as claimed. ■

Theorem 14.6.11 *Let $\gamma : [a,b] \to \mathbb{C}$ be of bounded variation. Let f be continuous on γ^*. For $z \notin \gamma^*$, define*

$$g(z) \equiv \int_\gamma \frac{f(w)}{w-z} dw$$

Then g is infinitely differentiable. Futhermore,

$$g^{(n)}(z) = n! \int_\gamma \frac{f(w)}{(w-z)^{n+1}} dw$$

$$(g(z+h) - g(z))/h = \frac{1}{h} \int_\gamma \left(\frac{f(w)}{(w-z-h)} - \frac{f(w)}{(w-z)} \right) dw$$

$$= \frac{1}{h} \int_\gamma f(w) \left(\frac{h}{(w-z-h)(w-z)} \right) dw = \int_\gamma f(w) \left(\frac{1}{(w-z-h)(w-z)} \right) dw$$

Consider only $h \in \mathbb{C}$ such that $2|h| < \operatorname{dist}(z, \gamma^*)$. The integrand converges to $1/(w-z)^2$. Then for these values of h,

$$\left| \frac{1}{(w-z-h)(w-z)} - \frac{1}{(w-z)^2} \right| = \left| \frac{h}{(w-z-h)(w-z)^2} \right|$$

$$\leq \frac{|h|}{\operatorname{dist}(z,\gamma^*)^3/2} = \frac{2|h|}{\operatorname{dist}(z,\gamma^*)^3}$$

and so the convergence of the integrand to

$$f(w)/(w-z)^2$$

is uniform for $|h| < \operatorname{dist}(z,\gamma^*)/2$. Using Theorem 14.2.4, it follows Lemma 14.6.10 applied to an arbitrary sequence corresponding to $h_k \to 0$,

$$\begin{aligned} g'(z) &= \lim_{h\to 0} \frac{g(z+h) - g(z)}{h} \\ &= \lim_{h\to 0} \frac{1}{h} \int_\gamma \left(\frac{f(w)}{(w-z-h)} - \frac{f(w)}{(w-z)} \right) dw \\ &= \int_\gamma \frac{f(w)}{(w-z)^2} dw. \end{aligned}$$

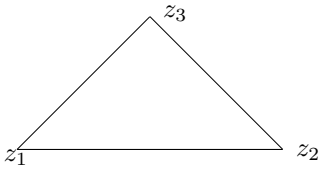
One can then differentiate the above expression using the same arguments. Continuing this way results in the following formula.

$$g^{(n)}(z) = n! \int_\gamma \frac{f(w)}{(w-z)^{n+1}} dw \blacksquare$$

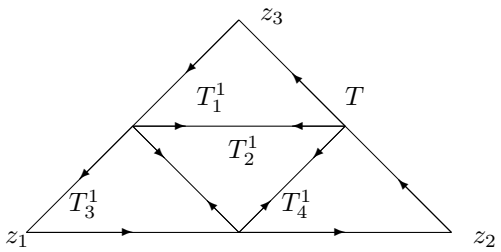
It turns out that in the definition of what it means for a function defined on an open set, U to be analytic it is not necessary to say that $z \to f'(z)$ is continuous. In fact, this comes for free. The statement that $z \to f'(z)$ is continuous is REDUNDANT! The key to understanding this is the Cauchy Goursat theorem.

14.6.4 The Cauchy Goursat Theorem

If you have two points in $\mathbb{C}$, z_1 and z_2, you can consider $\gamma(t) \equiv z_1 + t(z_2 - z_1)$ for $t \in [0,1]$ to obtain a continuous bounded variation curve from z_1 to z_2. More generally, if $z_1, \cdots, z_m$ are points in $\mathbb{C}$ you can obtain a continuous bounded variation curve from z_1 to z_m which consists of first going from z_1 to z_2 and then from z_2 to z_3 and so on, till in the end one goes from z_{m-1} to z_m. Denote this piecewise linear curve as $\gamma(z_1, \cdots, z_m)$. Now let T be a triangle with vertices z_1, z_2 and z_3 encountered in the counter clockwise direction as shown.



Denote by $\int_{\partial T} f(z)\,dz$, the expression, $\int_{\gamma(z_1,z_2,z_3,z_1)} f(z)\,dz$. Consider the following picture.



Thus

$$\int_{\partial T} f(z)\,dz = \sum_{k=1}^{4} \int_{\partial T_k^1} f(z)\,dz. \tag{14.45}$$

On the "inside lines" the integrals cancel because there are two integrals going in opposite directions for each of these inside lines.

Theorem 14.6.12 *(Cauchy Goursat) Let $f : \Omega \to \mathbb{C}$ have the property that $f'(z)$ exists for all $z \in \Omega$ and let T be a triangle contained in Ω. Then*

$$\int_{\partial T} f(w)\,dw = 0.$$

Proof: Suppose not. Then

$$\left| \int_{\partial T} f(w)\,dw \right| = \alpha \neq 0.$$

From 14.45 it follows

$$\alpha \leq \sum_{k=1}^{4} \left| \int_{\partial T_k^1} f(w)\,dw \right|$$

and so for at least one of these T_k^1, denoted from now on as T_1,

$$\left| \int_{\partial T_1} f(w)\,dw \right| \geq \frac{\alpha}{4}.$$

Now let T_1 play the same role as T. Subdivide as in the above picture, and obtain T_2 such that

$$\left|\int_{\partial T_2} f(w)\,dw\right| \geq \frac{\alpha}{4^2}.$$

Continue in this way, obtaining a sequence of triangles,

$$T_k \supseteq T_{k+1}, \operatorname{diam}(T_k) \leq \operatorname{diam}(T)\,2^{-k},$$

and

$$\left|\int_{\partial T_k} f(w)\,dw\right| \geq \frac{\alpha}{4^k}.$$

Then let $z \in \cap_{k=1}^{\infty} T_k$ and note that by assumption, $f'(z)$ exists. Therefore, for all k large enough,

$$\int_{\partial T_k} f(w)\,dw = \int_{\partial T_k} \left(f(z) + f'(z)(w-z) + g(w)\right) dw$$

where $|g(w)| < \varepsilon |w-z|$. Now observe that $w \to f(z) + f'(z)(w-z)$ has a primitive, namely,

$$F(w) = f(z)\,w + f'(z)(w-z)^2/2.$$

Therefore, by Proposition 14.6.5,

$$\int_{\partial T_k} f(w)\,dw = \int_{\partial T_k} g(w)\,dw.$$

From Theorem 14.2.4 applied to contour integrals or the definition of the contour integral,

$$\begin{aligned} \frac{\alpha}{4^k} &\leq \left|\int_{\partial T_k} g(w)\,dw\right| \leq \varepsilon \operatorname{diam}(T_k)\,(\text{length of } \partial T_k) \\ &\leq \varepsilon 2^{-k}\,(\text{length of } T)\operatorname{diam}(T)\,2^{-k}, \end{aligned}$$

and so

$$\alpha \leq \varepsilon\,(\text{length of } T)\operatorname{diam}(T).$$

Since ε is arbitrary, this shows $\alpha = 0$, a contradiction. Thus $\int_{\partial T} f(w)\,dw = 0$ as claimed. ■

This fundamental result yields the following important theorem.

Theorem 14.6.13 *(Morera[1]) Let Ω be an open set and let $f'(z)$ exist for all $z \in \Omega$. Let $D \equiv \overline{B(z_0, r)} \subseteq \Omega$. Then there exists $\varepsilon > 0$ such that f has a primitive on $B(z_0, r+\varepsilon)$. (Recall this is a function F such that $F'(z) = f(z)$.)*

Proof: Choose $\varepsilon > 0$ small enough that $B(z_0, r+\varepsilon) \subseteq \Omega$. Then for $w \in B(z_0, r+\varepsilon)$, define

$$F(w) \equiv \int_{\gamma(z_0, w)} f(u)\,du.$$

Then by the Cauchy Goursat theorem, Theorem 14.6.12, and $w \in B(z_0, r+\varepsilon)$, it follows that for $|h|$ small enough,

$$\frac{F(w+h) - F(w)}{h} = \frac{1}{h}\int_{\gamma(w, w+h)} f(u)\,du$$

[1] Giancinto Morera 1856-1909. This theorem or one like it dates from around 1886

$$= \frac{1}{h}\int_0^1 f(w+th)\,hdt = \int_0^1 f(w+th)\,dt$$

which converges to $f(w)$ due to the continuity of f at w. This proves the theorem. ■

The following is a slight generalization of the above theorem which is also referred to as Morera's theorem. It contains the proof that the condition of continuity of $z \to f'(z)$ is redundant.

Corollary 14.6.14 *Let Ω be an open set and suppose that whenever*

$$\gamma(z_1, z_2, z_3, z_1)$$

is a closed curve bounding a triangle T, which is contained in Ω, and f is a continuous function defined on Ω, it follows that

$$\int_{\gamma(z_1,z_2,z_3,z_1)} f(z)\,dz = 0,$$

then f is analytic on Ω. Also, if $f'(z)$ exists for $z \in \Omega$, then $z \to f'(z)$ is continuous.

Proof: As in the proof of Morera's theorem, let $\overline{B(z_0,r)} \subseteq \Omega$ and use the given condition to construct a primitive, F for f on $B(z_0,r)$. (As just shown in Theorem 14.6.13, the given condition is satisfied whenever $f'(z)$ exists for all $z \in \Omega$.) Then F is analytic and so by the Cauchy integral formula, for $z \in B(z_0,r)$

$$F(z) = \frac{1}{2\pi i}\int_{\partial B(z_0,r)} \frac{F(w)}{w-z}dw.$$

It follows from Theorem 14.6.11 that F and hence f have infinitely many derivatives, implying that f is analytic on $B(z_0,r)$. Since z_0 is arbitrary, this shows f is analytic on Ω. In particular $z \to f'(z)$ is continuous because actually this function is differentiable. This proves the corollary. ■

This shows that an equivalent definition of what it means for a function to be analytic is the following definition.

Definition 14.6.15 *Let U be an open set in $\mathbb{C}$ and suppose $f'(z)$ exists for all $z \in U$. Then f is called analytic.*

These theorems form the foundation for the study of functions of a complex variable. Some important theorems will be discussed in the exercises.

14.7 Exercises

1. Suppose $f : [a,b] \to [c,d]$ is continuous and one to one on (a,b). For $s \in (c,d)$, show
$$d(f,(a,b),s) = \pm 1$$
show it is 1 if f is increasing and -1 if f is decreasing. How can this be used to relate the degree to orientation?

2. In defining a simple curve the assumption was made that $\boldsymbol{\gamma}(t) \neq \boldsymbol{\gamma}(a)$ and $\boldsymbol{\gamma}(t) \neq \boldsymbol{\gamma}(b)$ if $t \in (a,b)$. Is this fussy condition really necessary? Which theorems and lemmas hold with simply assuming $\boldsymbol{\gamma}$ is one to one on (a,b)? Does the fussy condition follow from assuming $\boldsymbol{\gamma}$ is one to one on (a,b)?

3. Show that for many open sets in $\mathbb{R}^2$, Area of $U = \int_{\partial U} xdy$, and Area of $U = \int_{\partial U} -ydx$ and Area of $U = \frac{1}{2}\int_{\partial U} -ydx + xdy$. **Hint:** Use Green's theorem.

4. A closed polygon in the plane starts at (x_0, y_0), goes to (x_1, y_1), to (x_2, y_2) to $\cdots (x_n, y_n) = (x_0, y_0)$. Suppose the line segments never cross so that you have a simple closed curve. Using Green's theorem find a simple formula for the area of the parallelogram. You can use Problem 3. Get the area using a line integral obtained by adding the line integrals corresponding to the vertices of the polygon.

