Intermediate Calculus and Linear Algebra

Jerry L. Kazdan

Harvard University
Lecture Notes, 1964–1965

Preface

These notes will contain most of the material covered in class, and be distributed before each lecture (hopefully). Since the course is an experimental one and the notes written before the lectures are delivered, there will inevitably be some sloppiness, disorganization, and even egregious blunders—not to mention the issue of clarity in exposition. But we will try. Part of your task is, in fact, to catch and point out these rough spots. In mathematics, proofs are not dogma given by authority; rather a proof is a way of convincing one of the validity of a statement. If, after a reasonable attempt, you are not convinced, complain loudly.

Our subject matter is intermediate calculus and linear algebra. We shall develop the material of linear algebra and use it as setting for the relevant material of intermediate calculus. The first portion of our work—Chapter 1 on infinite series—more properly belongs in the first year, but is relegated to the second year by circumstance. Presumably this topic will eventually take its more proper place in the first year.

Our course will have a tendency to swallow whole two other more advanced courses, and consequently, like the duck in Peter and the Wolf, remain undigested until regurgitated alive and kicking. To mitigate—if not avoid—this problem, we shall often take pains to state a theorem clearly and then either prove only some special case, or offer no proof at all. This will be true especially if the proof involves technical details which do not help illuminate the landscape. More often than not, when we only prove a special case, the proof in the general case is essentially identical—the equations only becoming larger.

September 1964

Afterward

I have now taught from these notes for two years. No attempt has been made to revise them, although a major revision would be needed to bring them even vaguely in line with what I now believe is the "right" way to do things. And too, the last several chapters remain unwritten. Because the notes were written as a first draft under panic pressure, they contain many incompletely thought-out ideas and expose the whimsy of my passing moods.

It is with this—and the novelty of the material at the sophomore level—in mind, that the following suggestions and students' reactions are listed. There are three categories, A), Material that turned out to be too difficult (they found rigor hard, but not many of the abstractions), B), changes in the order of covering the stuff, and C), material—mainly supplementary at this level—which is not too hard, but should be omitted if one ever hopes to complete the "standard" topics within the confines of a year course.

(A) *It was too hard* (unless one took vast chunks of time).

(1) Completeness of reals. Only "monotone sequences converge" is needed for infinite series.

(2) Term-by-term differentiation and integration of power series. The statement of the main theorem should be fully intelligible—but the proof is too complicated.

(3) Cosets. This is apparently too abstract. It might be possible to do after finding general solutions of linear inhomogeneous O.D.E.'s.

(4) L_2 and uniform convergence of Fourier series. Again, all I ended up doing was to try to state what the issues were, and not to attempt the proof. The ambitious student should be warned that my proof of the Weierstrass theorem is opaque (one should explicitly introduce the idea of an approximate identity).

(5) Fundamental Theorem of Algebra. The students simply don't believe inequalities in such profusion.

(6) I you want to see rank confusion, try to teach the class how to compute higher order partial derivatives using the chain rule. That computation should be one of the headaches of advanced calculus.

(7) Existence of a determinant function. I don't know a simple proof except for the one involving permutations—and I hate that one.

(8) Dual spaces. As lovely as the ideas are, this topic is too abstract, and to my knowledge, unneeded at this level where almost all of the spaces are either finite dimensional or Hilbert spaces. One should, however, mention the words "vector" and "covector" to distinguish column from row vectors. I forgot to do so in these notes and it did cause some confusion.

(B) *Changes in Order and Timing.* The structure of the notes is to investigate bare linear spaces, then linear mappings between them, and finally non-linear mappings between them. It is with this in mind that *linear* O.D.E.'s came before *nonlinear* maps from $\mathbb{R}^n \to \mathbb{R}$. The course ended by treating the simplest problem in the calculus of variations as an example of a nonlinear map from an infinite dimensional space to the reals. My current feeling is to consider linear *and* non-linear maps between *finite* dimensional spaces before doing the infinite dimensional example of differential equations.

The first semester should get up to the generalities on solving $LX = Y$, p. 319 [incidentally, the material on inverses (p. 355 ff) belongs around p. 319]. Most students find the material on linear dependence difficult—probably for two reasons: 1) they are not used to formal definitions, and ii) they think they have learned a technique for doing something, not just a naked definition, and can't quite figure out just what they can do with it. In other words, they should feel these definitions about the anatomy of linear spaces are similar to those describing a football field and of little value until the game begins—i.e., until the operators between spaces make their grand entrance.

Because of time shortages, the sections on linear maps from $\mathbb{R}^1 \to \mathbb{R}^n$ and $\mathbb{R}^n \to \mathbb{R}^1$, pp. 320-41 were regrettably omitted both years I taught the course. The notes were written so that these sections can be skipped.

(C) *Supplementary Material.* A remarkable number of fascinating and important topics could have been included—if there were only enough time. For example:

(1) Change of bases for linear transformations (including the spectral theorem).

(2) Elementary differential geometry of curves and surfaces.

(3) Inverse and implicit function theorems. These should be stated as natural generalizations of the problems of a) inverting a linear map, b) finding the null space of a linear map, and c) generalizing $\dim D(L) = \dim R(L) + \dim N(L)$ all to local properties of nonlinear maps via the tangent map.

(4) Change of variable in multiple integration. Determinants were deliberately introduced as oriented volume to make the result obvious for linear maps and plausible for nonlinear maps.

(5) Constrained extrema using Lagrange multipliers.

(6) Line and surface integrals along with the theorems of Gauss, Green, and Stokes. The formal development of differential forms takes too much time to do here. Perhaps a satisfactory solution is to restrict oneself to line integrals and these theorems in the plane, where the topological difficulties are minimal.

(7) Elementary Morse Theory. One can prove the Morse inequalities easily for the real line, the circle, the plane, and S^2 merely by gradually flooding these sets and observing the number of lakes and shore line changes only at the critical points.

(8) Sturm-Liouville theory. An elegant fusion of the geometry of Hilbert spaces to differential equations.

(9) Translation-invariant operators with applications to constant coefficient difference and differential equations. The Laplace and Fourier transforms enter naturally here.

(10) The Calculus of Variations. The formalism of nonlinear functionals on $\mathbb{R}^n$, i.e., maps $f\colon \mathbb{R}^n \to \mathbb{R}$, generalizes immediately to nonlinear functionals defined on infinite dimensional spaces.

(11) The deleted rigor.

(12) Linear operators with finite dimensional (perhaps even compact) range.

One parting warning. When covering intermediate calculus from this viewpoint, it is all too natural to forget the innocence of the class, to enchant with glitter, and to numb with purity and formalism. Emphasis should be placed on developing insight and intuition along with routine computational facility.

My classes found frequent reviews of the mathematical edifice, backward glances at the previous months' work, not only helpful but mandatory if they were to have any conception of the vast canvas which was being etched in their minds over the course of the year. The question, "What are we doing now and how does it fit into the larger plan?" must constantly be raised and at least partially resolved.

May, 1966

Contents

Chapter 0

Remembrance of Things Past.

We shall treat a hodge-podge of topics in a hasty and incomplete fashion. While most of these topics should have been learned earlier, section 5 on the completeness of the real numbers has its more rightful place in advanced calculus. Do *not* take time to read this chapter unless the particular topic is needed; then read only the relevant portions. The chapter is included for reference.

0.1 Sets and Functions

A *set* is any collection of objects, called the *elements* of the set, together with a criterion for deciding if an object is in the set. For example, I) the set of all girls with blue eyes and blond hair, and ii) the less picturesque set of all positive even integers. We can also define a set by bluntly listing all of its elements. Thus, the set of all students in this class is defined by the list in the roll book.

Sets are often specified by a notation which is best described by examples.

i) $S = \{\, x\colon x \text{ is an integer } \}$ is the set of all integers.

ii) $T = \{\, (x, y)\colon x^2 + y^2 = 1 \,\}$ is the set of all points (x, y) on the unit circle $x^2 + y^2 = 1$.

iii) $A = \{\, 1, 2, 7, -3 \,\}$ is the set of integers $1, 2, 7$ and -3.

Our attitude toward set theory will be extremely casual; we shall mainly use it as a language and notation. Without further ado, let us introduce some notation.

$x \in S$, $\quad$ x is an element of the set S, or just x is in S.

$x \notin S$, $\quad$ x is not an element of the set S.

$\mathbb{Z}$, $\quad$ the set of all integers, positive, zero, and negative.

$\mathbb{Z}_+$, $\quad$ the set of all positive integers, excluding 0.

$\mathbb{R}$, $\quad$ the set of all real numbers (to be defined more precisely later).

$\mathbb{C}$, $\quad$ The set of all complex numbers (also to be defined more precisely later).

$\emptyset$, $\quad$ the set with no elements, the *empty* or *null* set. It is extremely uninteresting.

Definition: Given the two sets S and T, i) the set $S \cup T$, "S *union* T", is the set of elements which are in *either* S *or* T, or both.

ii) The set $S \cap T$, "S *intersection* T", is the set of elements in *both* S and T.

If we represent S by one blob and T by another, $S \cup T$ is the shaded region while $S \cap T$ is the cross-hatched region. Note that all elements in $S \cap T$ are also in $S \cup T$. Two sets are *disjoint* if $S \cap T = \emptyset$, that is, if their intersection is empty.

A *subset* of a set is another way of referring to a portion of a given set. Formally, A is the subset of S, written $A \subset S$, if every element in A is also an element of S. The set A is a subset of the set S if and only if either

$$A \cup S = S, \text{ or, equivalently, } A \cap S = A.$$

It is possible that $A = S$, or that $A = \emptyset$. If these degenerate cases are excluded, we say that A is a *proper subset* of S.

Given the two sets S and T, it is natural to form a new set $S \times T$, "S cross T", which consists of all pairs of elements, one from S and the other from T. For example, if S is the set of all men in this class, and T the set of all women in this class, then $S \times T$ is the set of all couples, a natural set to contemplate.

If $x \in S$ and $y \in T$, the standard notation for the induced element in $S \times T$ is (x, y). Note that the order in (x, y) is important. The element on the left is from S, while that on the right is from T. For this reason the pair of elements (x, y) is usually called an *ordered pair.* The whole set $S \times T$ is called the *product*, *direct product*, or *Cartesian product* of S and T, all three names being used interchangeably.

You have met this idea in graphing points in the plane. Since these points, (x, y), are determined by an ordered pair of real numbers, they are just the elements of $\mathbb{R} \times \mathbb{R}$. From this example it is clear that even though this set $\mathbb{R} \times \mathbb{R}$ is the product of a set with *itself*, the *order* of the pair (x, y) is still important. For example the point $(1, 2) \in \mathbb{R} \times \mathbb{R}$ is certainly not the same as $(2, 1) \in \mathbb{R} \times \mathbb{R}$.

Having defined the direct product of two sets S and T as ordered pairs, it is reasonable to define the direct product of three sets S, T, and U as the set of ordered triplets (x, y, z), where $x \in S$, $y \in T$, and $z \in U$. The extension to n sets, $S_1 \times S_2 \times \cdots \times S_n$, is done in the same way.

Let us now recall the ideas behind the notion of a function.

A *function* f from the set X into the set B is a rule which assigns to every $x \in X$ one and only one element $y = f(x) \in B$. We shall also say that f *maps* X into B, and write either

$$f \colon X \to B, \text{ or } X \xrightarrow{f} B.$$

This alternative notation is useful when X and B are more important than the specific nature of f. The set X is the *domain* of f, while the *range* of f is the subset $Y \subset B$ of all elements $y \in B$ which are the image of (at least) one point $x \in X$, so $y = f(x)$, or in suggestive notation, $Y = f(X)$.

Automobile license plates supply a nice example, for they assign to every license plate sold a unique car. The domain is the set of all license plates sold, while the range is not all cars, but rather the subset of all cars which are driven. Wrecks and museum pieces neither need nor have license plates since they are not on the roads. Some other examples are i) the function $f(n) \equiv \frac{1}{n}$, $n = 1, 2, 3, \ldots$ which assigns to every $n \in \mathbb{Z}_+$ the rational number $\frac{1}{n}$, and ii) the function $f(n, m) = \frac{m}{n}$, n, $m = 1, 2, 2, \ldots$, which assigns to every element of $\mathbb{Z}_+ \times \mathbb{Z}_+$ the rational number $\frac{m}{n}$.