5. Let Γ be a simple C^1 oriented curve having parameterization $\boldsymbol{\gamma}$ where t is the time and suppose $\mathbf{f}$ is a force defined on Γ. Then the work done by $\mathbf{f}$ on an object of mass m as it moves over the curve is defined by

$$\int_\gamma \mathbf{f} \cdot d\boldsymbol{\gamma}$$

Newton's second law states that $\mathbf{f} = m\frac{d\mathbf{v}}{dt}$ where $\mathbf{v} \equiv \boldsymbol{\gamma}'(t)$. Let $\mathbf{v}_b = \boldsymbol{\gamma}'(b)$ with $\mathbf{v}_a$ defined similarly. Thus these are the final and initial velocities. Show the work equals

$$\frac{1}{2}m|\mathbf{v}_b|^2 - \frac{1}{2}m|\mathbf{v}_a|^2.$$

6. In the situation of the above problem, show that if $\mathbf{f}(\mathbf{x}) = \nabla F(\mathbf{x})$ where F is a potential, then if the motion is governed by the Newton's law it follows that for $\boldsymbol{\gamma}(t)$ the motion,

$$-F(\boldsymbol{\gamma}(t)) + \frac{1}{2}m|\boldsymbol{\gamma}'(t)|^2$$

is constant.

7. Generalize Stoke's theorem, Theorem 14.4.4 to the case where $\mathbf{R}$ is only assumed C^1.

8. Given an example of a simple closed rectifiable curve Γ and a horizontal line which intersects this curve in infinitely many points.

9. Let Γ be a simple closed rectifiable curve and let U_i be its inside. Show you can remove any finite number of circular disks from U_i and what remains will still be a region for which Green's theorem holds. **Hint:** You might get some ideas from looking at the proof of Lemma 14.3.6. This is much harder than it looks because you only know Γ is a simple closed rectifiable curve. Begin by punching one circular hole and go from there.

10. Let $\gamma : [a, b] \to \mathbb{R}$ be of bounded variation. Show there exist increasing functions $f(t)$ and $g(t)$ such that

$$\gamma(t) = f(t) - g(t).$$

Hint: You might let $f(t) = V(\gamma; [a, t])$. Show this is increasing and then consider $g(t) = f(t) - \gamma(t)$.

11. Using Problem 10 describe another way to obtain the integral $\int_\gamma f d\gamma$ for f a real valued function and γ a real valued curve of bounded variation as just described using the theory of Lebesgue integration. What exactly is this integral in this simple case? Next extend to the case where $\boldsymbol{\gamma}$ has values in $\mathbb{R}^n$ and $\mathbf{f} : \boldsymbol{\gamma}^* \to \mathbb{R}^n$. What are some advantages of using this other approach?

12. Suppose f is continuous but not analytic and a function of $z \in U \subseteq \mathbb{C}$. Show f has no primitive. When functions of real variables are considered, there are function spaces $C^m(U)$ which specify how many continuous derivatives the function has. Why are such function spaces irrelevant when considering functions of a complex variable?

13. Analytic functions are all just long polynomials. Prove this disappointing result. More precisely prove the following. If $f : U \to \mathbb{C}$ is analytic where U is an open set and if $B(z_0, r) \subseteq U$, then

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n \tag{14.46}$$

for all $|z - z_0| < r$. Furthermore,

$$a_n = \frac{f^{(n)}(z_0)}{n!}. \tag{14.47}$$

Hint: You use the Cauchy integral formula. For $z \in B(z_0, r)$ and C the positively oriented boundary,

$$\begin{aligned} f(z) &= \frac{1}{2\pi i} \int_C \frac{f(w)}{w - z} = \frac{1}{2\pi i} \int_C \frac{f(w)}{w - z_0} \frac{1}{1 - \frac{z - z_0}{w - z_0}} dw \\ &= \frac{1}{2\pi i} \int_C \sum_{n=0}^{\infty} \frac{f(w)}{(w - z_0)^{n+1}} (z - z_0)^n \, dw \end{aligned}$$

Now explain why you can switch the sum and the integral. You will need to argue the sum converges uniformly which is what will justify this manipulation. Next use the result of Theorem 14.6.11.

14. Prove the following amazing result about the zeros of an analytic function. Let Ω be a connected open set (region) and let $f : \Omega \to X$ be analytic. Then the following are equivalent.

    (a) $f(z) = 0$ for all $z \in \Omega$

    (b) There exists $z_0 \in \Omega$ such that $f^{(n)}(z_0) = 0$ for all n.

    (c) There exists $z_0 \in \Omega$ which is a limit point of the set,

$$Z \equiv \{z \in \Omega : f(z) = 0\}.$$

    Hint: From Problem 13, if *c.*) holds, then for z near z_0

$$f(z) = \sum_{n=m}^{\infty} \frac{f^{(n)}(z_0)}{n!} (z - z_0)^n$$

    Say $f^{(n)}(z_0) \neq 0$. Then consider

$$\frac{f(z)}{(z - z_0)^m} = \frac{f^{(m)}(z_0)}{m!} + \sum_{n=m+1}^{\infty} \frac{f^{(n)}(z_0)}{n!} (z - z_0)^{n-m}$$

    Now let $z_n \to z_0, z_n \neq z_0$ but $f(z_n) = 0$. What does this say about $f^{(m)}(z_0)$? Clearly the first two conditions are equivalent and they imply the third.

15. You want to define e^z for z complex such that it is analytic on $\mathbb{C}$. Using Problem 14 explain why there is at most one way to do it and still have it coincide with e^x when $z = x + i0$. Then show using the Cauchy Riemann equations that

$$e^z \equiv e^x (\cos(y) + i \sin(y))$$

is analytic and agrees with e^x when $z = x + i0$. Also show

$$\frac{d}{dz}e^z = e^z.$$

Hint: For the first part, suppose two functions, f, g work. Then consider $f - g$. this is analytic and has a zero set, $\mathbb{R}$.

16. Do the same thing as Problem 15 for $\sin(z), \cos(z)$. Also explain with a very short argument why all identities for these functions continue to hold for the extended functions. This argument shouldn't require any computations at all. Why is $\sin(z)$ no longer bounded if z is allowed to be complex? **Hint:** You might try something involving the above formula for e^z to get the definition.

17. Show that if f is analytic on $\mathbb{C}$ and $f'(z) = 0$ for all z, then $f(z) \equiv c$ for some constant $c \in \mathbb{C}$. You might want to use Problem 14 to do this really quickly. Now using Theorem 14.6.11 prove Liouville's theorem which states that a function which is analytic on all of $\mathbb{C}$ which is also bounded is constant. **Hint:** By that theorem,

$$f'(z) = \frac{1}{2\pi i}\int_{C_r} \frac{f(w)}{(w-z)^2}dw$$

where C_r is the positively oriented circle of radius r which is centered at z. Now consider what happens as $r \to \infty$. You might use the corresponding version of Theorem 14.2.4 applied to contour integrals and note the total length of C_r is $2\pi r$.

18. Using Problem 15 prove the fundamental theorem of algebra which says every nonconstant polynomial having complex coefficients has at least one zero in $\mathbb{C}$. (This is the very best way to prove the fundamental theorem of algebra.) **Hint:** If $p(z)$ has no zeros, consider $1/p(z)$ and prove it must then be bounded and analytic on all of $\mathbb{C}$.

19. Let f be analytic on U_i, the inside of Γ, a rectifiable simple closed curve positively oriented with parameterization γ. Suppose also there are no zeros of f on Γ. Show then that the number of zeros, of f contained in U_i counted according to multiplicity is given by the formula

$$\frac{1}{2\pi i}\int_\gamma \frac{f'(z)}{f(z)}dz$$

Hint: You ought to first show $f(z) = \prod_{k=1}^m (z - z_k)\, g(z)$ where the z_k are the zeros of f in U_i and $g(z)$ is an analytic function which never vanishes in $U_i \cup \Gamma$. In the above product there might be some repeats corresponding to repeated zeros.

20. An open connected set U is said to be star shaped if there exists a point $z_0 \in U$ called a star center such that the for all $z \in U, \gamma(z_0, z)^*$ as described in before the proof of the Cauchy Goursat theorem is contained in U. For example, pick any complex number α and consider everything left after leaving out the ray $\{t\alpha : t \geq 0\}$. Show this is star shaped with a star center $t\alpha$ for $t < 0$. Now for U a star shaped open connected set, suppose g is analytic on U and $g(z) \neq 0$ for all $z \in U$. Show there exists an analytic function h defined on U such that

$$e^{h(z)} = g(z).$$

This function $h(z)$ is like $\log(g(z))$. **Hint:** Use an argument like that used to prove Morera's theorem and the Cauchy Goursat theorem to obtain a primitive for g'/g, h_1. Next consider the function

$$ge^{-h_1}$$

Using the chain rule and the product rule, show $\frac{d}{dz}\left(ge^{-h_1}\right)=0$. Using one of the results of Problem 17 show

$$g=ce^{h_1}$$

for some constant c. Tell why c can be written as e^{a+ib}. Then let $h=h_1+a+ib$.

21. One of the most amazing theorems is the open mapping theorem. Let U be an open connected set in $\mathbb{C}$ and suppose $f:U\to\mathbb{C}$ is analytic. Then $f(U)$ is either a point or an open connected set. In the case where $f(U)$ is an open connected set, it follows that for each $z_0\in U$, there exists an open set, V containing z_0 and $m\in\mathbb{N}$ such that for all $z\in V$,

$$f(z)=f(z_0)+\phi(z)^m \tag{14.48}$$

where $\phi:V\to B(0,\delta)$ is one to one, analytic and onto, $\phi(z_0)=0$, $\phi'(z)\neq 0$ on V and ϕ^{-1} analytic on $B(0,\delta)$. If f is one to one then $m=1$ for each z_0 and $f^{-1}:f(U)\to U$ is analytic. Consider the real valued function $f(x)=x^2$. $f(\mathbb{R})$ is neither a point nor an open connected set. This is a strictly complex analysis phenomenon. **Hint:**Work out the details of the following outline. Suppose $f(U)$ is not a point. Then using Problem 14 about the zeros of an analytic function there exists $r>0$ such that for $z\in B(z_0,r)\setminus\{z_0\}$,

$$f(z)-f(z_0)\neq 0.$$

Explain why there exists $g(z)$ analytic and nonzero on $B(z_0,r)$ such that for some positive integer m,

$$f(z)-f(z_0)=(z-z_0)^m g(z)$$

Next one tries to take the m^{th} root of $g(z)$. Using Problem 20 there exists h analytic such that

$$g(z)=e^{h(z)},\ g(z)=\left(e^{h(z)/m}\right)^m$$

Now let $\phi(z)=(z-z_0)e^{h(z)/m}$. This yields the formula 14.48. Also $\phi'(z_0)=e^{h(z_0)/m}\neq 0$. Now consider

$$\phi(x+iy)=u(x,y)+iv(x,y)$$

and the map

$$\begin{pmatrix} x\\ y\end{pmatrix}\to\begin{pmatrix} u(x,y)\\ v(x,y)\end{pmatrix}$$

Here u,v are C^1 because ϕ is analytic. Use the Cauchy Riemann equations to verify the Jacobian of this transformation at (x_0,y_0) is nonzero. This is where you use $\phi'(z_0)\neq 0$. Use inverse function theorem to verify ϕ maps some open set V containing z_0 one to one and onto $B(0,\delta)$. Thus also ϕ^m maps V onto $B(0,\delta^m)$. Explain why it follows from 14.48 and the fact that z_0 is arbitrary that f is an open map. Since f is continuous and U is connected, so is $f(U)$. However, if $m>1$, this mapping, f can't be one to one. To verify this,

$$e^{i2\pi/m}\phi(z_1)\neq\phi(z_1)$$

but both are in $B(0,\delta)$. Hence there exists $z_2\neq z_1$ such that $\phi(z_2)=e^{i2\pi/m}\phi(z_1)$ (ϕ is one to one) but $f(z_2)=f(z_1)$. If f is one to one, then the above shows that f^{-1} is continuous and for each z, the m in the above is always 1 so $f'(z)=e^{h(z)/1}\neq 0$. Hence

$$\begin{aligned}\left(f^{-1}\right)'(f(z)) &= \lim_{f(z_1)\to f(z)}\frac{f^{-1}(f(z_1))-f^{-1}(f(z))}{f(z_1)-f(z)}\\ &= \lim_{z_1\to z}\frac{z_1-z}{f(z_1)-f(z)}=\frac{1}{f'(z)}\end{aligned}$$

22. Let U be what is left when you leave out the ray $t\alpha$ for $t \geq 0$. This is a star shaped open set and $g(z) = z$ is nonzero on this set. Therefore, there exists $h(z)$ such that $z = e^{h(z)}$ by Problem 20. Explain why $h(z)$ is analytic on U. When $\alpha = -1$ this is called the principle branch of the logarithm. In this case define $Arg(z) \equiv \theta \in (-\pi, \pi)$ such that the given z equals $|z| e^{i\theta}$. Explain why this principle branch of the logarithm is

$$\log(z) = \ln(|z|) + iArg(z)$$

Note it follows from the open mapping theorem this is an analytic function on U. You don't have to fuss with any tedium in order to show this.

23. Suppose Γ is a simple closed curve and let U_i be the inside. Suppose f is analytic on U_i and continuous on $U_i \cup \Gamma$. Consider the function $z \to |f(z)|$. This is a continuous function. Show that if it achieves its maximum at any point of U_i then f must be a constant. **Hint:** You might use the open mapping theorem.

24. Let f, g be analytic on U_i, the inside of Γ, a rectifiable simple closed curve positively oriented with parameterization γ. Suppose either

$$|f(z) + g(z)| < |f(z)| + |g(z)| \text{ on } \Gamma$$

or

$$|f(z) - g(z)| < |f(z)| \text{ on } \Gamma$$

Let Z_f denote the number of zeros in U_i and let Z_g denote the number of zeros of g in U_i. Then neither f, g, nor f/g can equal zero anywhere on Γ and $Z_f = Z_g$. **Hint:** The first condition implies for all $z \in \Gamma$,

$$\frac{f(z)}{g(z)} \in \mathbb{C} \setminus [0, \infty)$$

Show there exists a primitive F for

$$\frac{(f/g)'}{f/g}.$$

and argue

$$0 = \int_\gamma \frac{(f/g)'}{f/g} dz = \int_\gamma \frac{f'}{g'} dz - \int_\gamma \frac{g'}{g} dz = Z_f - Z_g.$$

You could consider $F = L(f/g)$ where L is the analytic function defined on $\mathbb{C} \setminus [0, \infty)$ with the property that

$$e^{L(z)} = z.$$

Thus

$$e^{L(z)} L'(z) = 1, \; L'(z) = 1/z.$$

In the second case, show $g/f \notin (-\infty, 0]$ and so a similar thing can be done. This problem is a case of Rouche's theorem.

25. Use the result of Problem 24 to give another proof of the fundamental theorem of algebra as follows. Let $g(z)$ be a polynomial of degree n, $a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$ where $a_n \neq 0$. Now let $f(z) = a_n z^n$. Let Γ be a big circle, large enough that $|f(z) - g(z)| < |f(z)|$ on this circle. Then tell why g and f have the same number of zeros where they are counted according to multiplicity.

26. Let $p(z) = z^7 + 11z^3 - 5z^2 + 5$. Identify a ball $B(0,r)$ which must contain all the zeros of $p(z)$. Try to make r reasonably small. Use Problem 24.

27. Here is another approach to the open mapping theorem which I think might be a little easier and shorter which is based on Rouche's theorem and makes no reference to real variable techniques. Let $f : U \to \mathbb{C}$ where U is an open connected set. Then $f(U)$ is either an open connected set or a point. **Hint:** Suppose $f(U)$ is not a point. Then explain why for any $z_0 \in U$ there exists $r > 0$ such that

$$f(z) - f(z_0) = g(z)(z - z_0)^m$$

where g is analytic and nonzero on $\overline{B(z_0, r)}$. Now consider the function $z \to f(z) - w$. I would like to use Rouche's theorem to claim this function has the same number of zeros, namely m as the function $z \to f(z) - f(z_0)$. Let

$$\delta = \min\{|f(z) - f(z_0)| : |z - z_0| = r\}$$

Then if $|w - f(z_0)| < \delta$,

$$|w - f(z_0)| = |f(z) - f(z_0) - (f(z) - w)| < \delta \leq |f(z) - f(z_0)|$$

for each $z \in \partial B(z_0, r)$ and so you can apply Rouche's theorem. What does this say about when f is one to one? Why is $f(U)$ open? Why is $f(U)$ connected?

28. Let $\gamma : [a,b] \to \mathbb{C}$ be of bounded variation, $\gamma(a) = \gamma(b)$ and suppose $z \notin \gamma^*$. Define

$$n(\gamma, z) \equiv \frac{1}{2\pi i} \int_\gamma \frac{dw}{w - z}.$$

This is called the winding number. When γ^* is positively oriented and a simple closed curve, this number equals 1 by the Cauchy integral formula. However, it is always an integer. Furthermore, $z \to n(\gamma, z)$ is continuous and so is constant on every component of $\mathbb{C} \setminus \gamma^*$. For z in the unbounded component, $n(\gamma, z) = 0$. Most modern treatments of complex analysis feature the winding number extensively in the statement of all the major theorems. This is because it makes possible the most general form of the theorems. Prove the above properties of the winding number. **Hint:** The continuity is easy. It follows right away from a simple estimate and Theorem 14.2.4 applied to contour integrals. The tricky part is in showing it is an integer. This is where it is convenient to use Theorem 14.2.6 applied to contour integrals. There exists $\eta : [a,b] \to \mathbb{C}$ which is C^1 on $[a,b]$ and

$$\max\{|\eta(t) - \gamma(t)| : t \in [a,b]\} < \varepsilon,$$

$$\eta(a) = \eta(b) = \gamma(a) = \gamma(b)$$

$$\left| \frac{1}{2\pi i} \int_\gamma \frac{dw}{w - z} - \frac{1}{2\pi i} \int_\eta \frac{dw}{w - z} \right| < \varepsilon$$

where $\varepsilon < \operatorname{dist}(z, \gamma^*)$. Thus $z \notin \eta^*$. Consider the contour integral which involves η and show it is an integer. Then there exists a sequence of these C^1 contours $\{\eta_k\}$ such that

$$\left| \frac{1}{2\pi i} \int_\gamma \frac{dw}{w - z} - \frac{1}{2\pi i} \int_{\eta_k} \frac{dw}{w - z} \right| \to 0.$$

Consequently, for all k large enough there can be no change in

$$\frac{1}{2\pi i} \int_{\eta_k} \frac{dw}{w - z}$$

which shows

$$\frac{1}{2\pi i}\int_{\gamma}\frac{dw}{w-z}$$

is an integer as claimed. So how do you show the contour integral involving η yields an integer? As mentioned above,

$$\frac{1}{2\pi i}\int_{\eta_k}\frac{dw}{w-z}=\frac{1}{2\pi i}\int_a^b\frac{\eta'(t)}{\eta(t)-z}dt$$

Let

$$g(t)\equiv\int_a^t\frac{\eta'(s)}{\eta(s)-z}ds$$

Formally this is a lot like some sort of $\log(\eta(s)-z)$ (recall beginning calculus) so it is reasonable to consider

$$\left(\frac{e^{g(t)}}{\eta(t)-z}\right)'.$$

Show this equals 0. Explain why this requires the function which is differentiated must be constant. Thus

$$\frac{e^{g(a)}}{\eta(a)-z}=\frac{e^{g(b)}}{\eta(b)-z}$$

Now $\eta(a)=\eta(b), g(a)=0$, and so $e^{g(a)}=1=e^{g(b)}$. Explain why this requires $g(b)=2m\pi i$ for m an integer. Now this gives the desired result.

29. Let

$$B'(a,r)\equiv\{z\in\mathbb{C}\text{ such that }0<|z-a|<r\}.$$

Thus this is the usual ball without the center. A function is said to have an isolated singularity at the point $a\in\mathbb{C}$ if f is analytic on $B'(a,r)$ for some $r>0$.

An isolated singularity of f is said to be removable if there exists an analytic function, g analytic at a and near a such that $f=g$ at all points near a. A major theorem is the following.

Theorem 14.7.1 *Let $f:B'(a,r)\to X$ be analytic. Thus f has an isolated singularity at a. Suppose also that*

$$\lim_{z\to a}f(z)(z-a)=0.$$

Then there exists a unique analytic function, $g:B(a,r)\to X$ such that $g=f$ on $B'(a,r)$. Thus the singularity at a is removable.

Prove this theorem. **Hint:** Let $h(z)=f(z)(z-a)^2$. Then $h(a)=0$ and $h'(a)$ exists and equals 0. Show this. Also h is analytic near a. Therefore,

$$h(z)=\sum_{k=2}^{\infty}a_k(z-a)^k$$

Maybe consider $g(z)=h(z)/(z-a)^2$. Argue g is analytic and equals f for z near a.

30. Another really amazing theorem in complex analysis is the Casorati Weierstrass theorem.

Theorem 14.7.2 *Let a be an isolated singularity and suppose for some $r > 0$, $f(B'(a,r))$ is not dense in $\mathbb{C}$. Then either a is a removable singularity or there exist finitely many $b_1, \cdots, b_M$ for some finite number, M such that for z near a,*

$$f(z) = g(z) + \sum_{k=1}^{M} \frac{b_{-k}}{(z-a)^k} \tag{14.49}$$

where $g(z)$ is analytic near a. When the above formula holds, f is said to have a pole of order M at a.

Prove this theorem. **Hint:** Suppose a is not removable and $B(z_0, \delta)$ has no points of $f(B'(a,r))$. Such a ball must exist if $f(B'(a,r))$ is not dense in the plane. this means that for all $0 < |z-a| < r$,

$$|f(z) - z_0| \geq \delta > 0$$

Hence

$$\lim_{z \to a} \frac{1}{f(z) - z_0} (z-a) = 0$$

and so $1/(f(z) - z_0)$ has a removable sincularity at a. See Problem 29. Let $g(z)$ be analytic at and near a and agree with this function. Thus

$$g(z) = \sum_{n=0}^{\infty} a_n (z-a)^n .$$

There are two cases, $g(a) = 0$ and $g(a) \neq 0$. First suppose $g(a) = 0$. Then explain why

$$g(z) = h(z)(z-a)^m$$

where $h(z)$ is analytic and non zero near a. Then

$$f(z) - z_0 = \frac{1}{h(z)} \frac{1}{(z-a)^m}$$

Show this yields the desired conclusion. Next suppose $g(a) \neq 0$. Then explain why $g(z) \neq 0$ near a and this would contradict the assertion that a is not removable.