Quite often we shall use functions which map *part* of some set into *part* of some other set. In other words the function may be defined on only a subset of a given set and take on values in a subset of some other set. The function $f(n, m) \equiv \frac{m}{n}$ of the previous paragraph is of this nature for we defined it on a subset of $\mathbb{Z} \times \mathbb{Z}$ and takes its values on the positive subset of the set of all rational numbers.

There is some standard nomenclature (or fancy words, if you like) associated with mapping. Say $X \subset A$ and the function $f \colon X \to B$. Note that we know the definition of f only on X. It may not be defined for the remainder of A.

Definition: i) if *every* element of B is the image of (at least) one point in X, the map f is called *surjective* or *onto*. In other words $f \colon X \to B$ is a surjection if the range of f is *all* of B. Thus f is always surjective onto its range.

ii) If the map f has the property that for every $x_1, x_2 \in X$, we have $f(x_1) = f(x_2)$ when and only when $x_1 = x_2$, the map is called *injective* or *one to one* (1-1). This is the case if no two different elements in X are mapped into the *same* element in B.

iii) If the map f is both surjective and injective, that is, if it is both onto and 1-1, then f is called *bijective.*

Examples: For these, we have $f \colon X \to B$ where $X = B = \mathbb{Z}$.

(1) The map $f(n) = 2n$ is injective but not surjective since the range does not contain the odd integers in B.

(2) The map $f(n) = \begin{cases} \frac{n}{2} & \text{if n is even} \\ \frac{n+1}{2} & \text{if n is odd} \end{cases}$ is surjective but not injective since every element in B is the image of two distinct elements of X.

(3) The map $f(n) = n + 7$ is bijective.

Notational Remark: For functions whose domain is $\mathbb{Z}$ or $\mathbb{Z}_+$ it is customary to indicate the element of the range by a notation like a_n instead of $f(n)$. Thus $f(n) = \frac{1}{n}$, where $n \in \mathbb{Z}_+$, is written as $a_n = \frac{1}{n}$. Such a function is usually called a *sequence.*

The concepts we have just defined are useful if we try to define what we mean by the inverse of a function.

Definition: A function $f \colon X \to B$ is *invertible* if to every $b \in B$ there is one and only one $x \in X$ such that $b = f(x)$. Thus f is invertible if and only if it is bijective. If f is invertible, we denote the inverse function by f^{-1}, so $x = f^{-1}(b)$.

If $f \colon A \to B$, and $g \colon B \to C$, then when *composed* (put together) these two functions induce a mapping, $g \circ f$, of A into C. Slightly more generally, if $B \subset R$, and $f \colon A \to B$ while $g \colon R \to C$, th en $g \circ f \colon A \to C$.
You should be able to see why the composed map $g \circ f$ is only defined on A, and then understand that our stipulation that $B \subset R$ is a convenient requirement.

If $x \in A$ and $z \in C$, then $g \circ f$ maps x onto $z = (g \circ f)(x)$, or in more familiar notation, $z = g(f(x))$. Now an example. Say the distance s you have walked at time t is specified by the function $s = f(t)$, and the amount z of shoe leather worn out by walking the distance s is given by the function $z = g(s)$. Then the amount of shoe leather you have worn out at time t is given by the composed function $z = g(f(t))$. Here $t \in A$, $s \in B$, and $z \in C$. Hopefully you have by now recognized that the "chain rule" for derivatives is just the procedure for finding the derivative of composed functions from their constituent parts. In our example the chain rule would be used to find $\frac{dz}{dt}$ from $\frac{dg}{ds}$ and $\frac{df}{dt}$-if these functions were differentiable.

We conclude this section with more symbols—if you have not yet had enough. These are borrowed from logic. Although we shall use them only infrequently as a shorthand, they might have greater use to you in class notes.

$\forall$ "for every"
$\exists$ "there is", or "there exists"
$\ni$ "such that"
$A \Rightarrow B$ "the truth of statement A implies that of statement B".
$A \Leftrightarrow B$ "statement A is equivalent to statement B, that is, both $A \Rightarrow B$ and $B \Rightarrow A$.

Exercises

(1) If $R = \{1, 4\}$, $S = \{1, 2, 3, 4, \}$, and $T = \{2, 3, 7\}$, find the six other sets $R \cup S$, $R \cap S$, $R \cup T$, $R \cap T$, $S \cup T$, and $S \cap T$. Which of these nine sets are proper subsets of which other sets?

(2) If $S = \{x\colon |x - 1| \le 2\}$ and $T = \{x\colon |x| \le 2\}$, find $S \cup T$ and $S \cap T$. A sketch is adequate.

(3) If A, B, and C are any subsets of a set S, prove

 (a) $(A \cup B) \cup C = A \cup (B \cup C)$ —so that the parenthesis can be omitted without creating ambiguity.
 (b) $(A \cap B) \cap C = A \cap (B \cap C)$ —so that again the parentheses are superfluous.
 (c) $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$.
 (d) $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$.

 REMARK: two sets X and Y are proved equal by showing that both $X \subset Y$ and $Y \subset X$.

(4) If the function f has domain S, and both $A \subset C$ and $B \subset S$, prove that

 (i) $A \subset B \Rightarrow f(A) \subset f(B)$.
 (ii) $f(A \cap B) \subset f(A) \cap f(B)$ [We cannot hope to prove equality because of counterexamples like: let $A = \{-2, -1, 0, 1, 2, 3\}$ and $B = \{-4, -3, -2, -1\}$. Then with $f(n) = n^2$, we have $f(A) = \{0, 1, 4, 9\}$, $f(B) = \{1, 4, 9, 16\}$, and $f(A \cup B) = \{1, 4\} \ne f(A) \cap f(B)$].

 (iii) $f(A \cup B) = f(A) \cup f(B)$.

(5) For the following functions $f\colon X \to B$, classify as to injection, surjection, or bijection, or none of these.

 (i) $f(n) = n^2$ with $X = \mathbb{Z}_+$ and $B = \mathbb{Z}$.
 (ii) Let $X = \{\textit{all rational numbers}\}$, $B = \{\textit{all rational numbers}\}$, and $f(x) = \frac{1}{m}$, where $x = \frac{n}{m} \in X$ [Here $\frac{n}{m}$ is assumed to be reduced to lowest terms.]
 (iii) $f(x) = \frac{1}{x}$, where $x \in X$ and $X = B = \{\textit{all positive rational numbers}\}$.
 (iv) $X = \{\textit{all women born in May}\}$, $B = \{\textit{the thirty days in the month of June}\}$, and let f be the function assigning "her birthday" to each woman born in June.
 (v) $f(n) = |n|$, with $X = B = \mathbb{Z}$.

0.2 Relations

A relationship often exists between elements of sets. Some common examples are i) $a \ge b$, ii) $a \perp b$ (perpendicular to), iii) a loves b, and iv) $a \ne b$. Let S be a given set, a, $b \in S$, and let $\mathcal{R}$ be a relation defined on S (that is, $\forall\, a,\, b \in S$, either $a\mathcal{R}b$ or $a\not\mathcal{R}b$ with no third alternative possible). Most relations have at least one of the following properties.

 (i) *reflexive* $a\mathcal{R}a$ $\quad \forall a \in S$
 (ii) *symmetric* $a\mathcal{R}b \Rightarrow b\mathcal{R}a$
 (iii) *transitive* $(a\mathcal{R}b$ and $b\mathcal{R}c) \Rightarrow a\mathcal{R}c$.

EXAMPLES:

(1) perpendicular ($\perp$) is only symmetric.

(2) "loves" enjoys none of these (well, maybe it is reflexive).

(3) equality ($=$) has all three properties.

(4) geometric congruence ($\cong$) and geometric similarity ($\simeq$) both have all three.

(5) parallel ($\|$) has all three—if we are willing to agree that a line is parallel to itself.

(6) "is less than five miles from" is only reflexive and symmetric.

(7) for a, $b \in \mathbb{Z}_+$, the relation "a is divisible by b" is only reflexive and transitive but not symmetric.

(8) "less than" ($<$) is only transitive.

A relation which is reflexive, symmetric and transitive is called an *equivalence relation*. The standard examples are those of algebraic equality and of geometric congruence. An equivalence relation on a set S *partitions* the set into subsets of *equivalent elements*. Those terms are illustrated in the following.

EXAMPLES:

(1) In the set S of all triangles, the equivalence relation of geometric congruence partitions S into subsets of congruent triangles, any two triangles of S being in the same subset (or *equivalence class* as it is called) if and only i f they are congruent.

(2) In the set P of all people, consider the equivalence relation "has the same birthday," disregarding the year. This relation partitions P into 366 equivalence classes. Two people are in the same equivalence class if their birthdays fall on the same day of the year.

Notice that any two equivalence classes are either identical or disjoint, that is, they have either no elements in common or they coincide. This is particularly clear from the examples with birthdays.

By the fundamental theorem of calculus, we know that the indefinite integral of an integrable function f can be represented by any function F whose derivative is f. The mean value theorem told us that every other indefinite integral of f differs from F by only a constant. Thus, the indefinite integrals of a given function are an equivalence class of functions, differing from each other by constants. The equivalence relation is "equal up to an additive constant".

Exercises

(1) If $a,\ b,\ c,\ d\in\mathbb{Z}_+$, let us define the following equivalence relation between the elements of $\mathbb{Z}_+\times\mathbb{Z}_+$:

$$(a,b)\mathcal{R}(c,d) \quad \text{if and only if} \quad ad=bc.$$

Verify that $\mathcal{R}$ is an equivalence relation. [In real life, the pair (a,b) of this example is written as $\frac{a}{b}$, so all we have said is $\frac{a}{b}=\frac{c}{d}$ if and only if $ad=bc$. This equivalence relation partitions the set of rational numbers into very familiar equivalence classes. For example the equivalent rational numbers $\frac{1}{2},\ \frac{2}{4},\ \frac{3}{6},\ldots$ are in the same equivalence class, to no one's surprise].

(2) Explain the fallacy in the following argument by observing that equality "=" here is not the usual algebraic equality , but rather some other equivalence relation.

"Let $A=\int\frac{dx}{x}$. Integration by parts ($p=1/x,\ dq=dx$), gives

$$A=x(\frac{1}{x})-\int x(-\frac{1}{x^2})\,dx=1+A.$$

Hence $0=1$."

0.3 Mathematical Induction

You are familiar with a variety of proofs, viz. direct proofs and proofs by contradiction. There is, however, another type of proof which is not encountered very often in elementary mathematics: proof by *induction.*

Abstractly, you have a sequence of statements $P_1,\ P_2,\ P_3,\ldots$, and a guess for the nature of the general statement P_n. A proof by mathematical induction provides a method for showing the general statement P_n is correct. Here is how it is carried out. *First* verify that the statement is true in some special case, say for $n=1$, so you check the validity of P_1. *Second* you show that *if* it is true in some particular case $n=k$, then it is true for the next case $n=k+1$, that is, $P_k\Rightarrow P_{k+1}$. Now since P_1 is true, so is $P_{1+1}=P_2$, and consequently so is $P_{2+1}=P_3$, and so on up. Observe that the procedure does not tell you how in the world to guess the general statement P_n, but only shows how to verify it.

Let us carry out the procedure for an example. We guess the formula

$$1+2+\cdots+n=\frac{n(n+1)}{2} \tag{0-1}$$

STEP 1. Is the formula true for $n=1$? Yes, since both sides then equal 1.
STEP 2. Assuming the formula is true for $n=k$, we must show this implies the formula is true for $n=k+1$.

$$1+2+\cdots+k+(k+1)=\frac{(k+1)(k+2)}{2}.$$

The formula, assumed to be true, for $n=k$ is

$$1+2+\cdots+k=\frac{k(k+1)}{2}.$$

Adding $(k+1)$ to both sides we find that

$$1+2+\cdots+k+(k+1)=\frac{k(k+1)}{2}+(k+1)=\frac{(k+1)(k+2)}{2}$$

which is exactly the statement we wanted. This proves that formula (0.3) is true for all $n\geq 1$.

Exercises

Use mathematical induction to prove the given statements.

(1) $1^2+2^2+\cdots+n^2=\frac{n(n+1)(2n+1)}{6}$

(2) $\frac{d}{dx}(x^n)=nx^{n-1}$ (use the formula for the derivative of a product).

(3) Let $x_0=1$ and define $x_{n+1}=\sqrt{12+x_n}$ for $n\geq 0$. Show that $1<x_n<4$ for all $n\geq 1$.