31. One of the very important techniques in complex analysis is the method of residues. When a is a pole the residue of f at a denoted by $\operatorname{res}(f,a)$, is defined as b_{-1} in 14.49. Suppose a is a pole and Γ is a simple closed rectifiable curve containing a on the inside with no other singular points on Γ or anywhere else inside Γ. Show that under these conditions,

$$\int_\Gamma f(z)\, dz = 2\pi i (\operatorname{res}(f,a))$$

Also describe a way to find $\operatorname{res}(f,a)$ by multiplying by $(z-a)^m$ and differentiating. **Hint:** You should show $\int_\Gamma \frac{1}{(z-a)^m} dz = 0$ whenever $m > 1$. This is because the function has a primitive.

32. Using Problem 9 give a holy version of the Cauchy integral theorem. This is it. Let Γ be a positively oriented rectifiable simple closed curve with inside U_i and remove finitely many open discs $B(z_j, r_j)$ from U_i. Thus the result is a holy

region. Suppose f is analytic on some open set containing $\overline{U_i} \setminus \cup_{j=1}^{n} B(z_j, r_j)$. Then letting Γ_j denote the negatively oriented boundary of $B(z_j, r_j)$, show

$$0 = \int_\gamma f(z)\,dz + \sum_{j=1}^{n} \int_{\gamma_j} f(z)\,dz$$

where γ_j is a parameterization for Γ_j. **Hint:** The proof is the same as given earlier. You just use Green's theorem.

33. Let Γ be a simple closed curve and suppose on its inside there are finitely many poles for a function f which is analytic near Γ. Call these poles $\{z_k\}_{k=1}^{n}$. Then

$$\int_\gamma f(z)\,dz = 2\pi i \sum_{j=1}^{n} \operatorname{res}(f, z_j)$$

This is the very important residue theorem for computing line integrals. **Hint:** You should use Problem 32 and Problem 30, the Casorati Weierstrass theorem.

Chapter 15

Hausdorff Measures

15.1 Definition Of Hausdorff Measures

This chapter is on Hausdorff measures. First I will discuss some outer measures. In all that is done here, $\alpha(n)$ will be the volume of the ball in $\mathbb{R}^n$ which has radius 1. This volume is the usual Lebesgue measure and the balls will be determined by the usual norm on $\mathbb{R}^n$.

Definition 15.1.1 *For a set, E, denote by $r(E)$ the number which is half the diameter of E. Thus*

$$r(E) \equiv \frac{1}{2}\sup\{|\mathbf{x}-\mathbf{y}| : \mathbf{x},\mathbf{y} \in E\} \equiv \frac{1}{2}\operatorname{diam}(E)$$

Let $E \subseteq \mathbb{R}^n$.

$$\mathcal{H}^s_\delta(E) \equiv \inf\{\sum_{j=1}^{\infty} \beta(s)(r(C_j))^s : E \subseteq \cup_{j=1}^{\infty} C_j, r(C_j) \leq \delta\}$$

$$\mathcal{H}^s(E) \equiv \lim_{\delta \to 0} \mathcal{H}^s_\delta(E).$$

In the above definition, $\beta(s)$ is an appropriate positive constant depending on s. Later I will tell what this constant is but it is not important for now. It will be chosen in such a way that whenever n is a positive integer, $\mathcal{H}^n([0,1)^n) = 1 = m_n([0,1)^n)$. In fact, this is all you need to know about it.

Lemma 15.1.2 *$\mathcal{H}^s$ and $\mathcal{H}^s_\delta$ are outer measures.*

Proof: It is clear that $\mathcal{H}^s(\emptyset) = 0$ and if $A \subseteq B$, then $\mathcal{H}^s(A) \leq \mathcal{H}^s(B)$ with similar assertions valid for $\mathcal{H}^s_\delta$. Suppose $E = \cup_{i=1}^{\infty} E_i$ and $\mathcal{H}^s_\delta(E_i) < \infty$ for each i. Let $\{C^i_j\}_{j=1}^{\infty}$ be a covering of E_i with

$$\sum_{j=1}^{\infty} \beta(s)(r(C^i_j))^s - \varepsilon/2^i < \mathcal{H}^s_\delta(E_i)$$

and $\mathrm{diam}(C_j^i) \leq \delta$. Then

$$\begin{aligned}\mathcal{H}_\delta^s(E) &\leq \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\beta(s)(r(C_j^i))^s \\ &\leq \sum_{i=1}^{\infty}\mathcal{H}_\delta^s(E_i)+\varepsilon/2^i \\ &\leq \varepsilon+\sum_{i=1}^{\infty}\mathcal{H}_\delta^s(E_i).\end{aligned}$$

It follows that since $\varepsilon > 0$ is arbitrary,

$$\mathcal{H}_\delta^s(E) \leq \sum_{i=1}^{\infty}\mathcal{H}_\delta^s(E_i)$$

which shows $\mathcal{H}_\delta^s$ is an outer measure. Now notice that $\mathcal{H}_\delta^s(E)$ is increasing as $\delta \to 0$. Picking a sequence δ_k decreasing to 0, the monotone convergence theorem implies

$$\mathcal{H}^s(E) \leq \sum_{i=1}^{\infty}\mathcal{H}^s(E_i).$$

This proves the lemma. ■

The outer measure $\mathcal{H}^s$ is called s dimensional Hausdorff measure when restricted to the σ algebra of $\mathcal{H}^s$ measurable sets. Recall these are the sets E such that for all S,

$$\mathcal{H}^s(S) = \mathcal{H}^s(S\cap E)+\mathcal{H}^s(S\setminus E).$$

Next I will show the σ algebra of $\mathcal{H}^s$ measurable sets includes the Borel sets. This is done by the following very interesting condition known as Caratheodory's criterion.

15.1.1 Properties Of Hausdorff Measure

Definition 15.1.3 *For two sets A, B in a metric space, define*

$$\mathrm{dist}(A,B) \equiv \inf\{||\mathbf{x}-\mathbf{y}|| : \mathbf{x}\in A, \mathbf{y}\in B\}.$$

Theorem 15.1.4 *Let μ be an outer measure on the subsets of X, a closed subset of a normed vector space and suppose*

$$\mu(A\cup B) = \mu(A)+\mu(B)$$

whenever $\mathrm{dist}(A,B) > 0$, *then the σ algebra of measurable sets contains the Borel sets.*

Proof: It suffices to show that closed sets are in $\mathcal{F}$, the σ-algebra of measurable sets, because then the open sets are also in $\mathcal{F}$ and consequently $\mathcal{F}$ contains the Borel sets. Let K be closed and let S be a subset of Ω. Is $\mu(S) \geq \mu(S\cap K)+\mu(S\setminus K)$? It suffices to assume $\mu(S) < \infty$. Let

$$K_n \equiv \{x : \mathrm{dist}(x,K) \leq \frac{1}{n}\}$$

By Lemma 7.4.4 on Page 167, $x \to \mathrm{dist}(x,K)$ is continuous and so K_n is closed. By the assumption of the theorem,

$$\mu(S) \geq \mu((S\cap K)\cup(S\setminus K_n)) = \mu(S\cap K)+\mu(S\setminus K_n) \tag{15.1}$$

since $S \cap K$ and $S \setminus K_n$ are a positive distance apart. Now

$$\mu(S \setminus K_n) \leq \mu(S \setminus K) \leq \mu(S \setminus K_n) + \mu((K_n \setminus K) \cap S). \tag{15.2}$$

If $\lim_{n\to\infty} \mu((K_n \setminus K) \cap S) = 0$ then the theorem will be proved because this limit along with 15.2 implies $\lim_{n\to\infty} \mu(S \setminus K_n) = \mu(S \setminus K)$ and then taking a limit in 15.1, $\mu(S) \geq \mu(S \cap K) + \mu(S \setminus K)$ as desired. Therefore, it suffices to establish this limit.

Since K is closed, a point, $x \notin K$ must be at a positive distance from K and so

$$K_n \setminus K = \cup_{k=n}^{\infty} K_k \setminus K_{k+1}.$$

Therefore

$$\mu(S \cap (K_n \setminus K)) \leq \sum_{k=n}^{\infty} \mu(S \cap (K_k \setminus K_{k+1})). \tag{15.3}$$

If

$$\sum_{k=1}^{\infty} \mu(S \cap (K_k \setminus K_{k+1})) < \infty, \tag{15.4}$$

then $\mu(S \cap (K_n \setminus K)) \to 0$ because it is dominated by the tail of a convergent series so it suffices to show 15.4.

$$\sum_{k=1}^{M} \mu(S \cap (K_k \setminus K_{k+1})) =$$

$$\sum_{k \text{ even},\ k\leq M} \mu(S \cap (K_k \setminus K_{k+1})) + \sum_{k \text{ odd},\ k\leq M} \mu(S \cap (K_k \setminus K_{k+1})). \tag{15.5}$$

By the construction, the distance between any pair of sets $S \cap (K_k \setminus K_{k+1})$ for different even values of k is positive and the distance between any pair of sets $S \cap (K_k \setminus K_{k+1})$ for different odd values of k is positive. Therefore,

$$\sum_{k \text{ even},\ k\leq M} \mu(S \cap (K_k \setminus K_{k+1})) + \sum_{k \text{ odd},\ k\leq M} \mu(S \cap (K_k \setminus K_{k+1})) \leq$$

$$\mu(\bigcup_{k \text{ even}} S \cap (K_k \setminus K_{k+1})) + \mu(\bigcup_{k \text{ odd}} S \cap (K_k \setminus K_{k+1})) \leq 2\mu(S) < \infty$$

and so for all M, $\sum_{k=1}^{M} \mu(S \cap (K_k \setminus K_{k+1})) \leq 2\mu(S)$ showing 15.4 and proving the theorem. ■

The next theorem applies the Caratheodory criterion above to $\mathcal{H}^s$.

Theorem 15.1.5 *The σ algebra of $\mathcal{H}^s$ measurable sets contains the Borel sets and $\mathcal{H}^s$ has the property that for all $E \subseteq \mathbb{R}^n$, there exists a Borel set $F \supseteq E$ such that $\mathcal{H}^s(F) = \mathcal{H}^s(E)$.*

Proof: Let $\text{dist}(A, B) = 2\delta_0 > 0$. Is it the case that

$$\mathcal{H}^s(A) + \mathcal{H}^s(B) = \mathcal{H}^s(A \cup B)?$$

This is what is needed to use Caratheodory's criterion.

Let $\{C_j\}_{j=1}^{\infty}$be a covering of $A \cup B$ such that $r(C_j) \leq \delta < \delta_0/2$ for each j and

$$\mathcal{H}^s_\delta(A \cup B) + \varepsilon > \sum_{j=1}^{\infty} \beta(s)(r(C_j))^s.$$

Thus

$$\mathcal{H}_\delta^s(A \cup B)\dot{} + \varepsilon > \sum_{j\in J_1} \beta(s)(r(C_j))^s + \sum_{j\in J_2} \beta(s)(r(C_j))^s$$

where

$$J_1 = \{j : C_j \cap A \neq \emptyset\},\ J_2 = \{j : C_j \cap B \neq \emptyset\}.$$

Recall $\text{dist}(A,B) = 2\delta_0$ and so $J_1 \cap J_2 = \emptyset$. It follows

$$\mathcal{H}_\delta^s(A \cup B) + \varepsilon > \mathcal{H}_\delta^s(A) + \mathcal{H}_\delta^s(B).$$

Letting $\delta \to 0$, and noting $\varepsilon > 0$ was arbitrary, yields

$$\mathcal{H}^s(A \cup B) \geq \mathcal{H}^s(A) + \mathcal{H}^s(B).$$

Equality holds because $\mathcal{H}^s$ is an outer measure. By Caratheodory's criterion, $\mathcal{H}^s$ is a Borel measure.

To verify the second assertion, note first there is no loss of generality in letting $\mathcal{H}^s(E) < \infty$. Let

$$E \subseteq \cup_{j=1}^\infty C_j,\ r(C_j) < \delta,$$

and

$$\mathcal{H}_\delta^s(E) + \delta > \sum_{j=1}^\infty \beta(s)(r(C_j))^s.$$

Let

$$F_\delta = \cup_{j=1}^\infty \overline{C_j}.$$

Thus $F_\delta \supseteq E$ and

$$\begin{aligned} \mathcal{H}_\delta^s(E) &\leq \mathcal{H}_\delta^s(F_\delta) \leq \sum_{j=1}^\infty \beta(s)(r(\overline{C_j}))^s \\ &= \sum_{j=1}^\infty \beta(s)(r(C_j))^s < \delta + \mathcal{H}_\delta^s(E). \end{aligned}$$

Let $\delta_k \to 0$ and let $F = \cap_{k=1}^\infty F_{\delta_k}$. Then $F \supseteq E$ and

$$\mathcal{H}_{\delta_k}^s(E) \leq \mathcal{H}_{\delta_k}^s(F) \leq \mathcal{H}_{\delta_k}^s(F_\delta) \leq \delta_k + \mathcal{H}_{\delta_k}^s(E).$$

Letting $k \to \infty$,

$$\mathcal{H}^s(E) \leq \mathcal{H}^s(F) \leq \mathcal{H}^s(E)$$

This proves the theorem. ■

A measure satisfying the first conclusion of Theorem 15.1.5 is sometimes called a Borel regular measure.

15.1.2 $\mathcal{H}^n$ And m_n

Next I will compare $\mathcal{H}^n$ and m_n. To do this, recall the following covering theorem which is a summary of Corollaries 9.7.5 and 9.7.4 found on Page 236.

Theorem 15.1.6 *Let $E \subseteq \mathbb{R}^n$ and let $\mathcal{F}$, be a collection of balls of bounded radii such that $\mathcal{F}$ covers E in the sense of Vitali. Then there exists a countable collection of disjoint balls from $\mathcal{F}$, $\{B_j\}_{j=1}^\infty$, such that $\overline{m_n}(E \setminus \cup_{j=1}^\infty B_j) = 0$.*

In the next lemma, the balls are the usual balls taken with respect to the usual distance in $\mathbb{R}^n$.

Lemma 15.1.7 *If $m_n(S) = 0$ then $\mathcal{H}^n(S) = \mathcal{H}^n_\delta(S) = 0$. Also, there exists a constant, k such that $\mathcal{H}^n(E) \leq k m_n(E)$ for all E Borel. Also, if $Q_0 \equiv [0,1)^n$, the unit cube, then $\mathcal{H}^n([0,1)^n) > 0$.*

Proof: Suppose first $m_n(S) = 0$. First suppose S is bounded. Then by outer regularity, there exists a bounded open V containing S and $m_n(V) < \varepsilon$. For each $\mathbf{x} \in S$, there exists a ball $B_{\mathbf{x}}$ such that $\widehat{B_{\mathbf{x}}} \subseteq V$ and $\delta > r\left(\widehat{B_{\mathbf{x}}}\right)$. By the Vitali covering theorem there is a sequence of disjoint balls $\{B_k\}$ such that $\left\{\widehat{B_k}\right\}$ covers S. Then letting $\alpha(n)$ be the Lebesgue measure of the unit ball in $\mathbb{R}^n$

$$\begin{aligned}\mathcal{H}^n_\delta(S) &\leq \sum_k \beta(n)\, r\left(\widehat{B_k}\right)^n = \frac{\beta(n)}{\alpha(n)} 5^n \sum_k \alpha(n)\, r(B_k)^n \\ &\leq \frac{\beta(n)}{\alpha(n)} 5^n m_n(V) < \frac{\beta(n)}{\alpha(n)} 5^n \varepsilon\end{aligned}$$

Since ε is arbitrary, this shows $\mathcal{H}^n_\delta(S) = 0$ and now it follows $\mathcal{H}^n(S) = 0$. In case S is not bounded, let $S_m = B(\mathbf{0},m) \cap S$. Then $\mathcal{H}^n_\delta(S_m) = 0$ and so letting $m \to \infty, \mathcal{H}^n_\delta(S) = 0$ also. Then as before, $\mathcal{H}^n(S) = 0$.

Letting U be an open set and $\delta > 0$, consider all balls, B contained in U which have diameters less than δ. This is a Vitali covering of U and therefore by Theorem 15.1.6, there exists $\{B_i\}$, a sequence of disjoint balls of radii less than δ contained in U such that $\cup_{i=1}^\infty B_i$ differs from U by a set of Lebesgue measure zero. Let $\alpha(n)$ be the Lebesgue measure of the unit ball in $\mathbb{R}^n$. Then from what was just shown,

$$\begin{aligned}\mathcal{H}^n_\delta(U) &= \mathcal{H}^n_\delta(\cup_i B_i) \leq \sum_{i=1}^\infty \beta(n)\, r(B_i)^n = \frac{\beta(n)}{\alpha(n)} \sum_{i=1}^\infty \alpha(n)\, r(B_i)^n \\ &= \frac{\beta(n)}{\alpha(n)} \sum_{i=1}^\infty m_n(B_i) = \frac{\beta(n)}{\alpha(n)} m_n(U) \equiv k m_n(U).\end{aligned}$$

Now letting E be Borel, it follows from the outer regularity of m_n there exists a decreasing sequence of open sets $\{V_i\}$ containing E such such that $m_n(V_i) \to m_n(E)$. Then from the above,

$$\mathcal{H}^n_\delta(E) \leq \lim_{i\to\infty} \mathcal{H}^n_\delta(V_i) \leq \lim_{i\to\infty} k m_n(V_i) = k m_n(E).$$

Since $\delta > 0$ is arbitrary, it follows that also

$$\mathcal{H}^n(E) \leq k m_n(E).$$

This proves the first part of the lemma.

To verify the second part, note that it is obvious $\mathcal{H}^n_\delta$ and $\mathcal{H}^n$ are translation invariant because diameters of sets do not change when translated. Therefore, if $\mathcal{H}^n([0,1)^n) = 0$, it follows $\mathcal{H}^n(\mathbb{R}^n) = 0$ because $\mathbb{R}^n$ is the countable union of translates of $Q_0 \equiv [0,1)^n$. Since each $\mathcal{H}^n_\delta$ is no larger than $\mathcal{H}^n$, the same must hold for $\mathcal{H}^n_\delta$. Therefore, there exists a sequence of sets $\{C_i\}$ each having diameter less than δ such that the union of these sets equals $\mathbb{R}^n$ but

$$1 > \sum_{i=1}^\infty \beta(n)\, r(C_i)^n.$$

Now let B_i be a ball having radius equal to $\operatorname{diam}(C_i) = 2r(C_i)$ which contains C_i. It follows

$$m_n(B_i) = \alpha(n)\, 2^n r(C_i)^n = \frac{\alpha(n)\, 2^n}{\beta(n)} \beta(n)\, r(C_i)^n$$

which implies

$$1 > \sum_{i=1}^{\infty} \beta(n)\, r(C_i)^n = \sum_{i=1}^{\infty} \frac{\beta(n)}{\alpha(n)\, 2^n} m_n(B_i) = \infty,$$

a contradiction. This proves the lemma. ■

Lemma 15.1.8 *Every open set in $\mathbb{R}^n$ is the countable disjoint union of half open boxes of the form*

$$\prod_{i=1}^{n} (a_i, a_i + 2^{-k}]$$

where $a_i = l2^{-k}$ for some integers, l, k. The sides of these boxes are of equal length. One could also have half open boxes of the form

$$\prod_{i=1}^{n} [a_i, a_i + 2^{-k})$$

and the conclusion would be unchanged.

Proof: Let

$$\mathcal{C}_k = \{\text{All half open boxes } \prod_{i=1}^{n} (a_i, a_i + 2^{-k}] \text{ where}$$

$$a_i = l2^{-k} \text{ for some integer } l.\}$$

Thus $\mathcal{C}_k$ consists of a countable disjoint collection of boxes whose union is $\mathbb{R}^n$. This is sometimes called a tiling of $\mathbb{R}^n$. Think of tiles on the floor of a bathroom and you will get the idea. Note that each box has diameter no larger than $2^{-k}\sqrt{n}$. This is because if

$$\mathbf{x}, \mathbf{y} \in \prod_{i=1}^{n} (a_i, a_i + 2^{-k}],$$

then $|x_i - y_i| \leq 2^{-k}$. Therefore,

$$|\mathbf{x} - \mathbf{y}| \leq \left(\sum_{i=1}^{n} \left(2^{-k}\right)^2 \right)^{1/2} = 2^{-k}\sqrt{n}.$$

Let U be open and let $\mathcal{B}_1 \equiv$ all sets of $\mathcal{C}_1$ which are contained in U. If $\mathcal{B}_1, \cdots, \mathcal{B}_k$ have been chosen, $\mathcal{B}_{k+1} \equiv$ all sets of $\mathcal{C}_{k+1}$ contained in

$$U \setminus \cup \left(\cup_{i=1}^{k} \mathcal{B}_i \right).$$

Let $\mathcal{B}_\infty = \cup_{i=1}^{\infty} \mathcal{B}_i$. In fact $\cup \mathcal{B}_\infty = U$. Clearly $\cup \mathcal{B}_\infty \subseteq U$ because every box of every $\mathcal{B}_i$ is contained in U. If $p \in U$, let k be the smallest integer such that p is contained in a box from $\mathcal{C}_k$ which is also a subset of U. Thus

$$p \in \cup \mathcal{B}_k \subseteq \cup \mathcal{B}_\infty.$$

Hence $\mathcal{B}_\infty$ is the desired countable disjoint collection of half open boxes whose union is U. The last assertion about the other type of half open rectangle is obvious. This proves the lemma. ■

Theorem 15.1.9 *By choosing $\beta(n)$ properly, one can obtain $\mathcal{H}^n = m_n$ on all Lebesgue measurable sets.*

Proof: I will show $\mathcal{H}^n$ is a positive multiple of m_n for any choice of $\beta(n)$. Define

$$k = \frac{m_n(Q_0)}{\mathcal{H}^n(Q_0)}$$

where $Q_0 = [0,1)^n$ is the half open unit cube in $\mathbb{R}^n$. I will show $k\mathcal{H}^n(E) = m_n(E)$ for any Lebesgue measurable set. When this is done, it will follow that by adjusting $\beta(n)$ the multiple can be taken to be 1.