(4) Let $I(n) = \int_0^{\frac{\pi}{2}} \sin^n x\,dx$

(a) Prove the following formula is correct when n is an odd integer ≥ 3,

$$I(n) = \frac{2\cdot 4\cdot 6\cdot\cdots\cdot(n-1)}{1\cdot 3\cdot 5\cdot\cdots\cdot n}$$

(b) Guess and prove the formula when n is an even integer ≥ 2.

(5) Let $\Gamma(s) = \int_0^{\infty} e^{-t}t^{s-1}\,dt$, where $s > 0$ (this is the famous *gamma function*).

(a) Show $\Gamma(s+1) = s\Gamma(s)$ (Hint: integrate by parts)

(b) If $n \in \mathbb{Z}_+$, guess and prove the formula for $\Gamma(n+1)$.

(6) Say you have a square that has n straight lines drawn across it. The lines must begin and terminate at the boundary of the square but need not be parallel to the sides of the square. This partitions the square into regions all of whose sides are line segments and no line segments are inside the region (a simple example is a chess board whose lines are not necessarily parallel to the sides).

Show the resulting square can be colored unambiguously using only two colors,just as on a chess board. [Here, *unambiguously* means that no two adjacent regions have the same color.]

0.4 The Real Numbers: Algebraic and Order Properties.

The set of all real numbers can be characterized by a set of axioms. These properties are of three different types, i) algebraic properties, ii) order properties, and iii) the completeness property. Of these, the last is by far the most difficult to grasp. But that is getting ahead of our story. Let S be a set with the following properties.

I. Algebraic Properties

A. Addition. To every pair of elements $a, b \in S$, is associated another element, denoted by $a + b$, with the properties

A - 0. $(a+b) \in S$

A - 1. Associative: for every $a, b, c \in S, a + (b+c) = (a+b)+c$.

A - 2. Commutative: $a + b = b + a$

A - 3. There is an *additive identity*, that is, an element "0" $\in S$ such that $0 + a = a$ for all $a \in S$.

A - 4. For every $a \in S$, there is also a $b \in S$ such that $a + b = 0$. b is the *additive inverse* of a, usually written $-a$.

M. Multiplication. To every pair $a, b \in S$, there is associated another element, denoted by ab, with the properties

M - 0. $ab \in S$

M - 1. Associative. For every $a, b, c \in S$, $a(bc) = (ab)c$.

M - 2. Commutative. $ab = ba$.

M - 3. There is a *multiplicative identity*, that is, an element "l" $\in S$ such that $la = a$ for all $a \in S$. Moreover $1 \neq 0$.

M - 4. For every $a \in S, a \neq 0$, there is also a $b \in S$ such that $ab = 1$. b is the *multiplicative inverse* of a, usually written $\frac{1}{a}$ or a^{-1}.

D. Connection between Addition and Multiplication.

D - 1. *Distributive.* For every $a, b, c \in S$, $a(b + c) = ab + ac$.

Some sample - and simple—consequences of these nine axioms are i) $a + 0 = a$, ii) $a \cdot 1 = a$, and iii) $a + b = a + c \Rightarrow b = c$.

Any set whose elements satisfy the axioms A-0 to A-4 is called a *commutative*(or *abelian*) *group*. The *group operation* here is addition. In this language, we see that the multiplication axioms just state that the elements of S —with the additive identity 0 excluded—also form a commutative group, with the group operation being multiplication. These additive and multiplicative structures are connected by the distributive axiom. Most of high school algebra takes place in this setting; however, the possibility of non-integer exponents is not yet specifically included; in particular the square root of an element of S is not necessarily also in S.

Our axioms, or some part of them, are satisfied by sets other than the real numbers. The set of even integers form a commutative group with the group operation being addition, while numbers of the form 2^n, $n \in \mathbb{Z}$, form a commutative group under multiplication. The set of rational numbers satisfies all nine axioms. Any such set which satisfies all nine axioms is called a *field*. Both the real numbers and the rational numbers (a subset of the real numbers) are fields. A more thorough investigation of groups and fields is carried out in courses in modern algebra.

II. Order Axioms

Besides the above algebraic rules, we shall introduce an *order relation*, intuitively, the notion of 'greater than". To do this we need to use an undefined concept of positivity for elements of S and use it to state our axioms.

O -1. If $a \in S$ and $b \in S$ are positive, so are $a + b$ and ab.

O -2. The additive identity 0 is not positive.

O - 3. For every $a \in S, a \neq 0$, either a or $-a$ is positive, but not both. If $-a$ is positive, we shall say that a is negative.

Trichotomy Theorem. For any two numbers $a, b \in S$, exactly one of the following three statements is true, i) $a - b$ is positive, ii) $b - a$ is positive, or iii) $b - a$ is zero. If the notation $a < b$ is used to mean "$b - a$ is positive," and $a > b$ means $b < a$, then this theorem reads, either $a > b$, $a < b$, or $a = b$. The proof—which you should do—is a simple consequence of our axioms.

Some other consequences are

$a < b$ and $b < c \Rightarrow a < c$ (transitivity of "$<$")

$a < b$ and $c > 0 \Rightarrow ac < bc$

$a \neq 0 \Rightarrow a^2 > 0$. (Since $1 = 1^2$, this implies $1 > 0$).

The set of rational numbers as well as the set $\mathbb{R}$ of real numbers satisfy all twelve axioms. Any set which satisfies these twelve axioms is called an *ordered field*.

Exercises

(1) Let T be a set whose elements are of the form $a + b\sqrt{2}$, where a and b are rational numbers (and so are elements of a field). Show that T is also a field.

(2) Consider the set of all integers $\mathbb{Z}$ with the following equivalence relation: $m \in \mathbb{Z}$ and $n \in \mathbb{Z}$ are equivalent if they have the same remainder when divided by 2. The notation for this equivalence is

$$m \equiv n \quad (\text{mod} 2)$$

This equivalence relation partitions $\mathbb{Z}$ into two equivalence classes which we may denote respectively by 0 if the number is even, and 1 if the number is odd . Thus $8 \equiv -22 \pmod 2$ and $7 \equiv 13 \pmod 2$. Prove that the set $\mathbb{Z}$ with ordinary addition and multiplication but with this equivalence relation forms a field.

(3) Prove the trichotomy theorem.

(4) Prove that if $a \neq 0$, then $a^2 > 0$. Use it to prove that $1 > 0$ and then to conclude that all of the 'positive integers" are, in fact, positive.

0.5 The Real Numbers: Completeness Property.

III. Completeness Axiom.

So far our axioms do not insure that we can take fractional powers like the square root, of an element of an ordered field S and still obtain an element of the same field. The issue here is not merely that of fractional powers or other algebraic operations, but a more serious one. Imagine the (as yet undefined) real number line. Although the rational numbers are an infinite number of points on the line, there are many "holes" between the rationals. We already know of one "hole" at $\sqrt{2}$, there is another at $\sqrt{3}$, at π, and at e. In fact, in a sense which can be made precise, almost all of the points on the real number line represent irrational numbers.

The completeness axiom is designed to eliminate the possibility of "holes" in the real number line. It does so by more or less bluntly stating that there are no holes. This is the "Dedekind cut" form of the completeness axiom. we have chosen it over other equivalent axioms because it is easy to visualize—even though the "Cauchy sequence" form is perhaps preferable for more advanced analysis courses. A definition is needed before the axiom can be stated.

Definition: Let S_1 and S_2 be subsets of an ordered field S. Then the set S_1 *precedes* S_2 if for every $a \in S_1$ and $b \in S_2$, we have $a \leq b$.

If you imagine the real number line, "S_1 precedes S_2" should be thought of as meaning that all of S_1 is to the left of all of S_2. S_1 and S_2 of course might touch, or might just miss touching.

Completeness Axiom. Let S_1 and S_2 be nonempty subsets of an ordered field S. If S_1 precedes S_2, then there is at least one number $c \in S$ such that c precedes S_2 and is preceded by S_1. In other words, there is (at least) one element of S between S_1 and S_2.

Definition: The set of *real numbers*, $\mathbb{R}$, is a set which satisfies the above axioms of algebra, order, and completeness. Thus, the real numbers is a *complete ordered field.*

This type of definition of $\mathbb{R}$ amounts to saying "we don't know or care what the real numbers *are*, but in any event they have the required properties." If we had used the Cauchy sequence version of the completeness axiom, we would have begun the rational numbers—which we do know—and then defined the real numbers as the set of limits of rational numbers. This would have been somewhat more concrete, but would have involved the difficult concept of limit before we even get off the ground.

From the picture associated with the completeness axiom, we see that it exactly states that the real number line has no holes, for - emotionally speaking—if there were a hole, let S_1 be the set of real numbers to the left of the hole, and S_2 the s et to the right of the hole. Then there would be no real number between S_1 and S_2, since the hole is there, contradicting the completeness axiom.

Let us use the idea of the last paragraph to show that the rational numbers, an ordered field, are *not* complete by exhibiting two sets, one preceding the other, which have no rational number between them. Just let

$$S_1 = \{\, x \colon x > 0,\, x^2 < 2 \,\} \text{ and } S_2 = \{\, x \colon x > 0,\, x^2 > 2 \,\}.$$

The only possible number between S_1 and S_2 is $\sqrt{2}$ —which is irrational. This construction is just what we need to prove the following sample.

Theorem 0.1 *Every non-negative real number $a \in \mathbb{R}$ has a unique non-negative square root.*

Proof: If $a = 0$, then 0 is the square root. If $a > 0$, let $S_1 = \{\, x \colon x > 0,\, x^2 < a \,\}$ and $S_2 = \{\, x \colon x > 0,\, x^2 > a \,\}$. We first show that neither S_1 nor S_2 is empty. Since $(1 + \frac{a}{2})^2 = 1 + a + \frac{a^2}{4} > a$, we know that $(1 + \frac{a}{2}) \in S_2$, so $S_2 \neq \emptyset$. Also $(\frac{a}{1+\frac{a}{2}})^2 < a$ (check this) so that $\frac{a}{1+\frac{a}{2}} \in S_1$ and hence $S_1 \neq \emptyset$. Because S_1 precedes S_2, by the completeness axiom there is a $c \in \mathbb{R}$ between S_1 and S_2. Notice that $c > 0$, since c is preceded by S_1.

It remains to show that $c^2 = a$. By the trichotomy theorem, either $c^2 > a$, $c^2 < a$, or $c^2 = a$. The first two possibilities will be shown to give contradictions. If $c^2 > a$, since $a < (\frac{c^2+a}{2c})^2 < c^2$, we see that $\frac{c^2+a}{2c} \in S_2$ an d precedes c^2, contradicting the property specified in the completeness axiom that c^2 precedes every element of S_2. Similarly the assumption $c^2 < a$, with the inequality $c^2 < (\frac{2ac}{c^2+a})^2 < a$, leads to a contradiction. The only remaining possibility is $c^2 = a$, which shows that c is the desired positive square root of a.

Let us now prove that the positive square root c of a is unique. Assume that there are two positive numbers c_1 and c_2 such that both $c_1^2 = a$ and $c_2^2 = a$. Then

$$0 = c_1^2 - c_2^2 = (c_1 - c_2)(c_1 + c_2)$$

Since $c_1 + c_2 > 0$, we conclude that $c_1 - c_2 = 0$, so $c_1 = c_2$, completing the proof of the theorem.

DEFINITION: The real number M is an *upper bound* for the set $A \subset \mathbb{R}$ if for every $a \in A$, we have $a \leq M$. The number $\mu \subset \mathbb{R}$ is a *least upper bound* (l.u.b) for A if μ is an upper bound for A and no smaller number is also an upper bound for A. *Lower bound* and *greatest lower bound* (g.l.b) are defined similarly. A set $A \subset \mathbb{R}$ is *bounded* if it has both upper and lower bounds.

Theorem 0.2 *Every non-empty bounded set $A \subset \mathbb{R}$ has both a greatest lower bound and a least upper bound.*

PROOF: Observe first that this theorem utilizes the completeness property in that without it, there might have been a "hole" just where the g.l.b. and l.u.b. should be. Since the proofs for the g.l.b. and l.u.b. are almost identical we only prove there is a g.l.b. Let

$$S_1 = \{\, x \colon x \text{ precedes } A \,\}, \text{ and } S_2 = A.$$

By hypothesis $S_2 \neq \emptyset$. Since A is bounded, it has a lower bound m, $m \in S_1$ so $S_1 \neq \emptyset$. By the completeness axiom, there is a $c \in \mathbb{R}$ between S_1 and S_2. It should be obvious that c is both greater than or equal to every element of S_1, and less than or equal to every element of S_2 - so it is the required g.l.b.

DEFINITION: The *closed interval* [a,b] is the set $\{\, x \in \mathbb{R} \colon a \leq x \leq b \,\}$.

The *open interval* (a,b) is the set $\{\, x \in \mathbb{R} \colon a < x < b \,\}$. All we can do is apologize for the multiple use of the parentheses in notation. Please note that sets are not like doors. Some sets, like $(a, b) = \{\, x \in \mathbb{R} \colon a \leq x < b \,\}$ are neither open nor closed.

Theorem 0.3 *(Nested set property). Let $I_1, I_2, \dots$ be a sequence of non-empty closed bounded intervals, $I_n = \{\, x \colon a_n \leq x \leq b_n \,\}$, which are* nested *in the sense $I_1 \supset I_2 \supset I_3 \dots$, so each covers all that follow it. Then there is at least one point $c \in \mathbb{R}$ which lies in all of the intervals, that is, c is in their intersection $c \in \cap_{k=1}^{\infty} I_k$.*

PROOF: Let $S_1 = \{\, x \colon x \text{ precedes some } I_n, \text{ and so all } I_k,\ k \geq n \,\}$
$S_2 = \{\, x \colon x \text{ preceded by some } I_n, \text{ and so all } I_k,\ k \geq n \,\}$.
First, neither S_1 nor S_2 are empty since $a_1 \in S_1$ and $b_1 \in S_2$. Thus by the completeness axiom, there is at least one $c \in \mathbb{R}$ between S_1 and S_2. This c is the required number (complete the reasoning).

If the intervals I_k do not get smaller after, say I_N because $a_N = a_{N+1} = \dots$ and $b_N = b_{N+1} = \dots$, then the whole interval $a_N \leq x \leq b_N$ is caught by the preceding argument. The more common case is there the a_k's strictly increase and the b_k's strictly

decrease. This is what happens when approximating a real number to successively greater accuracy by the decimal expansion. In the case of $\sqrt{2}$ for example,

$I_1 = \{\, x \colon 1 \leq x \leq 2 \,\}$,
$I_2 = \{\, x \colon 1.4 \leq x \leq 1.5 \,\}$,
$I_3 = \{\, x \colon 1.41 \leq x \leq 1.42 \,\}$,
$I_4 = \{\, x \colon 1.414 \leq x \leq 1.415 \,\}$,

and so on, gradually squeezing down on $\sqrt{2}$ to any desired accuracy.

DEFINITION: The sequence $a_n \in \mathbb{R}$, $n = 1, 2, \dots$. of real numbers *converges to the real number c* if, given any $\epsilon > 0$, there is an integer N such that $|a_n - c| < \epsilon$ for all $n > N$. We will then write $a_n \to c$. [In practice no confusion arises for the use of $\to$ to denote both convergence and mappings (cf. 1)].

Again ordinary decimals supply an example, for they allow us to get arbitrarily close to any real number. We could have *defined* the real numbers as all decimals; however there would be a mess avoiding the built-in ambiguity illustrated by $1.9999 \dots = 2.0000 \dots$

Theorem 0.4 *Under the hypotheses of the previous theorem, if in addition the length of I_n tends to zero, $(b_n - a_n) \to 0$, then the number $c \in \mathbb{R}$ found is unique. Furthermore, if $u_k \in I_k$ for all k, that is if $a_k \leq u_k \leq b_k$, then $u_k \to c$ too.*

PROOF: Suppose there were two real numbers c and $\tilde{c}$ in all of the intervals,

$$a_k \leq c \leq b_k \quad \text{and} \quad a_k \leq \tilde{c} \leq b_k \quad \text{for all } k.$$

Rewriting the second inequality as $-b_k \leq -\tilde{c} \leq -a_k$, and adding this to the first inequality, we find that $a_k - b_k \leq c - \tilde{c} \leq b_k - a_k$. Since both sides of this inequality tend to zero, if $c - \tilde{c} \neq 0$, we would have a contradiction.

To prove $u_k \to c$, repeat the above reasoning with $\tilde{c}$ replaced by u_k. We find that $a_k - b_k \leq c - u_k \leq b_k - a_k$. Again both sides of this inequality tend to zero. Now let us fiddle with the ϵ, N definition of limit t o complete the proof. Since $b_n - a_n \to 0$, given any $\epsilon > 0$, there is an N such that $|a_n - b_n| < \epsilon$ for all $n > N$. Thus for any $\epsilon > 0$ and the same N, $|u_n - c| < \epsilon$ for $n > N$, which is the definition of $u_n \to c$.

Theorem 0.5 BOLZANO-WEIERSTRASS. *Every infinite sequence of real numbers $\{\, u_k \,\}$ in a bounded interval I has at least one* subsequence *which converges to a number $c \in \mathbb{R}$.*

PROOF: This one is very clever and picturesque. Watch. Bisect I into two intervals I_1 and $\tilde{I}_1$ of equal length. At least one of I_1 or $\tilde{I}_1$ must contain an infinite number of the $\{\, u_k \,\}$'s. Continuing in this way we obtain a set of nested intervals $I \supset I_1 \supset I_2 \supset \dots$ each of which have an infinite number of the $\{\, u_k \,\}$'s, and the length of I_n tending to zero. From Theorem 3 we conclude that there must be a $c \in \mathbb{R}$ common to all of the intervals. We must now select the subsequence $\{\, u_{k_n} \,\}$ of the $\{\, u_k \,\}$'s which converge to c. Since each I_n contains an infinite number of points of the sequence, we can certainly pick one, say $u_{k_n} \in I_n$. This sequence $\{\, u_{k_n} \,\}$ satisfies the hypotheses of Theorem 4. Thus $u_{k_n} \to c$.

REMARKS: 1. If we also assume I is closed, then we can further assert that $c \in I$. If I is not closed, c may be an end point $\ni I$.

2. If a sequence u_k converges to a $c \in \mathbb{R}$, then every infinite *subsequence* u_{k_n} also converges, and to the same number c.

Theorem 0.6 . *If the sequence* $\{u_k\}$ *converges, it is bounded.*

PROOF: Say $u_k \to \alpha$, and let $\epsilon = 1$ in the definition of convergence. Then there is an N such that $|u_n - \alpha| < 1$ for all $n > N$. Thus, when $n > N$,

$$|u_n| = |u_n - \alpha + \alpha| \leq |u_n - \alpha| + |\alpha| < 1 + |a|$$

Therefore for any k the number $|u_k|$ is bounded by the largest of the $N+1$ numbers $|u_1|$, $|u_2|$, ..., $|u_N|$ and $(1 + |a|)$.

The following theorem shows how to handle algebraic combinations of convergent sequences.

Theorem 0.7 *If* $a_n \to \alpha$ *and* $b_n \to \beta$, *then*
i) $a_n + b_n \to \alpha + \beta$
ii) $a_n b_n \to \alpha\beta$
iii) $\frac{a_n}{b_n} \to \frac{\alpha}{\beta}$ *if both* $b_n \neq 0$, *for all* n, *and if* $\beta \neq 0$.

PROOF: Since the proofs are all similar, we only prove ii). Observe that

$$|a_n b_n - \alpha\beta| = |(a_n b_n - \alpha b_n) + (\alpha b_n - \alpha\beta)| \leq |a_n - \alpha|\,|b_n| + |alpha|\,|b_n - \beta|$$

By Theorem 6, the $|b_n|$'s are bounded, say by B. Since $a_n \to \alpha$, given any $\varepsilon > 0$, there is an N_1 such that $|a_n - \alpha| < \frac{\varepsilon}{2B}$ for all $n > N_1$, and since $b_n \to \beta$, for the same ε there is an N_2 such that $|b_n - \beta| < \frac{\varepsilon}{2|\alpha|}$ for all $n > N_2$. Thus, if n is greater than the larger of N_1 and N_2, $n > \max(N_1, N_2)$, we find that

$$|a_n b_n - \alpha\beta| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

which does the job.

DEFINITION: The sequence $a_1, a_2, \ldots$ of real numbers is said to be *monotone increasing* if $a_1 \leq a_2 \leq a_3 \leq \ldots$, and *monotone decreasing* $a_1 \geq a_2 \geq a_3 \geq \ldots$. Both kinds are called *monotone* sequences.

Theorem 0.8 *Every bounded monotone sequence* $a_1, a_2, \ldots$ *of real numbers converges. In other words, there is an* $\alpha\varepsilon\mathbb{R}$ *such that* $a_n \to \alpha$.

PROOF: We assume the sequence is increasing. The proof for decreasing sequence is identical. Since the sequence is bounded, by Theorem 2 it has a least upper bound $\alpha\varepsilon\mathbb{R}$. We maintain $a_n \to \alpha$. Given any $\varepsilon > 0$, we know that for all n, $a_n < a + \varepsilon$ because α is an upper bound. Since $\alpha - \varepsilon < \alpha$, and α is the l.u.b. of the sequence, we can find an N such that $\alpha - \varepsilon < a_N$. But then, because the sequence is increasing $\alpha - \varepsilon < a_n$ for *all* $n \geq N$. Thus for all $n \geq N$, $a - \varepsilon < a_n < a + \varepsilon$; that is, $|a_n - \alpha| < \varepsilon$ for all $n \geq N$, proving the convergence to α.

We shall close this difficult section with a wonderful procedure for computing the square root of a positive real number. I use it all of the time. It is much easier to understand than the hair-raising method taught in public school.

Theorem 0.9 *For any positive real numbers* A *and* a_0 *the infinite sequence defined by*

$$a_{n+1} = \frac{1}{2}(a_n + \frac{A}{a_n}),\ n = 0, 1, 2, \ldots, \tag{0-2}$$

is monotone decreasing and converges to $\sqrt{A}$. *Moreover, if we let* $b_n = \frac{A}{a_n}$, *then the* b_n*'s are monotone increasing and also converge to* $\sqrt{A}$:

$$b_a \leq b_2 \leq \ldots \leq \sqrt{A} \leq \ldots \leq a_2 \leq a_1$$

PROOF: We first show that $a_k^2 \geq A$ and that $a_{k+1} \leq a_k$,

$$a_k^2 - A = \frac{1}{4}(a_{k-1} + \frac{A}{a_k - 1})^2 - A = \frac{1}{4}(a_{k-1} + \frac{A}{a_k - 1})^2 \geq 0, \text{ so } a_k^2 \geq A.$$

From this, it is easy to see that $a_{k+1} \leq a_k$, for

$$a_k - a_{k+1} = a_k - \frac{1}{2}(a_k + \frac{A}{a_k}) = \frac{a_k^2 - A}{2a_k} \geq 0.$$

Thus $a_1 \geq a_2 \geq \cdots . \geq \sqrt{A}$.

That the a_k^2 converge is an immediate consequence of Theorem 8, since the sequence $\{a_k^2\}$ is a bounded (by A) monotone decreasing sequence. Denoting the limit by α, $a_k^2 \to \alpha$, the proof that $\alpha = A$ is identical to th e reasoning which gave a unique limit in Theorem 4.

Since $b_n = \frac{A}{a_n}$, and the a_n's decrease and are $\geq \sqrt{A}$, then the b_n's increase and are $\leq \sqrt{A}$. This also shows that $b_n \leq a_n$. Since $a_n \to \sqrt{A}$, we have $b_n = \frac{A}{a_n} \to \sqrt{A}$ too.

APPLICATION: We compute $\sqrt{8}$. Take $a_0 = 3$. Then $a_1 = \frac{1}{2}(3 + \frac{8}{3}) = \frac{17}{6}$, and $b_1 = 8 \cdot \frac{6}{17} = \frac{48}{17}$. Similarly, $a_2 = \frac{577}{204}$, $b^2 = \frac{1632}{577}$. This gives $\frac{1632}{577} \leq \sqrt{8} \leq \frac{577}{204}$, or in decimal form

$$2.82842 < \sqrt{8} < 2.82843,$$

astounding accuracy after only two steps. I carried the computations one step further and found

$$2.828427124\ldots. \leq \sqrt{8} \leq 2.828427124\ldots.,$$

where the dots indicate I gave up on the arithmetic, having obtained the exact value as far as the approximation went. Digital computers use this method and related ones for similar computations. It is particularly well adapted to them (and me) since only simple arithmetic operations are involved.