Let $Q = \prod_{i=1}^n [a_i, a_i + 2^{-k})$ be a half open box where $a_i = l2^{-k}$. Thus Q_0 is the union of $\left(2^k\right)^n$ of these identical half open boxes. By translation invariance, of $\mathcal{H}^n$ and m_n

$$\left(2^k\right)^n \mathcal{H}^n(Q) = \mathcal{H}^n(Q_0) = \frac{1}{k} m_n(Q_0) = \frac{1}{k}\left(2^k\right)^n m_n(Q).$$

Therefore, $k\mathcal{H}^n(Q) = m_n(Q)$ for any such half open box and by translation invariance, for the translation of any such half open box. It follows from Lemma 15.1.8 that $k\mathcal{H}^n(U) = m_n(U)$ for all open sets. It follows immediately, since every compact set is the countable intersection of open sets that $k\mathcal{H}^n = m_n$ on compact sets. Therefore, they are also equal on all closed sets because every closed set is the countable union of compact sets. Now let F be an arbitrary Lebesgue measurable set. I will show that F is $\mathcal{H}^n$ measurable and that $k\mathcal{H}^n(F) = m_n(F)$. Let $F_l = B(\mathbf{0}, l) \cap F$. Then there exists H a countable union of compact sets and G a countable intersection of open sets such that

$$H \subseteq F_l \subseteq G \tag{15.6}$$

and $m_n(G \setminus H) = 0$ which implies by Lemma 15.1.7

$$m_n(G \setminus H) = k\mathcal{H}^n(G \setminus H) = 0. \tag{15.7}$$

To do this, let $\{G_i\}$ be a decreasing sequence of bounded open sets containing F_l and let $\{H_i\}$ be an increasing sequence of compact sets contained in F_l such that

$$k\mathcal{H}^n(G_i \setminus H_i) = m_n(G_i \setminus H_i) < 2^{-i}$$

Then letting $G = \cap_i G_i$ and $H = \cup_i H_i$ this establishes 15.6 and 15.7. Then by completeness of $\mathcal{H}^n$ it follows F_l is $\mathcal{H}^n$ measurable and

$$k\mathcal{H}^n(F_l) = k\mathcal{H}^n(H) = m_n(H) = m_n(F_l).$$

Now taking $l \to \infty$, it follows F is $\mathcal{H}^n$ measurable and $k\mathcal{H}^n(F) = m_n(F)$. Therefore, adjusting $\beta(n)$ it can be assumed the constant, k is 1. This proves the theorem. ■

The exact determination of $\beta(n)$ is more technical. You can skip it if you want. Just remember $\beta(n)$ is chosen such that $\mathcal{H}^n([0,1)^n) = 1$. It turns out this will require $\beta(n) = \alpha(n)$ where $\alpha(n)$ is the volume of the unit ball taken with respect to the usual norm. The optional sections are starred.

15.2 Technical Considerations*

Let $\alpha(n)$ be the volume of the unit ball in $\mathbb{R}^n$. Thus the volume of $B(\mathbf{0}, r)$ in $\mathbb{R}^n$ is $\alpha(n)r^n$ from the change of variables formula. There is a very important and interesting inequality known as the isodiametric inequality which says that if A is any set in $\mathbb{R}^n$, then

$$\overline{m}(A) \leq \alpha(n)(2^{-1}\mathrm{diam}(A))^n.$$

This inequality may seem obvious at first but it is not really. The reason it is not is that there are sets which are not subsets of any sphere having the same diameter as the set. For example, consider an equilateral triangle.

Lemma 15.2.1 *Let $f : \mathbb{R}^{n-1} \to [0,\infty)$ be Borel measurable and let*

$$S = \{(\mathbf{x},y) : |y| < f(\mathbf{x})\}.$$

Then S is a Borel set in $\mathbb{R}^n$.

Proof: Set s_k be an increasing sequence of Borel measurable functions converging pointwise to f.

$$s_k(\mathbf{x}) = \sum_{m=1}^{N_k} c_m^k \mathcal{X}_{E_m^k}(\mathbf{x}).$$

Let

$$S_k = \cup_{m=1}^{N_k} E_m^k \times (-c_m^k, c_m^k).$$

Then $(\mathbf{x},y) \in S_k$ if and only if $f(\mathbf{x}) > 0$ and $|y| < s_k(\mathbf{x}) \leq f(\mathbf{x})$. It follows that $S_k \subseteq S_{k+1}$ and

$$S = \cup_{k=1}^{\infty} S_k.$$

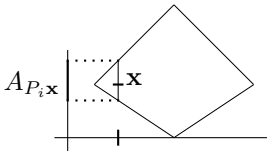
But each S_k is a Borel set and so S is also a Borel set. This proves the lemma. ■

Let P_i be the projection onto

$$\text{span}\,(\mathbf{e}_1, \cdots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \cdots, \mathbf{e}_n)$$

where the $\mathbf{e}_k$ are the standard basis vectors in $\mathbb{R}^n$, $\mathbf{e}_k$ being the vector having a 1 in the k^{th} slot and a 0 elsewhere. Thus $P_i\mathbf{x} \equiv \sum_{j\neq i} x_j \mathbf{e}_j$. Also let

$$A_{P_i\mathbf{x}} \equiv \{x_i : (x_1, \cdots, x_i, \cdots, x_n) \in A\}$$



$$P_i\mathbf{x} \in \text{span}\{\mathbf{e}_1, \cdots, \mathbf{e}_{i-1}\mathbf{e}_{i+1}, \cdots, \mathbf{e}_n\}.$$

Lemma 15.2.2 *Let $A \subseteq \mathbb{R}^n$ be a Borel set. Then $P_i\mathbf{x} \to m(A_{P_i\mathbf{x}})$ is a Borel measurable function defined on $P_i(\mathbb{R}^n)$.*

Proof: Let $\mathcal{K}$ be the π system consisting of sets of the form $\prod_{j=1}^n A_j$ where A_i is Borel. Also let $\mathcal{G}$ denote those Borel sets of $\mathbb{R}^n$ such that if $A \in \mathcal{G}$ then

$$P_i\mathbf{x} \to m((A \cap R_k)_{P_i\mathbf{x}}) \text{ is Borel measurable.}$$

where $R_k = (-k,k)^n$. Thus $\mathcal{K} \in \mathcal{G}$. If $A \in \mathcal{G}$

$$P_i\mathbf{x} \to m\left(\left(A^C \cap R_k\right)_{P_i\mathbf{x}}\right)$$

is Borel measurable because it is of the form

$$m\left((R_k)_{P_i\mathbf{x}}\right) - m\left((A \cap R_k)_{P_i\mathbf{x}}\right)$$

and these are Borel measurable functions of $P_i\mathbf{x}$. Also, if $\{A_i\}$ is a disjoint sequence of sets in $\mathcal{G}$ then

$$m\left((\cup_i A_i \cap R_k)_{P_i\mathbf{x}}\right) = \sum_i m\left((A_i \cap R_k)_{P_i\mathbf{x}}\right)$$

and each function of $P_i\mathbf{x}$ is Borel measurable. Thus by the lemma on π systems $\mathcal{G} = \mathcal{B}(\mathbb{R}^n)$ and This proves the lemma. ■

Now let $A \subseteq \mathbb{R}^n$ be Borel. Let P_i be the projection onto

$$\text{span}\left(\mathbf{e}_1, \cdots, \mathbf{e}_{i-1}, \mathbf{e}_{i+1}, \cdots, \mathbf{e}_n\right)$$

and as just described,

$$A_{P_i\mathbf{x}} = \{y \in \mathbb{R} \ : P_i\mathbf{x} + y\mathbf{e}_i \in A\}$$

Thus for $\mathbf{x} = (x_1, \cdots, x_n)$,

$$A_{P_i\mathbf{x}} = \{y \in \mathbb{R} : (x_1, \cdots, x_{i-1}, y, x_{i+1}, \cdots, x_n) \in A\}.$$

Since A is Borel, it follows from Lemma 15.2.1 that

$$P_i\mathbf{x} \to m(A_{P_i\mathbf{x}})$$

is a Borel measurable function on $P_i\mathbb{R}^n = \mathbb{R}^{n-1}$.

15.2.1 Steiner Symmetrization*

Define

$$S(A, \mathbf{e}_i) \equiv \{\mathbf{x} = P_i\mathbf{x} + y\mathbf{e}_i : |y| < 2^{-1}m(A_{P_i\mathbf{x}})\}$$

Lemma 15.2.3 *Let A be a Borel subset of $\mathbb{R}^n$. Then $S(A, \mathbf{e}_i)$ satisfies*

$$P_i\mathbf{x} + y\mathbf{e}_i \in S(A, \mathbf{e}_i) \text{ if and only if } P_i\mathbf{x} - y\mathbf{e}_i \in S(A, \mathbf{e}_i),$$

$$S(A, \mathbf{e}_i) \text{ is a Borel set in } \mathbb{R}^n,$$

$$m_n(S(A, \mathbf{e}_i)) = m_n(A), \tag{15.8}$$

$$diam(S(A, \mathbf{e}_i)) \leq \ diam(A). \tag{15.9}$$

Proof: The first assertion is obvious from the definition. The Borel measurability of $S(A, \mathbf{e}_i)$ follows from the definition and Lemmas 15.2.2 and 15.2.1. To show Formula 15.8,

$$\begin{aligned} m_n(S(A, \mathbf{e}_i)) &= \int_{P_i\mathbb{R}^n} \int_{-2^{-1}m(A_{P_i\mathbf{x}})}^{2^{-1}m(A_{P_i\mathbf{x}})} dx_i dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n \\ &= \int_{P_i\mathbb{R}^n} m(A_{P_i\mathbf{x}}) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n \\ &= m(A). \end{aligned}$$

Now suppose $\mathbf{x}_1$ and $\mathbf{x}_2 \in S(A, \mathbf{e}_i)$

$$\mathbf{x}_1 = P_i\mathbf{x}_1 + y_1\mathbf{e}_i, \ \mathbf{x}_2 = P_i\mathbf{x}_2 + y_2\mathbf{e}_i.$$

For $\mathbf{x} \in A$ define

$$l(\mathbf{x}) = \sup\{y : P_i\mathbf{x} + y\mathbf{e}_i \in A\}.$$

$$g(\mathbf{x}) = \inf\{y : P_i\mathbf{x} + y\mathbf{e}_i \in A\}.$$

Then it is clear that

$$l(\mathbf{x}_1) - g(\mathbf{x}_1) \geq m(A_{P_i\mathbf{x}_1}) \geq 2|y_1|, \tag{15.10}$$

$$l(\mathbf{x}_2) - g(\mathbf{x}_2) \geq m(A_{P_i\mathbf{x}_2}) \geq 2|y_2|. \tag{15.11}$$

Claim: $|y_1 - y_2| \leq |l(\mathbf{x}_1) - g(\mathbf{x}_2)|$ or $|y_1 - y_2| \leq |l(\mathbf{x}_2) - g(\mathbf{x}_1)|$.

Proof of Claim: If not,

$$\begin{aligned} 2|y_1 - y_2| &> |l(\mathbf{x}_1) - g(\mathbf{x}_2)| + |l(\mathbf{x}_2) - g(\mathbf{x}_1)| \\ &\geq |l(\mathbf{x}_1) - g(\mathbf{x}_1) + l(\mathbf{x}_2) - g(\mathbf{x}_2)| \\ &= l(\mathbf{x}_1) - g(\mathbf{x}_1) + l(\mathbf{x}_2) - g(\mathbf{x}_2). \\ &\geq 2\,|y_1| + 2\,|y_2| \end{aligned}$$

by 15.10 and 15.11 contradicting the triangle inequality.

Now suppose $|y_1 - y_2| \leq |l(\mathbf{x}_1) - g(\mathbf{x}_2)|$. From the claim,

$$\begin{aligned} |\mathbf{x}_1 - \mathbf{x}_2| &= (|P_i\mathbf{x}_1 - P_i\mathbf{x}_2|^2 + |y_1 - y_2|^2)^{1/2} \\ &\leq (|P_i\mathbf{x}_1 - P_i\mathbf{x}_2|^2 + |l(\mathbf{x}_1) - g(\mathbf{x}_2)|^2)^{1/2} \\ &\leq (|P_i\mathbf{x}_1 - P_i\mathbf{x}_2|^2 + (|z_1 - z_2| + 2\varepsilon)^2)^{1/2} \\ &\leq \operatorname{diam}(A) + O(\sqrt{\varepsilon}) \end{aligned}$$

where z_1 and z_2 are such that $P_i\mathbf{x}_1 + z_1\mathbf{e}_i \in A$, $P_i\mathbf{x}_2 + z_2\mathbf{e}_i \in A$, and

$$|z_1 - l(\mathbf{x}_1)| < \varepsilon \text{ and } |z_2 - g(\mathbf{x}_2)| < \varepsilon.$$

If $|y_1 - y_2| \leq |l(\mathbf{x}_2) - g(\mathbf{x}_1)|$, then we use the same argument but let

$$|z_1 - g(\mathbf{x}_1)| < \varepsilon \text{ and } |z_2 - l(\mathbf{x}_2)| < \varepsilon,$$

Since $\mathbf{x}_1, \mathbf{x}_2$ are arbitrary elements of $S(A, \mathbf{e}_i)$ and ε is arbitrary, this proves 15.9. ■

The next lemma says that if A is already symmetric with respect to the j^{th} direction, then this symmetry is not destroyed by taking $S(A, \mathbf{e}_i)$.