This Theorem 9 gives another proof that every positive real number has a unique positive square root. It is valuable to compare this proof with that of Theorem 1. The main distinction is that the second proof just given is *constructive* it actually shows a way to compute successive approximations to the square root of any number. However, you are justified in asking how we ever found the procedure of equation (0.9) in the first place. The secret is that this formula is a statement of *Newton's method* for finding roots of $f(x) = 0$, applied to the particular function $f(x) = x^2 - A$. See most calculus books for more information about this method. Hopefully, we will have time to discuss this topic later, for it is a constructive way o f proving the existence of a sought after object. The standard existence theorem for ordinary differential equations is a close relative of Newton's method.

Exercises

(1) For the sequences defined below, find which converge, which do not converge but do have at least one convergent subsequence, and which have neither. In all cases $n \in \mathbb{Z}_+$.

(a) $a_n = \frac{1}{n} + 1$

(b) $a_n = \frac{(-1)^n}{n}$

(c) $b_n = e^n$

(d) $a_n = e^{-2n+1}$

(e) $a_n = 1 + n$

(f) $a_n = 2 + (-1)^n$

(g) $a_n = \sqrt{n+1} - \sqrt{n}$

(h) $a_n = \frac{2-3n}{5n+1}$

(i) $a_n = \frac{7^n}{n!}$ (tough, isn't it?)

(j) $s_n = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^n}$.

(2) Prove that if $a_n \to \alpha$ and $b_n \to \beta$, then $(a_n + b_n) \to \alpha + \beta$, where all the letters represent real numbers.

(3) a). Prove *Bernoulli's inequality*

$$(1+h)^n > 1 + nh,\ h \neq 0,\ h > -1,\ n \geq 2.$$

Here $h \in \mathbb{R}$ and $n \in \mathbb{Z}$. I suggest proof by induction.

b). If $s \in \mathbb{R}$, use part a) to prove that

$$a_n \equiv s^n \to \begin{cases} 0 & \text{if } |s| < 1 \\ \infty & \text{if } |s| > 1. \end{cases}$$

[Hint: If $|s| < 1$, write $|s| = \frac{1}{1+h}$, $h > 0$, while if $|s| > 1$, write $|s| = 1 + h$, $h > 0$].

0.6 Appendix: Continuous Functions and the Mean Value Theorem

DEFINITION: : The function $f(x)$ is *continuous at the point* x_0 if, given any $\epsilon > 0$, there is a $\delta(\epsilon) > 0$ such that

$$|f(x) - f(x_0)| < \varepsilon \text{ when } 0 < |x - x_0| < \delta(\epsilon).$$

REMARK: This may be rephrased as

$$\lim_{x \to x_0} f(x) = f(x_0).$$

Note that either statement requires

(1) f be defined at x_0.

(2) $\lim f(x)$ exists.
$x \to x_0$
$x \neq x_0$

(3) the limiting value of f at x_0 is equal to the defined value of f at x_0.

If a function is discontinuous at x_0, it has at least one of the four troubles

(1) Jump discontinuity

(2) Infinite discontinuity

(3) Infinite oscillations

(4) Removable discontinuity.

Here are examples of each trouble at the point $x = 0$.

(1) $f(x) = \begin{cases} 1, & 0 \leq x \\ -1, & x < 0 \end{cases}$

(2) $f(x) = \begin{cases} \frac{1}{x} & x \neq 0 \\ \text{anything}, & \text{say } 1,\ \text{x} = 0 \end{cases}$

(3) $f(x) = \begin{cases} \sin\frac{1}{x} & x \neq 0 \\ \text{anything}, & \text{say } 0,\ \text{x} = 0 \end{cases}$

(4) $f(x) = \begin{cases} x, & x \neq 0 \\ 1, & x = 0 \end{cases}$

Note that a function may oscillate infinitely about a point and still be continuous there. This is illustrated by the everywhere continuous function

$$f(x) = \begin{cases} x \sin \frac{1}{x} & , \quad x \neq 0 \\ 0 & , \quad x = 0 \end{cases}$$

Theorem 0.10 *I. If $f(x)$ is continuous at $x = c$, and $f(c) = A \neq 0$, then $f(x)$ will keep the same sign as $f(c)$ in a suitably small neighborhood of $x = c$.*

Proof: : We construct the desired neighborhood. Assume A is positive. The proof if $A < 0$ is essentially the same. In the definition of continuity, take $\epsilon = A$. Then there is a $\delta > 0$ such that

$$|f(x) - A| < A \quad \text{when} \quad |x - c| < \delta,$$

that is,

$$0 < f(x) < 2A, \quad \text{when} \quad |x - c| < \delta.$$

In other words, $f(x)$ is positive in the interval $|x - c| < \delta$.

Theorem 0.11 *II. If $f(x)$ is continuous at every point of a closed and bounded interval, then there is a constant M such that $|f(x)| \leq M$ throughout the interval. Thus a continuous function in a closed and bounded interval is bounded.*

Proof: : By contradiction. If f is not bounded, there is a sequence of points x_n such that $|f(x^n)| > n$. From that sequence by Theorem 5 (Bolzano-Weierstrass) we can select a subsequence x_{n_k} which converges to some point x_0 in t he interval, $x_{n_k} \to x_0$. Thus

$$|f(x_{n_k})| \to \infty.$$

But we know from the continuity of f that $|f(x_{n_k})| \to |f(x_0)|$. A contradiction.

Moreover, with the *same hypotheses*, we can conclude more.

Theorem 0.12 *III. If f is continuous at every point of a closed and bounded interval, then there are points $x = \alpha$ and $x = \beta$ in the interval where f* assumes *its greatest and least values, respectively.*

Proof: : We show that f assumes its greatest value. The proof for the least value is essentially identical. Let S be the set of all upper bounds for f. By Theorem II S is not empty. Therefore by Theorem 2, S has a g.l.b., call it M_0. Since M_0 is the greatest lower bound of upper bounds for f, there is a sequence x_n such that $\lim_{n\to\infty} f(x_n) \to M_0$. Use Bolzano-Weierstrass to pick a subsequence x_{n_k} of the x_n such that the x_{n_k} converges, say to c. By continuity of f, $\lim_{n_k\to\infty} f(x_{n_k}) = f(x)$. Thus $f(c) = M_0$, so f does assume its greatest value at $x = c$.

Remark: This theorem refers to the *absolute maximum* and *absolute minimum* values.

Examples: The following show that the theorem is not necessarily true if any of the hypotheses are omitted.

(1) $f(x) = x$, $0 < x \leq 1$. No min. (interval not closed).

(2) $f(x) = x$, $x \leq 0$, and $f(x) = \frac{1}{1+x^2}$, all x, both have no min. (the interval is unbounded.)

(3) $f(x) = \begin{cases} x, & 0 \leq x < 3. \\ x - 2, & 3 \leq x \leq 4 \end{cases}$ No max. (function is discontinuous.)

Theorem 0.13 *If $f(x)$ is continuous at every point of a closed and bounded interval $[a, b]$, and if $f(a)$ and $f(b)$ have opposite sign, then there is at least one point $c \in (a, b)$ such that $f(c) = 0$.*

Proof: : Say $f(a) < 0$, $f(b) > 0$. We find one point c, "the largest x such that $f(x) = 0$". Let $S = \{ x \in [a, b] \colon f(x) \leq 0 \}$.

Since $f(a) < 0$, S is not empty. It thus has a l.u.b., c. We prove that $f(c) = 0$. Either $f(c) > 0$, $f(c) < 0$, or $f(c) = 0$. The first two possibilities cannot happen, since by Theorem I, if they did, f would be positive (or negative) in a whole neighborhood of c-violating the fact that c is the l.u.b. of S.

Corollary 0.14 *(*Intermediate value theorem*). Let $f(x)$ be continuous at every point of a closed and bounded interval $[a, b]$, with $f(a) = A$, and $f(b) = B$. Then if C is any number between A and B, there is at least one point c, $a \leq c < b$, such that $f(c) = C$. Thus, f assumes every value between A and B at least once.*

Proof: : Apply Theorem IV to the function $\varphi(x) = C - f(x)$.

Remark: The function may assume values other than just those between A and B. An example is the function $f(x) = x^2$, $-1 \leq x \leq 3$. The theorem requires that it assume all values between $f(-1) = 1$ and $f(3) = 9$. Besides those values , this function also *happens* to assume all values between 0 and 1.

We can offer another proof of

Corollary 0.15 *Every positive number k has a unique positive square root.*

Proof: : Consider $f(x) = x^2 - k$, which is clearly continuous everywhere. Since $f(0) < 0$, and $f(1+\frac{k}{2}) = (1+\frac{k}{2})^2 - k = 1 + \frac{k^2}{4} > 0$, Theorem IV shows that f must vanish somewhere in the interval $0 < x < 1 + \frac{k}{2}$. This is the root. It is the unique *positive* square root, for say there were two positive numbers x and y such that $x^2 - k = 0$ and $y^2 - k = 0$. then $x^2 - y^2 = 0$. Thus, $0 = x^2 - y^2 = (x - 1)(x + y)$. Since $x + y > 0$, we conclude $x - y = 0$, or $x = y$.

Remark: It appears that if a function has the property of Corollary 1, the intermediate value property, then it must be continuous. *This is false.* An example is given by the discontinuous (trouble 3) function

$$f(x) = \begin{cases} \sin \frac{1}{x} & , \quad x \neq 0 \\ 0 & , \quad x = 0 \end{cases}$$

about the point $x = 0$. If a is any number < 0, and b any number > 0, then $f(x)$ assumes every value between $f(a)$ and $f(b)$, but $f(x)$ is not continuous throughout the interval since it is not continuous at $x = 0$.

DEFINITION: The function $f(x)$ has a relative *maximum* (minimum) at the point x_0, if, for all x in a sufficiently small interval containing x_0 as an interior point, we have

$$f(x) \le f(x_0) \qquad (f(x) \ge f(x_0)).$$

REMARK: By convention, we shall agree *not* to call the possible max (or min) at the end point of an interval a relative max (or min). This does lead to the possibility of an absolute max (or min) not being a relative max (or min). However, if the absolute max (or min) does occur at an *interior point* of an interval, it is also a relative max (or min).

DEFINITION: The function $f(x)$ is *differentiable at the point* x_0 if the following limit

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

exists. There are the usual notations: $f'(x_0)$, $\frac{df}{dx}\Big|_{x=x_0}$, $Df(x_0)$.

Theorem 0.16 *If $f(x)$ is differentiable at x_0, then it is continuous there.*

PROOF: : Now if the limit

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

exists, as we have assumed, then the numerator must approach zero as x tends to x_0. Thus f is continuous at x_0.

Theorem 0.17 *If $f(x)$ is differentiable at x_0 and has a relative maximum or minimum at x_0, then $f'(x_0) = 0$.*

PROOF: : Assume f has a relative min at x_0. Then for all x near x_0, $f(x) \ge f(x_0)$.

(i) if $x < x_0$ $\frac{f(x)-f(x_0)}{x-x_0} \le 0$

so $\lim_{\substack{x \to x_0 \\ x < x_0}} \frac{f(x)-f(x_0)}{x-x_0} \le 0$

(ii) if $x > x_0$ $\frac{f(x)-f(x_0)}{x-x_0} \ge 0$

so $\lim_{\substack{x \to x_0 \\ x > x_0}} \frac{f(x)-f(x_0)}{x-x_0} \ge 0$

Because the function is differentiable at x_0, the two limiting values are $f'(x_0)$. Thus $f'(x_)) \le 0$ and $f'(x_0) \ge 0$. Both statements can be true only if $f'(x_0) = 0$. The trick here was, the slope must be negative to the left, and positive to the right of x_0. Since there is a unique slope (the derivative) at x_0, the slope must be zero there. At a relative max., the same proof holds with obvious modifications.

EXAMPLES: 1. Although the function $f(x) = |x|$ has a relative minimum at $x = 0$, the conclusion of the theorem does not hold since f is not differentiable there. Note that both (i) and (ii) of the proof still do hold.

2. The differentiable function (for all x)

$$f(x) = \begin{cases} x^4 \sin \frac{1}{x} & , \quad x \ne 0 \\ 0 & , \quad x = 0 \end{cases}$$

has an infinite number of relative max and min in any interval including the origin.

Theorem 0.18 *(ROLLE). If*

(i) $f(x)$ is continuous at every point of the closed and bounded interval $[a, b]$

(ii) $f(x)$ is differentiable at every point of the open interval (a, b) and

(iii) $f(a) = f(b)$,

then there is at least one point c, $a < c < b$, where $f'(c) = 0$.