Lemma 15.2.4 *Suppose A is a Borel set in $\mathbb{R}^n$ such that $P_j\mathbf{x} + \mathbf{e}_j x_j \in A$ if and only if $P_j\mathbf{x} + (-x_j)\mathbf{e}_j \in A$. Then if $i \neq j$, $P_j\mathbf{x} + \mathbf{e}_j x_j \in S(A, \mathbf{e}_i)$ if and only if $P_j\mathbf{x} + (-x_j)\mathbf{e}_j \in S(A, \mathbf{e}_i)$.*

Proof: By definition,

$$P_j\mathbf{x} + \mathbf{e}_j x_j \in S(A, \mathbf{e}_i)$$

if and only if

$$|x_i| < 2^{-1} m(A_{P_i(P_j\mathbf{x} + \mathbf{e}_j x_j)}).$$

Now

$$x_i \in A_{P_i(P_j\mathbf{x} + \mathbf{e}_j x_j)}$$

if and only if

$$x_i \in A_{P_i(P_j\mathbf{x} + (-x_j)\mathbf{e}_j)}$$

by the assumption on A which says that A is symmetric in the $\mathbf{e}_j$ direction. Hence

$$P_j\mathbf{x} + \mathbf{e}_j x_j \in S(A, \mathbf{e}_i)$$

if and only if

$$|x_i| < 2^{-1} m(A_{P_i(P_j\mathbf{x} + (-x_j)\mathbf{e}_j)})$$

if and only if

$$P_j\mathbf{x} + (-x_j)\mathbf{e}_j \in S(A, \mathbf{e}_i).$$

This proves the lemma. ■

15.2.2 The Isodiametric Inequality*

The next theorem is called the isodiametric inequality. It is the key result used to compare Lebesgue and Hausdorff measures.

Theorem 15.2.5 *Let A be any Lebesgue measurable set in $\mathbb{R}^n$. Then*

$$m_n(A) \leq \alpha(n)(r(A))^n.$$

Proof: Suppose first that A is Borel. Let $A_1 = S(A, \mathbf{e}_1)$ and let $A_k = S(A_{k-1}, \mathbf{e}_k)$. Then by the preceding lemmas, A_n is a Borel set, $\operatorname{diam}(A_n) \leq \operatorname{diam}(A)$, $m_n(A_n) = m_n(A)$, and A_n is symmetric. Thus $\mathbf{x} \in A_n$ if and only if $-\mathbf{x} \in A_n$. It follows that

$$A_n \subseteq \overline{B(\mathbf{0}, r(A_n))}.$$

(If $\mathbf{x} \in A_n \backslash \overline{B(\mathbf{0}, r(A_n))}$, then $-\mathbf{x} \in A_n \backslash \overline{B(\mathbf{0}, r(A_n))}$ and so $\operatorname{diam}(A_n) \geq 2|\mathbf{x}| > \operatorname{diam}(A_n)$.) Therefore,

$$m_n(A_n) \leq \alpha(n)(r(A_n))^n \leq \alpha(n)(r(A))^n.$$

It remains to establish this inequality for arbitrary measurable sets. Letting A be such a set, let $\{K_n\}$ be an increasing sequence of compact subsets of A such that

$$m(A) = \lim_{k\to\infty} m(K_k).$$

Then

$$\begin{aligned} m(A) &= \lim_{k\to\infty} m(K_k) \leq \limsup_{k\to\infty} \alpha(n)(r(K_k))^n \\ &\leq \alpha(n)(r(A))^n. \end{aligned}$$

This proves the theorem. ■

15.2.3 The Proper Value Of $\beta(n)$*

I will show that the proper determination of $\beta(n)$ is $\alpha(n)$, the volume of the unit ball. Since $\beta(n)$ has been adjusted such that $k = 1$, $m_n(B(\mathbf{0},1)) = \mathcal{H}^n(B(\mathbf{0},1))$. There exists a covering of $B(\mathbf{0},1)$ of sets of radii less than δ, $\{C_i\}_{i=1}^{\infty}$ such that

$$\mathcal{H}^n_\delta(B(\mathbf{0},1)) + \varepsilon > \sum_i \beta(n)\, r(C_i)^n$$

Then by Theorem 15.2.5, the isodiametric inequality,

$$\begin{aligned} \mathcal{H}^n_\delta(B(\mathbf{0},1)) + \varepsilon &> \sum_i \beta(n)\, r(C_i)^n = \frac{\beta(n)}{\alpha(n)} \sum_i \alpha(n)\, r\left(\overline{C}_i\right)^n \\ &\geq \frac{\beta(n)}{\alpha(n)} \sum_i m_n\left(\overline{C}_i\right) \geq \frac{\beta(n)}{\alpha(n)} m_n(B(\mathbf{0},1)) = \frac{\beta(n)}{\alpha(n)} \mathcal{H}^n(B(\mathbf{0},1)) \end{aligned}$$

Now taking the limit as $\delta \to 0$,

$$\mathcal{H}^n(B(\mathbf{0},1)) + \varepsilon \geq \frac{\beta(n)}{\alpha(n)} \mathcal{H}^n(B(\mathbf{0},1))$$

and since $\varepsilon > 0$ is arbitrary, this shows $\alpha(n) \geq \beta(n)$.

By the Vitali covering theorem, there exists a sequence of disjoint balls, $\{B_i\}$ such that $B(\mathbf{0},1) = (\cup_{i=1}^{\infty} B_i) \cup N$ where $m_n(N) = 0$. Then $\mathcal{H}_\delta^n(N) = 0$ can be concluded because $\mathcal{H}_\delta^n \leq \mathcal{H}^n$ and Lemma 15.1.7. Using $m_n(B(\mathbf{0},1)) = \mathcal{H}^n(B(\mathbf{0},1))$ again,

$$\begin{aligned}
\mathcal{H}_\delta^n(B(\mathbf{0},1)) &= \mathcal{H}_\delta^n(\cup_i B_i) \leq \sum_{i=1}^{\infty} \beta(n) r(B_i)^n \\
&= \frac{\beta(n)}{\alpha(n)} \sum_{i=1}^{\infty} \alpha(n) r(B_i)^n = \frac{\beta(n)}{\alpha(n)} \sum_{i=1}^{\infty} m_n(B_i) \\
&= \frac{\beta(n)}{\alpha(n)} m_n(\cup_i B_i) = \frac{\beta(n)}{\alpha(n)} m_n(B(\mathbf{0},1)) = \frac{\beta(n)}{\alpha(n)} \mathcal{H}^n(B(\mathbf{0},1))
\end{aligned}$$

which implies $\alpha(n) \leq \beta(n)$ and so the two are equal. This proves that if $\alpha(n) = \beta(n)$, then the $\mathcal{H}^n = m_n$ on the measurable sets of $\mathbb{R}^n$.

This gives another way to think of Lebesgue measure which is a particularly nice way because it is coordinate free, depending only on the notion of distance.

For $s < n$, note that $\mathcal{H}^s$ is not a Radon measure because it will not generally be finite on compact sets. For example, let $n = 2$ and consider $\mathcal{H}^1(L)$ where L is a line segment joining $(0,0)$ to $(1,0)$. Then $\mathcal{H}^1(L)$ is no smaller than $\mathcal{H}^1(L)$ when L is considered a subset of $\mathbb{R}^1, n = 1$. Thus by what was just shown, $\mathcal{H}^1(L) \geq 1$. Hence $\mathcal{H}^1([0,1] \times [0,1]) = \infty$. The situation is this: L is a one-dimensional object inside $\mathbb{R}^2$ and $\mathcal{H}^1$ is giving a one-dimensional measure of this object. In fact, Hausdorff measures can make such heuristic remarks as these precise. Define the Hausdorff dimension of a set, A, as

$$\dim(A) = \inf\{s : \mathcal{H}^s(A) = 0\}$$

15.2.4 A Formula For $\alpha(n)^*$

What is $\alpha(n)$? Recall the gamma function which makes sense for all $p > 0$.

$$\Gamma(p) \equiv \int_0^\infty e^{-t} t^{p-1} dt.$$

Lemma 15.2.6 *The following identities hold.*

$$p\Gamma(p) = \Gamma(p+1),$$

$$\Gamma(p)\Gamma(q) = \left(\int_0^1 x^{p-1}(1-x)^{q-1} dx \right) \Gamma(p+q),$$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

Proof: Using integration by parts,

$$\begin{aligned}
\Gamma(p+1) &= \int_0^\infty e^{-t} t^p dt = -e^{-t} t^p|_0^\infty + p \int_0^\infty e^{-t} t^{p-1} dt \\
&= p\Gamma(p)
\end{aligned}$$

Next

$$\begin{aligned}\Gamma(p)\Gamma(q) &= \int_0^\infty e^{-t}t^{p-1}dt\int_0^\infty e^{-s}s^{q-1}ds\\ &= \int_0^\infty\int_0^\infty e^{-(t+s)}t^{p-1}s^{q-1}dtds\\ &= \int_0^\infty\int_s^\infty e^{-u}(u-s)^{p-1}s^{q-1}duds\\ &= \int_0^\infty\int_0^u e^{-u}(u-s)^{p-1}s^{q-1}dsdu\\ &= \int_0^\infty\int_0^1 e^{-u}(u-ux)^{p-1}(ux)^{q-1}udxdu\\ &= \int_0^\infty\int_0^1 e^{-u}u^{p+q-1}(1-x)^{p-1}x^{q-1}dxdu\\ &= \Gamma(p+q)\left(\int_0^1 x^{p-1}(1-x)^{q-1}dx\right).\end{aligned}$$

It remains to find $\Gamma\left(\frac{1}{2}\right)$.

$$\Gamma\left(\frac{1}{2}\right)=\int_0^\infty e^{-t}t^{-1/2}dt=\int_0^\infty e^{-u^2}\frac{1}{u}2udu=2\int_0^\infty e^{-u^2}du$$

Now

$$\begin{aligned}\left(\int_0^\infty e^{-x^2}dx\right)^2 &= \int_0^\infty e^{-x^2}dx\int_0^\infty e^{-y^2}dy=\int_0^\infty\int_0^\infty e^{-\left(x^2+y^2\right)}dxdy\\ &= \int_0^\infty\int_0^{\pi/2}e^{-r^2}rd\theta dr=\frac{1}{4}\pi\end{aligned}$$

and so

$$\Gamma\left(\frac{1}{2}\right)=2\int_0^\infty e^{-u^2}du=\sqrt{\pi}$$

This proves the lemma. ■

Next let n be a positive integer.