PROOF: : If $f(x) \equiv$ constant throughout $[a, b]$, take c to be any point in (a, b). Otherwise $f(x)$ must go either above or below (or both) the value $f(a)$. Assume it goes above. Then by Theorem III there is a point $x = c$ where f has its absolute maximum. Since we assumed $f(x)$ goes above $f(a)$, the point $x = c$ is an interior point. Thus there is a relative maximum. Since f is differentiable in (a, b), we may apply Theorem VI to conclude that $f'(c) = 0$. If we had assumed f went below $f(a)$, then there would have been an absolute (and relative) min. etc.

REMARKS: 1. From the proof of the theorem, we see that if f has values both greater and less than $f(a)$, then there would be at least two points in (a, b) where $f' = 0$.

2. You should be able to construct examples showing the theorem is not true if any of the hypotheses are dropped.

Corollary 0.19 *(MEAN VALUE THEOREM) If*

(i) $f(x)$ is continuous at every point of the closed and bounded interval $[a, b]$ and

(ii) $f(x)$ is differentiable at every point of the open interval (a, b), then there is at least one point c in (a, b) where

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

PROOF: : "Shift and apply Rolle's Theorem". In more detail, consider

$$F(x) = f(x) - f(a) - \frac{x - a}{b - a}(f(b) - f(a)).$$

$F(x)$ satisfies all of the assumption of Rolle's Theorem. Therefore there is a point c where $F'(c) = 0$. Since

$$F'(x) = f'(x) - \frac{f(b) - f(a)}{b - a},$$

at $x = c$, we have

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

REMARKS: 1. The function $f(x) = |x|$ in the interval $[a, b]$, $a < 0$, $b > 0$, shows what happens if the function fails to be differentiable at even one point of the open interval (a, b).

2. An alternative form of the conclusion is: there is a number θ, $0 < \theta < 1$, such that

$$f(b) - f(a) = f'(a + \theta(b - a))(b - a).$$

This is because every point in the interval (a, b) is of the form $a + \theta(b - a)$, for some θ, $0 < \theta < 1$.

We shall now give some applications of the Mean Value Theorem. The first one is a specific example, while the others have great significance in themselves.

EXAMPLE: The function $f(x) = a_1 \sin x + a_2 \sin 2x + b \cos x + b_2 \cos 2x$ has at least one zero in the interval $[0, 2\pi]$, no matter what the coefficients a_1, a_2, b_1 and b_2 are. To show this, we shall show f is the derivative of a function $g(x)$ which satisfies the hypotheses of Rolle's theorem. This function g is just an anti-derivative of f: $g'(x) = f(x)$

$$g(x) = -a \cos x - \frac{a_2}{2} \cos 2x + b_1 \sin x + \frac{b_2}{2} \sin 2x.$$

Since g is clearly continuous and differentiable everywhere, we must only see if $g(0) = g(2\pi)$, which as also easy.

Theorem 0.20 *If $f(x)$ is continuous and differentiable throughout $[a, b]$, and $|f'| < N$ there too, then the $\delta(\epsilon)$ in the definition of continuity can be chosen as $\delta(c) = \frac{\epsilon}{N}$. This δ works for* every *x in $[a, b]$.*

PROOF: : Use the form of the mean value theorem in Remark 2. Then for any points x, x_0 in (a, b),

$$f(x) - f(x_0) = f'(\tilde{x})(x - x_0),$$

where $\tilde{x}$ is somewhere between x and x_0. Thus

$$|f(x) - f(x_0)| \leq N\,|x - x_0|\,.$$

We see now that if $\delta(\epsilon) = \frac{\epsilon}{N}$, then for any $\epsilon > 0$,

$$|f(x) - f(x_0)| < \epsilon \text{ if } |x - x_0| < \delta.$$

Theorem 0.21 *If f satisfies the hypotheses of the mean value theorem and if in addition $f'(x) \equiv 0$ throughout (a, b), then $f(x) \equiv$ const.*

PROOF: : Let x_1 and x_2 be any points on (a, b). Then by the form of the mean value theorem in Remark 2

$$f(x_2) - f(x_1) = 0 \cdot (x_2 - x_1) = 0.$$

Thus $f(x_2) = f(x_1)$ for any two points in (a, b), that is, f is identically constant.

Corollary 0.22 *If $f(x)$ and $g(x)$ both satisfy the hypotheses of the mean value theorem, and if in addition $f'(x) \equiv g'(x)$ for all x in (a, b), then $f(x) = g(x) + c$, where c is some constant.*

PROOF: : consider the function $F(x) = f(x) - g(x)$. It satisfies the hypothesis of Theorem VII, so $F(x) \equiv c$, c constant. Thus $f(x) - g(x) = c$.
REMARK: Theorem IX is the converse of the theorem: "the derivative of a constant function is zero."

A FIGURE GOES HERE

Exercises

(1) Look over all the theorems (and corollaries) here and be sure you can find examples showing that the theorems are not true if any of the hypotheses are relaxed.

(2) Let $f(x) = \begin{cases} 1, & \text{if } x \text{ is a rational number} \\ 0, & \text{if } x \text{ is an irrational number.} \end{cases}$

Is f continuous anywhere?

(3) Let $f(x)$ be an everywhere differentiable function which is zero at $x = a_j$, $j = 1, 2, \ldots, n$. Find a function which vanishes at least once between each of the zeros of f.

(4) Use Theorem VIII to find a $\delta(\epsilon)$ for the given functions.

(a) $f(x) = x^4 - 7$, $-2 \leq x \leq 3$.
(b) $f(x) = x^2 \sin x$, $-4 \leq x \leq 3$
(c) $f(x) = \frac{1}{1+x^2}$, $-2 \leq x \leq 1$
(d) $f(x) = x^{\frac{4}{3}} + 7$, $-2 \leq x \leq 8$
(e) $f(x) = x\sqrt{x^2 + 1}$, $-2 \leq x \leq 2$

(5) (a) The function $f(x)$ satisfies the following condition

$$|f(x) - f(x_0)| \leq 2\,|x - x_0|^3$$

for every pair of points x, x_0 in the interval $[a, b]$. Prove $f(x) \equiv$ constant in this interval.

(b) Generalize your proof to the case when f satisfies

$$|f(x) - f(x_0)| \leq c\,|x - x_0|^{\alpha}\,,$$

where $c > 0$ is some constant and α is any number > 1.

(6) Consider the function $f(x) = x^{\frac{2}{3}}$, in the interval $[-8, 8]$.

Sketch a graph. Note that $f(-8) = f(8) = 4$ but there is no point where $f' = 0$; which hypothesis of Rolle's theorem is violated?

(7) In a trip, the average speed of a car is 180 miles per hour. Prove that at some time during the trip, the speedometer must have registered precisely 180 miles per hour.

(8) Let $P_1 := (x_1, y_1)$ and $P_2 := (x_2, y_2)$ be any two points on the parabola $y = ax^2 + bx + c$, and let $P_3 := (x_3, y_3)$ be the point on the arc P_1P_2 where the tangent is parallel to the chord P_1P_2. Show that

$$x_3 = \frac{x_1 + x_2}{2}.$$

(9) Prove that every polynomial of *odd* degree

$$P(x) = x^{2n+1} + a_{2n}x^{2n} + \cdots + a_1x + a_0$$

has at least one real root.

(10) If f is a nice function and $f' < 0$ everywhere, prove that f is strictly decreasing.

0.7 Complex Numbers: Algebraic Properties

.

In high school, to be able to find the roots of all quadratic equations $ax^2 + 2bx + c = 0$, we were forced to introduce the symbol $i \equiv \sqrt{-1}$, in other words, introduce a special symbol for a root of $x^2 + 1 = 0$. Before going any further, we should prove that no real number c can satisfy $c^2 + 1 = 0$. By contradiction, assume that there is such a c. Then necessarily either $c > 0$, $c < 0$, or $c = 0$. If $c = 0$, we have the immediate contradiction that $1 = 0$. If $c > 0$, or $c < 0$, $0 < c^2$. Consequently $0 < c^2 + 1$ too, which again contradicts $0 = c^2 + 1$, and proves our contention that no real number can satisfy $x^2 + 1 = 0$.

Observe that our proof also shows that if we introduce a new symbol for a root of $x^2 + 1 = 0$, that symbol cannot be an element of an ordered field, for only the ordered field properties of the real numbers were used in the above proof. we shall see that "i" is an element of a field, but not an ordered field.

It is difficult to overestimate the importance of complex numbers for all of mathematics, both from an esthetic as well as from a practical viewpoint. With them we can prove that every quadratic polynomial has exactly two roots (which may coincide). What is more surprising is that every polynomial of order n

$$a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 = 0,\ a_n \neq 0,$$

has exactly n complex roots. This result, the*fundamental theorem of algebra*, was first proved by Gauss in his doctoral dissertation (1799). It is one of the crown jewels of mathematics. The difficult part is proving that every polynomial has *at least one* complex root,

from which the general result follows using only the "factor theorem" of high school algebra. Later on in the semester we shall discuss this more fully and offer a proof. It is not simpleminded, for the proof is non-constructive pure existence proof, giving absolutely no method of finding the roots. Perhaps we shall even prove some more exotic results.

Having gotten carried away, let us retreat and obtain the algebraic rules governing the set $\mathbb{C}$ of complex numbers. In order to reveal the algebraic structure most clearly, we shall denote a complex number z by an ordered pair of real numbers: $z = (x, y)$, x, $y \in \mathbb{R}$. Thus $\mathbb{C}$ is $\mathbb{R} \times \mathbb{R}$ with the following additional algebraic structure.

Definition: If $z_1 = (x_1, y_1)$ and $z_2 = (x^2, y^2)$ are any two complex numbers, then we define

Addition: $z_1 + z_2 = (x_1 + x_2,\ y_1 + y_2)$,

and

Multiplication: $z_1 \cdot z_2 = (x_1x_2 - y_1y_2,\ x_1y_2 + y_1x_2)$.

Equality: $z_1 = z_2$ if and only if both $x_1 = x_2$ and $y_1 = y_2$.

Thus, the complex number zero—the additive identity—is $(0, 0)$, while the complex number one—the multiplicative identity—is $(1, 0)$. Using the fact that the real numbers $\mathbb{R}$ form a field, we can now prove the

Theorem 0.23 *The complex numbers $\mathbb{C}$ form a field.*

Proof: Since the verification of the field axioms are entirely straightforward we give only a smattering. Note that we shall rely heavily on the field properties of $\mathbb{R}$. Addition is commutative:

$$\begin{aligned} z_1 + z_2 &= (x_1, y_1) + (x^2, y^2) = (x_1 + x_2,\ y_1 + y_2) \\ &= (x_2 + x_1,\ y_2 + y_1) = (x_2, y_2) + (x_1, y_1) = z_2 + z_1. \end{aligned} \tag{0-3}$$

Additive identity:

$$0 + z = (0, 0) + (x, y) = (0 + x,\ 0 + y) = (x, y) = z.$$

Multiplicative inverse: For any $z \in \mathbb{C}$, $z \neq (0, 0)$, we must find a $\hat{z} = (\hat{x}, \hat{y}) \in \mathbb{C}$ such that $z\hat{z} = 1$, that is, find real numbers $\hat{x}$ and $\hat{y}$ such that $(x, y)(\hat{x}, \hat{y}) = (1, 0)$. Using the definition o f complex multiplication, this means we must solve the two linear algebraic equations

$$\left.\begin{array}{l} x\hat{x} - y\hat{y} = 1 \\ y\hat{x} + x\hat{y} = 0 \end{array}\right\} \qquad x,\ y \in \mathbb{R},$$

for $\hat{x}$ and $\hat{y} \in \mathbb{R}$. The result is

$$\hat{z} = (\hat{x}, \hat{y}) = (\frac{x}{x^2 + y^2}, \frac{-y}{x^2 + y^2}).$$

We will denote this multiplicative inverse, which we have just proved does exist, by $\frac{1}{z}$ or z^{-1}.

It is interesting to notice that complex numbers of the form $(x,0)$ have the same arithmetic definitions as the real numbers, viz.

$$(x_1,0)+(x_2,0)=(x_1+x_2,0)$$

$$(x_1,0)(x_2,0)=(x_1x_2,0).$$

We can easily verify that all complex numbers of this form $(x,0)$ also form a field, a *subfield* of the field $\mathbb{C}$. On the basis of these last two equations, we can identify a real number x with the complex number $(x,0)$ in the sense th at if we perform any computation with these complex numbers of this form, the result will be the same as if the computation had been performed with the real numbers alone. Thus, numbers of the form $(x,0)\in\mathbb{C}$ are *algebraically equivalent* to the numbers $x\in\mathbb{R}$. The technical term for such an algebraic equivalence is *isomorphic*, much as a term for geometric equivalence is congruent. After identifying the real numbers with complex numbers of the form $(x,0)$, we can say that the field of real numbers $\mathbb{R}$ is *embedded* as a subfield in the field of complex numbers, $\mathbb{R}\subset\mathbb{C}$.

After all this chatter, let us at least convince ourselves that every quadratic equation is solvable if we use complex numbers. First we solve $z^2+1=0$, which may be written as $(x,y)(x,y)+(1,0)=(0,0)$, or as the two real equations $x^2-y^2=-1, 2xy=0$. The last equation says that either $x=0$ or $y=0$. Now if $y=0$, we are left to solve $x^2+1=0$, $x\in\mathbb{R}$, which we know is impossible. Therefore $x=0$ and then $y^2=1$. Thus the two complex numbers $(0,1)$ and $(0,-1)$ both satisfy $z^2+1=0$. The general case, $az^2+bz+c=0$ is easily reduced to the special one by completing the square.

One by-product of the above demonstration is that we see it is foolhardy to try to define an order relation on $\mathbb{C}$ to obtain an ordered field. This is because the equation $x^2+1=0$ cannot be solved in any ordered field, as was shown earlier, whereas we have just solved it in $\mathbb{C}$.

Observe that every $(x,y)\in\mathbb{C}$ can be written as

$$(x,y)=(x,0)(1,0)+(y,0)(0,1),$$

where the complex number $(0,1)$ is called the imaginary unit and is denoted by i. If we now utilize the isomorphism between the real number a and complex numbers $(a,0)$, the last equation shows that (x,y) may be thought of as $x+iy$. Thus, we have obtained the usual notation for complex numbers. From our development, the algebraic role of i as the *symbol* for the imaginary unit $(0,1)$ is hopefully clarified. The number x is called the *real part*, and y the *imaginary part* of the complex number $z=x+iy$. In symbols, $x=Re\{z\}$ and $y=Im\{z\}$.

Our introduction of complex numbers suggests a geometric interpretation. We have defined complex numbers $\mathbb{C}$ as ordered pairs of real numbers, elements of $\mathbb{R}\times\mathbb{R}$, with an additional algebraic structure. Since the points in the plane are also elements of $\mathbb{R}\times\mathbb{R}$, it is clear that there is a one to one correspondence between the complex numbers and the points in the plane. If we plot the point $z=(x,y)$, the real number $|z|$, the "*absolute value* or *modulus* of z" is the distance of the point z from the origin. Its value is computed by the Pythagorean theorem

$$|z|=\sqrt{x^2+y^2}.$$

Here are several formulas which are easily verified:

$$\left.\begin{array}{r}|z_1z_2|=|z_1|+|z_2|\\ |x|\le|z|,|y|\le|z|\\ |z_1+z_2|\le|z_1|+|z_2|\quad\text{(triangle inequality)}\end{array}\right\}\tag{0-4}$$

If the line joining the point z to the origin is drawn, the angle θ between that line and the positive real $(=x)$ axis is called the *argument* or *amplitude* or z. The absolute value r and argument θ of a complex number determine it uniquely, since we have

$$z=r(\cos\theta+i\sin\theta)\tag{0-5}$$

This is the polar coordinate form of the complex number z. Note that conversely, z determines its argument only to within an additive multiple of 2π. This observation will prove of value to us shortly.

Associated with every complex number, $z=x+iy$ there is another complex number $\overline{z}=x-iy$, the *complex conjugate* of z. It is the reflection of z in the real axis. Probably the main reason for introducing $\overline{z}$ is that we can solve for x and y in terms of z and $\overline{z}$:

$$x=\frac{z+\overline{z}}{2},\qquad y=\frac{z-\overline{z}}{2i}.$$

Again some simple formulas:

$$\left.\begin{array}{rr}|\bar{z}|=|z|, & |z|^2=|\bar{z}|^2=z\overline{z}.\\ \overline{(z_1+z_2)}=\overline{z_1}+\overline{z_2}, & \overline{(z_1z_2)}=\overline{z_1}\,\overline{z_2}.\end{array}\right\}\tag{0-6}$$

To illustrate the value of this notation, let us leave the main road to prove the interesting

Theorem 0.24 . *If the complex number γ is a root of the polynomial*

$$P(t)=a_nt^n+a_{n-1}t^{n-1}+\cdots+a_1t+a_0,$$

where the coefficients $a_0,a_1,\ldots,a_n$ are real *numbers, then $\overline{\gamma}$ is also a root of $P(t)$. In other words, the roots of* real *equations occur in conjugate pairs.*

PROOF: Since γ is a root, the complex number

$$P(\gamma)=a_n\gamma^n+\cdots+a_1\gamma+a_0$$

is zero, $P(\gamma)=0$. This implies that its conjugate is also 0, $\overline{P(\gamma)}=0$. By using equations (0.7), we have that

$$\overline{P(\gamma)}=\overline{a_n}\overline{\gamma}^n+\cdots+\overline{a_1}\overline{\gamma}+\overline{a_0},$$

since the coefficients a_j are real, $\overline{a_j}=a_j$. Thus

$$0=\overline{P(\gamma)}=a_n\overline{\gamma}^n+\cdots a_1\overline{\gamma}+a_0=P(\overline{\gamma}),$$

that is, the complex number $\overline{\gamma}$ is a root of the same polynomial.

Now if the proof looks like it was done with mirrors, go over each step carefully. This type of reasoning is somewhat typical of modern mathematics in that it yields information about an object (the roots of a polynomial in this case) *without* first obtaining an explicit formula for the object.

After this digression let us return and find a geometric interpretation for the arithmetic operations on complex numbers. First, addition. The three points z_1, z_2 and $z_1 + z_2$ together with the origin determine a parallelogram. (check this). Thus addition of complex numbers is sometimes called the *parallelogram rule* for additions. Given the points z_1 and z_2, the point $z_1 + z_2$ can be constructed using compass and straight-edge. Subtraction is just $z_1 + (-z_2)$.

Multiplication is much more difficult to interpret geometrically. We shall use equation (0.7) and write $z_j = |z_j|\,(\cos\theta_j + i\sin\theta_j), j = 1, 2$. Then

$$z_1 z_2 = |z_1|\,(\cos\theta_1 + i\sin\theta_1)\,|z_2|\,(\cos\theta_2 + i\sin\theta_2)$$
$$z_1 z_2 = |z_1 z_2|\,[\cos\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2)]. \tag{0-7}$$

Thus the product of z_1 and z_2 has modulus $|z_1 z_2|$ and argument $\theta_1 + \theta_2$: multiply the moduli and add the arguments. This too may be carried out using compass and straight-edge. Since $\frac{1}{z_2} = \frac{1}{|z_2|}(\cos\theta_2 - i\sin\theta_2)$, division reads

$$\frac{z_1}{z_2} = \left|\frac{z_1}{z_2}\right|[\cos(\theta_1 - \theta_2) + i\sin(\theta_1 - \theta_2)],$$

so the moduli are divided while the arguments are subtracted.

We will exploit the multiplication formula (0.7) to find all n complex roots of the specific polynomial

$$z^n = A,$$

for any $A \in \mathbb{C}$. This equation is one of the few whose roots can always be found explicitly. The trick is to write A in its polar coordinate form

$$A = |A|\,[\cos(\alpha + 2k\pi) + i\sin(\alpha + 2k\pi)],$$

where α is the argument of A and k is any integer. Although we get the same A no matter what k is used, as was observed following equation (0.7), we shall retain the arbitrary k since it is the heart of the process we have in mind. From equation (0.7) we see that

$$A^{\frac{1}{n}} = |A|^{\frac{1}{n}}\,[\cos\frac{\alpha + 2k\pi}{n} + i\sin\frac{\alpha + 2k\pi}{n}]$$

in the sense that for any value of the integer $k, (A^{\frac{1}{n}})^n = A$. As k runs through the integers, we get only n different angles of the form $\frac{\alpha+2k\pi}{n}$, since the other angles differ from these n angles by multiples of 2π . For each of these n different angles we obtain a different complex number $A^{\frac{1}{n}}$. These n numbers for $A^{\frac{1}{n}}$ are the desired n roots of $z^n = A$. It is

usually convenient to obtain the angles by letting $k = 0, 1, 2, \ldots, n - 1$, although any n integers which do not differ by multiples of n will do.

An example should help clear the air. We shall find the three cube roots of -2, that is, solve $z^3 = -2$. First,

$$-2 = 2[\cos(\pi + 2k\pi) + i\sin(\pi + 2k\pi)],$$

since the argument of -2 is π while its modulus is 2. Thus, the roots are

$$z = 2^{\frac{1}{3}}[\cos\frac{\pi + 2k\pi}{3} + i\sin\frac{\pi + 2k\pi}{3}],\ k = 0, t_1, t_2 \ldots$$

There are only three values of z possible, no matter what k's are used. These three cube roots of -2 are

$k = 0, 3, 6, \ldots\ z_1 = 2^{\frac{1}{3}}[\cos(\frac{\pi}{3}) + i\sin(\frac{\pi}{3})] = 2^{\frac{1}{3}}(\frac{1}{2} + i\frac{\sqrt{3}}{2})$ $k = 1, 4, 7, \ldots\ z_2 = 2^{\frac{1}{3}}[\cos(\pi) + i\sin(\pi)] = -2^{\frac{1}{3}}$

$k = 2, 5, 8, \ldots\ z_3 = 2^{\frac{1}{3}}[\cos(\frac{5\pi}{3}) + i\sin(\frac{5\pi}{3})] = 2^{\frac{1}{3}}(\frac{1}{2} + i\frac{\sqrt{3}}{2})$.

It is time-saving to observe that the n roots of unity, that is, of $z^n = 1$, can be written down immediately by utilizing the geometric interpretation of multiplication. All of the roots have modulus 1, and so must lie on the unit circle $|z| = 1$. Bisecting the circle into n equal sectors by the radii, the first beginning on the positive x-axis, we find the roots of unity, w_j, at the n successive intersections of these radii with the unit circle. The roots w_j, $j = 1, 2, 3$, of $z^3 = 1$ are illustrated in the figure as the intersections of $\theta = 0$, $\theta = \frac{2\pi}{3}$, and $\theta = \frac{4\pi}{3}$ with $|z| = 1$. Thus $w_1 = \cos 0 + i\sin 0 = 1$, $w_2 = \cos\frac{2\pi}{3} + i\sin\frac{2\pi}{3} = -\frac{1}{2} + i\frac{\sqrt{3}}{2}$, $w_3 = \cos\frac{4\pi}{3} + i\sin\frac{4\pi}{3} = -\frac{1}{2} - i\frac{\sqrt{3}}{2}$.

Exercises

(1) Express the following complex numbers in the form $a + bi$.

  (a) $(1 - i)^2$
  (b) $(2 + i)(3 - i)$
  (c) $\frac{1}{i}$
  (d) $\frac{1+i}{2-i}$
  (e) $\frac{1+i}{1+2i}$
  (f) $i^3 + i^4 + i^{271}$

(2) Compute the absolute values of the complex numbers in Ex. 1.

(3) a) Add $(1+i)$ and $(1+2i)$ using compass and straight-edge.

b) Multiply $(1+i)$ and $(1+2i)$ using compass and straight-edge.

(4) Express in the form $r(\cos\theta + i\sin\theta)$, with $0 \le \theta < 2\pi$:

(a) i
(b) $2i$
(c) $-2i$
(d) 4
(e) -1
(f) $-1+i$
(g) $(1-i)^3$
(h) $\frac{1}{(1+i)^2}$
(i) $\frac{1}{2}(\sqrt{3}+i)$

(5) Determine the

(a) three cube roots of $i, -i$, and of $1+i$,
(b) four fourth roots of -1 and $+2$
(c) six roots of $z^6 = 1$.

(6) Let A be any complex number, $A = |A|\,[\cos\alpha + i\sin\alpha]$, and let $w_1, \ldots, w_n$ be the n roots of $z^n = 1$. Prove that the n roots of $z^n = A$ are

$$z_1 = A^{\frac{1}{n}} w_1,\; z_2 = A^{\frac{1}{n}} w_2, \ldots z_n = A^{\frac{1}{n}} w_n,$$

where

$$A^{\frac{1}{n}} = |A|^{\frac{1}{n}}\,(\cos\frac{\alpha}{n} + i\sin\frac{\alpha}{n})$$

is the principal nth root of A. This shows that the problem of finding the roots of a complex number is essentially reduced to the simpler problem of finding the roots of unity.