Theorem 15.2.7 *$\alpha(n)=\pi^{n/2}(\Gamma(n/2+1))^{-1}$ where $\Gamma(s)$ is the gamma function*

$$\Gamma(s)=\int_0^\infty e^{-t}t^{s-1}dt.$$

Proof: First let $n=1$.

$$\Gamma(\frac{3}{2})=\frac{1}{2}\Gamma\left(\frac{1}{2}\right)=\frac{\sqrt{\pi}}{2}.$$

Thus

$$\pi^{1/2}(\Gamma(1/2+1))^{-1}=\frac{2}{\sqrt{\pi}}\sqrt{\pi}=2=\alpha(1).$$

and this shows the theorem is true if $n=1$.

Assume the theorem is true for n and let B_{n+1} be the unit ball in $\mathbb{R}^{n+1}$. Then by the result in $\mathbb{R}^n$,

$$m_{n+1}(B_{n+1})=\int_{-1}^1\alpha(n)(1-x_{n+1}^2)^{n/2}dx_{n+1}$$

$$= 2\alpha(n)\int_0^1 (1-t^2)^{n/2}dt.$$

Doing an integration by parts and using Lemma 15.2.6

$$\begin{aligned}
&= 2\alpha(n)n\int_0^1 t^2(1-t^2)^{(n-2)/2}dt \\
&= 2\alpha(n)n\frac{1}{2}\int_0^1 u^{1/2}(1-u)^{n/2-1}du \\
&= n\alpha(n)\int_0^1 u^{3/2-1}(1-u)^{n/2-1}du \\
&= n\alpha(n)\Gamma(3/2)\Gamma(n/2)(\Gamma((n+3)/2))^{-1} \\
&= n\pi^{n/2}(\Gamma(n/2+1))^{-1}(\Gamma((n+3)/2))^{-1}\Gamma(3/2)\Gamma(n/2) \\
&= n\pi^{n/2}(\Gamma(n/2)(n/2))^{-1}(\Gamma((n+1)/2+1))^{-1}\Gamma(3/2)\Gamma(n/2) \\
&= 2\pi^{n/2}\Gamma(3/2)(\Gamma((n+1)/2+1))^{-1} \\
&= \pi^{(n+1)/2}(\Gamma((n+1)/2+1))^{-1}.
\end{aligned}$$

This proves the theorem. ■

From now on, in the definition of Hausdorff measure, it will always be the case that $\beta(s) = \alpha(s)$. As shown above, this is the right thing to have $\beta(s)$ to equal if s is a positive integer because this yields the important result that Hausdorff measure is the same as Lebesgue measure. Note the formula, $\pi^{s/2}(\Gamma(s/2+1))^{-1}$ makes sense for any $s \geq 0$.

15.3 Hausdorff Measure And Linear Transformations

Hausdorff measure makes possible a unified development of n dimensional area including in one theory length and surface area. Imagine the boundary of an open set in $\mathbb{R}^3$. You would tend to think of this as something two dimensional. The way to measure it is with $\mathcal{H}^2$. Length can be measured by $\mathcal{H}^1$ and the boundary of an open set in $\mathbb{R}^4$ is measured in terms of $\mathcal{H}^3$ etc.

As in the case of Lebesgue measure, the first step in this is to understand basic considerations related to linear transformations. Recall that for $L \in \mathcal{L}\left(\mathbb{R}^k, \mathbb{R}^l\right)$, L^* is defined by

$$(L\mathbf{u}, \mathbf{v}) = (\mathbf{u}, L^*\mathbf{v}).$$

Also recall the right polar decomposition, Theorem 3.9.3 on Page 68. This theorem says you can write a linear transformation as the composition of two linear transformations, one which preserves length and the other which distorts, the right polar decomposition. The one which distorts is the one which will have a nontrivial interaction with Hausdorff measure while the one which preserves lengths does not change Hausdorff measure. These ideas are behind the following theorems and lemmas.

Lemma 15.3.1 *Let $R \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, $n \leq m$, and $R^*R = I$. Then if $A \subseteq \mathbb{R}^n$,*

$$\mathcal{H}^n(RA) = \mathcal{H}^n(A).$$

In fact, if $P : \mathbb{R}^n \to \mathbb{R}^m$ satisfies $|P\mathbf{x} - P\mathbf{y}| = |\mathbf{x} - \mathbf{y}|$, then

$$\mathcal{H}^n(PA) = \mathcal{H}^n(A).$$

Proof: Note that

$$|R(\mathbf{x}-\mathbf{y})|^2 = (R(\mathbf{x}-\mathbf{y}), R(\mathbf{x}-\mathbf{y})) = (R^*R(\mathbf{x}-\mathbf{y}), \mathbf{x}-\mathbf{y}) = |\mathbf{x}-\mathbf{y}|^2$$

Thus R preserves lengths.

Now let P be an arbitrary mapping which preserves lengths and let A be bounded, $P(A) \subseteq \cup_{j=1}^{\infty} C_j$, $r(C_j) < \delta$, and

$$\mathcal{H}^n_\delta(PA) + \varepsilon > \sum_{j=1}^{\infty} \alpha(n)(r(C_j))^n.$$

Since P preserves lengths, it follows P is one to one on $P(\mathbb{R}^n)$ and P^{-1} also preserves lengths on $P(\mathbb{R}^n)$. Replacing each C_j with $C_j \cap (PA)$,

$$\begin{aligned} \mathcal{H}^n_\delta(PA) + \varepsilon &> \sum_{j=1}^{\infty} \alpha(n) r(C_j \cap (PA))^n \\ &= \sum_{j=1}^{\infty} \alpha(n) r\left(P^{-1}(C_j \cap (PA))\right)^n \\ &\geq \mathcal{H}^n_\delta(A). \end{aligned}$$

Thus $\mathcal{H}^n_\delta(PA) \geq \mathcal{H}^n_\delta(A)$.

Now let $A \subseteq \cup_{j=1}^{\infty} C_j, \operatorname{diam}(C_j) \leq \delta$, and

$$\mathcal{H}^n_\delta(A) + \varepsilon \geq \sum_{j=1}^{\infty} \alpha(n) (r(C_j))^n$$

Then

$$\begin{aligned} \mathcal{H}^n_\delta(A) + \varepsilon &\geq \sum_{j=1}^{\infty} \alpha(n) (r(C_j))^n \\ &= \sum_{j=1}^{\infty} \alpha(n) (r(PC_j))^n \\ &\geq \mathcal{H}^n_\delta(PA). \end{aligned}$$

Hence $\mathcal{H}^n_\delta(PA) = \mathcal{H}^n_\delta(A)$. Letting $\delta \to 0$ yields the desired conclusion in the case where A is bounded. For the general case, let $A_r = A \cap B(\mathbf{0}, r)$. Then $\mathcal{H}^n(PA_r) = \mathcal{H}^n(A_r)$. Now let $r \to \infty$. This proves the lemma. ■

Lemma 15.3.2 *Let $F \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m), n \leq m$, and let $F = RU$ where R and U are described in Theorem 3.9.3 on Page 68. Then if $A \subseteq \mathbb{R}^n$ is Lebesgue measurable,*

$$\mathcal{H}^n(FA) = \det(U) m_n(A).$$

Proof: Using Theorem 9.8.8 on Page 241 and Theorem 15.1.9,

$$\mathcal{H}^n(FA) = \mathcal{H}^n(RUA)$$

$$= \mathcal{H}^n(UA) = m_n(UA) = \det(U) m_n(A).$$

Definition 15.3.3

Define J to equal $\det(U)$. *Thus*

$$J = \det((F^*F)^{1/2}) = (\det(F^*F))^{1/2}.$$

Bibliography

[1] **Apostol, T. M.,** *Calculus second edition,* Wiley, 1967.

[2] **Apostol T.M.** *Calculus Volume II Second edition,* Wiley *1969.*

[3] **Apostol, T. M.,** *Mathematical Analysis,* Addison Wesley Publishing Co., 1974.

[4] **Baker, Roger**, *Linear Algebra*, Rinton Press 2001.

[5] **Bartle R.G.,** *A Modern Theory of Integration,* Grad. Studies in Math., Amer. Math. Society, Providence, RI, 2000.

[6] **Bartle R. G. and Sherbert D.R.** *Introduction to Real Analysis* third edition, Wiley 2000.

[7] **Chahal J. S.** , *Historical Perspective of Mathematics* 2000 B.C. - 2000 A.D.

[8] **Davis H. and Snider A.,** *Vector Analysis* Wm. C. Brown 1995.

[9] **Deimling K.** *Nonlinear Functional Analysis,* Springer-Verlag, 1985.

[10] **D'Angelo, J. and West D**. *Mathematical Thinking Problem Solving and Proofs*, Prentice Hall 1997.

[11] **Edwards C.H.** *Advanced Calculus of several Variables,* Dover 1994.

[12] **Euclid**, *The Thirteen Books of the Elements*, Dover, 1956.

[13] **Evans L.C.** and **Gariepy,** *Measure Theory and Fine Properties of Functions,* CRC Press, 1992.

[14] **Evans L.C.** *Partial Differential Equations,* Berkeley Mathematics Lecture Notes. 1993.

[15] **Fitzpatrick P. M.,** *Advanced Calculus a course in Mathematical Analysis,* PWS Publishing Company 1996.

[16] **Federer H.,** *Geometric Measure Theory,* Springer-Verlag, New York, 1969.

[17] **Fleming W**., *Functions of Several Variables,* Springer Verlag 1976.

[18] **Fonesca I. and Gangbo W.** *Degree theory in analysis and applications* Clarendon Press 1995.

[19] **Greenberg, M**. Advanced Engineering Mathematics, Second edition, Prentice Hall, 1998

[20] **Gromes W.** Ein einfacher Beweis des Satzes von Borsuk. *Math. Z.* 178, pp. 399 -400 (1981)

[21] **Gurtin M.** *An introduction to continuum mechanics,* Academic press 1981.

[22] **Hardy G.,** *A Course Of Pure Mathematics, Tenth edition,* Cambridge University Press 1992.

[23] **Heinz, E.**An elementary analytic theory of the degree of mapping in n dimensional space. *J. Math. Mech. 8, 231-247* 1959

[24] **Henstock R**. *Lectures on the Theory of Integration,* World Scientific Publiching Co. 1988.

[25] **Horn R. and Johnson C.** *matrix Analysis,* Cambridge University Press, 1985.

[26] **Karlin S. and Taylor H.** *A First Course in Stochastic Processes,* Academic Press, 1975.

[27] **Kuttler K. L.**, *Basic Analysis*, Rinton

[28] **Kuttler K.L.,** *Modern Analysis* CRC Press 1998.

[29] **Lang S.** *Real and Functional analysis* third edition Springer Verlag 1993. Press, 2001.

[30] **McLeod R.** *The Generalized Riemann Integral,* Mathematical Association of America, Carus Mathematical Monographs number 20 1980

[31] **McShane E. J.** *Integration,* Princeton University Press, Princeton, N.J. 1944.

[32] **Nobel B. and Daniel J.** *Applied Linear Algebra,* Prentice Hall, 1977.

[33] **Rose, David, A.,** The College Math Journal, vol. 22, No.2 March 1991.

[34] **Rudin, W.,** *Principles of mathematical analysis,* McGraw Hill third edition 1976

[35] **Rudin W.,** *Real and Complex Analysis,* third edition, McGraw-Hill, 1987.

[36] **Salas S. and Hille E.,** *Calculus One and Several Variables,* Wiley 1990.

[37] **Sears and Zemansky,** *University Physics, Third edition, Addison Wesley 1963.*

[38] **Tierney John,** *Calculus and Analytic Geometry,* fourth edition, Allyn and Bacon, Boston, 1969.

[39] **Yosida K.,** *Functional Analysis, Springer Verlag, 1978.*

Index