(7) Draw a sketch of the following sets of points in the complex plane.

(a) $\{\, z \in \mathbb{C}\colon\, |z-2| \le 1 \,\}$
(b) $\{\, z \in \mathbb{C}\colon\, |z-1+i| \le 2 \,\}$
(c) $\{\, z \in \mathbb{C}\colon\, |z-2| > 3 \,\}$
(d) $\{\, z \in \mathbb{C}\colon 1 \le |z-2| \le 3 \,\}$
(e) $\{\, z \in \mathbb{C}\colon 1 \le |z+i| < 2 \,\}$

0.8 Complex numbers: Completeness Properties, Complex Functions.

We have just considered the algebraic properties of complex numbers. Now we look at infinite sequences of complex numbers. To develop the desired properties of $\mathbb{C}$, we shall utilize those of $\mathbb{R}$.

DEFINITION: The sequence z_n of complex numbers *converges to the complex number* z if, given any $\epsilon > 0$, there is an N such that $|z_n - z| < \epsilon$ for all $n > N$. We shall again write $z_n \to z$.

In order to apply the theorem known for real sequences to complex sequences, the following is vital.

Theorem 0.25 *Let $z_n = x_n + iy_n$, and $z = x+iy$. Then z_n converges to z if and only if both the real and imaginary parts converge to their respective limits. In symbols,*

$$z_n \to z \Longleftrightarrow x_n \to x \text{ and } y_n \to y.$$

PROOF: Since $z_n \to z$, given any $\epsilon > 0$, we can find an N etc. for the z_n's. Now by equation (0.7)

$$|x_n - x| \le |z_n - z| < \epsilon \text{ and } |y_n - y| \le |z_n - z| < \epsilon$$

so both $x_n \to x$ and $y_n \to y$.

Conversely, given any $\epsilon > 0$, we can find an N_1 for the x_n's and an N_2 for the y_n's. Let N be the larger of N_1 and N_2, $N = \max(N_1, N_2)$. This N works for both the x_n and y_n. But

$$|z_n - z| = |x_n + iy_n - x - iy| \le |x_n - x| + |y_n - y| < 2\epsilon.$$

Therefore $z_n \to z$, completing the proof.

This theorem states that a definition is equivalent to some other property. We could thus have used either property as a definition.

Recall that the real numbers were defined so that there would be no"hole" in the real line. This was the completeness property. It guaranteed that if a sequence of real numbers a_n "looked like" they were approaching a limiting value, then indeed th ere is some $a \in \mathbb{R}$ such that $a_n \to a$. The issue here was to avoid the problem of a sequence of rational numbers approaching an irrational number—which is a "hole" if our set just consisted of the rationals. One consequence of the las t theorem is that the set of complex numbers $\mathbb{C}$ is also complete.

Theorem 0.26 . *Every bounded infinite sequence of complex numbers $\{\, z_k \,\}$ has at least one subsequence which converges to a number $z \in \mathbb{C}$. (By bounded, we mean that there is some $r \in \mathbb{R}$ such that $|z_k| < r$ for all k).*

PROOF: Since the $\{\, z_k \,\}$ are bounded, we know $\{\, x_k \,\}$ and $\{\, y_k \,\}$ are also bounded sequences of real numbers. The conclusion is now a consequence of the Bolzano-Weierstrass theorem 5 applied to $\{\, x_k \,\}$ and $\{\, y_k \,\}$, and of theorem 12 just proved. There is a fine point though:

how to get a subsequence of the z_k whose real and imaginary parts both converge. The trick is first to select a subsequence $\{x_{k_j}\} = \{Rez_{k_j}\}$ of the $\{x_k\}$ which converge to some $x \in \mathbb{R}$. Then, from the related subsequence $\{y_{k_j}\} = \{Imz_{k_j}\}$, select a subsequence $\{y_{k_{j_n}}\}$ which converges to some $y \in \mathbb{R}$. Then $\{x_{k_{j_n}}\}$ also converges to $x \in \mathbb{R}$ so $z_{k_{j_n}} \to z$, and we a re done.

With these technical results under our belts, sequences in $\mathbb{C}$ become no more difficult than those in $\mathbb{R}$.

Let us briefly examine the elements of functions of a complex variable. A complex-valued function $f(z)$ of the complex variable z is a mapping of some subset $z \quad U \subset \mathbb{C}$ into the complex numbers $\mathbb{C}$, $f\colon U \to \mathbb{C}$. Two examples are $f(z) = z^2$, and $f(z) = \frac{1}{z}$. Both the domain and range of $f(z) = z^2$ are all of $\mathbb{C}$, while the domain and range of $f(z) = \frac{1}{z}$ are all of $\mathbb{C}$ with the exception of 0.

If f maps $\mathbb{R} \to \mathbb{R}$, like $f(x) = 1 + x$ or $f(x) = e^x$, since $\mathbb{R} \subset \mathbb{C}$, one asks how the domain of definition of f can be extended from $\mathbb{R}$ to $\mathbb{C}$. Of course there are many possible ways to do this, but most of them are entirely artificial. For $f(x) = 1 + x$, the natural extension is $f(z) = 1 + z$, $z \in \mathbb{C}$. Similarly, if $P(x) = \sum_{k=0}^N a_k x^k$ is any polynomial defined for $x \in \mathbb{R}$, the natural extension to $z \in \mathbb{C}$ is $P(z) = \sum_{k=0}^N a_k z^k$. We are thus led to extend $f(x) = e^x$ for $x \in \mathbb{R}$, to $z \in \mathbb{C}$ by defining $f(z) = e^z$. The only problem is that we have absolutely no idea what it means to raise a real number, e, to a complex power. Taylor (power) series are needed to resolve this issue. This will be carried out at the end of Chapter 1.

Continuity of complex functions is defined in a natural way. Let z_0 be an interior point of the set $U \subset \mathbb{C}$ (that is, z_0 is not on the boundary of U).

Definition: The function $f\colon U \to \mathbb{C}$ is continuous at the interior point $a_0 \epsilon U$ if, given any $\epsilon > 0$ there is a $\delta > 0$ such that $|f(z) - f(z_0)| < \epsilon$ for all z in $0 < |z - z_0| < \delta$.

Reasonable theorems like, if f and g are continuous at the interior point $z_0 \epsilon U$, so is the function $f + g$, are true too - with the same proof as was given for real-valued functions of a real variable.

Although we could go on and define the derivative and integral for complex-valued functions $f(z)$ of a *complex* variable, the development would take too much work. For our future purposes, it will be sufficient to define the derivative and integral of a complex-valued function $f(x)$ of the *real* variable x. The first step is to split $f(x)$ into its real and imaginary parts, that is, find real valued functions $u(x)$ and $v(x)$ such that $f(x) = u(x) + iv(x)$. This decomposition c an always be done by taking

$$u(x) = \frac{f(x) + \overline{f(x)}}{2},\ v(x) = \frac{f(x) + \overline{f(x)}}{2i}.$$

Since $u(x) = \overline{u(x)}$ and $v(x) = \overline{v(x)}$, both $u(x)$ and $v(x)$ are real-valued functions. It is clear that $f(x) = u(x) + iv(x)$.

Example: For the functions $f(x) = 1 + 2ix$, we have $\overline{f(x)} = 1 - 2ix$, so

$$u(x) = \frac{(1 + 2ix) + (1 - 2ix)}{2} = 1,\ v(x) = \frac{(1 + 2ix) - (1 - 2ix)}{2i} = 2x.$$

as expected.

Because $f(x)$ is a complex number for every x in the domain where f is defined, we

$$|f(x)| = \sqrt{u^2(x) + v^2(x)}.$$

With this notion of absolute value, the definitions of continuity and differentiability read just as if f were itself real-valued. For example

Definition: : The complex-valued function $f(x)$ of the real variable x is *differentiable* at the point x_0 if

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

exists.

A more convenient way of dealing with the derivative is supplied by the following

Theorem 0.27 . *The function $f(x) = u(x) + iv(x)$ is differentiable at a point x_0 if and only if both $u(x)$ and $v(x)$ are differentiable there, and*

$$\frac{df}{dx} = \frac{du}{dx} + i\frac{dv}{dx}.$$

Proof: We shall use Theorem 12. Let $\{x_n\}$ be any sequence whose limit is x_0. Define the sequences $\{a_n\}, \{\alpha_n\}$, and $\{\beta_n\}$ by

$$a_n = \frac{f(x_n) - f(x_0)}{x_n - x_0},$$

$$\alpha_n = \frac{u(x_n) - u(x_0)}{x_n - x_0}, \text{ and } \beta_n = \frac{v(x_n) - v(x_0)}{x_n - x_0}.$$

We must show that $\lim_{n\to\infty} a_n$ exists if and only if both limits $\lim_{n\to\infty} \alpha_n$ and $\lim_{n\to\infty} \beta_n$ exist, for the existence of these limits is equivalent to the existence of the respective derivatives. But notice that $a_n = \alpha_n + i\beta_n$, since

$$a_n = \frac{f(x_n) - f(x_0)}{x_n - x_0} = \frac{u(x_n) + iv(x_n) - (u(x_0) + iv(x_0))}{x_n - x_0} = \alpha_n + i\beta_n.$$

Thus we can appeal to Theorem 12 to conclude that $\lim a_n$ exists if and only if both $\lim \alpha_n$ and $\lim \beta_n$ exist. The formula $f' = u' + iv'$ is an immediate consequence since

$$a_n \to f'(x_0),\ \alpha_n \to u'(x_0),\ \text{ and } \beta_n \to v'(x_0)$$

Examples:

a) If $f(x) = 1 + 2ix$, $\frac{df}{dx} = \frac{d}{dx}1 + i\frac{d}{dx}2x = 2i$

b) If $f(\theta) = \cos 7\theta + i \sin 7\theta + 2\theta - i\theta^2$

$\frac{df}{d\theta} = \frac{d}{d\theta}[2\theta + \cos 7\theta] + i\frac{d}{d\theta}[-\theta^2 + \sin 7\theta] = 2 - 7\sin 7\theta + i[-2\theta + 7\cos 7\theta]$

A related result which is even easier to prove is

Theorem 0.28 . *The complex-valued function $f(x) = u(x) + iv(x)$, $x\epsilon\mathbb{R}$ is continuous at $x_0\epsilon\mathbb{R}$ if and only if both $u(x)$ and $v(x)$ are continuous at x_0.*

PROOF: An exercise.

Integration is defined more directly.

DEFINITION: Let $f(x) = u(x)+iv(x)$, $x\epsilon\mathbb{R}$. If the real-valued functions $u(x)$, and $v(x)$ are integrable for $x\epsilon[a, b]$, we define the *definite integral* of $f(x)$ by

$$\int_a^b f(x)\,dx = \int_a^b u(x)\,dx + i\int_a^b v(x)\,dx.$$

The standard theorems, like if c is any complex constant, then

$$\int_a^b cf(x)\,dx = c\int_a^b f(x)\,dx, \text{ and, if } a \le b, \left|\int_a^b f(x)\,dx\right| \le \int_a^b |f(x)|\,dx$$

are proved by using the definition above and the corresponding theorems for real functions. We shall, however, need the more difficult

Theorem 0.29 . *If the complex-valued function $f(t) = u(t)+iv(t)$, $t\epsilon\mathbb{R}$, is continuous for all $t\epsilon[a, b]$, then there is a constant K such that $|f(t)| \le K$ for all $t\epsilon[a, b]$. Furthermore if x, $x_0\epsilon[a, b]$, the n*

$$\left|\int_{x_0}^x f(t)\,dt\right| \le K\,|x - x_0|\,. \tag{0-8}$$

Notice that the left-hand side absolute value is in the sense of complex numbers.

PROOF: Since $f(t)$ is continuous in $[a, b]$, by Theorem 15 so are both $u(t)$ and $v(t)$. But a real-valued function which is continuous in a closed and bounded interval is bounded. Thus there are constants K_1 and K_2 such that $|u(t)| \le K_1$, $|v(t)| \le K_2$ for all $t\epsilon[a, b]$. then

$$|f(t)| = \sqrt{u^2(t) + v^2(t)} \le \sqrt{K_1^2 + K_2^2} \equiv K.$$

To prove the inequality (0.29), we use the inequality mentioned before the theorem to see that if $x_0 \le x$

$$\left|\int_{x_0}^x f(t)\,dt\right| \le \int_{x_0}^x |f(t)|\,dt.$$

Since $|f(t)| \le K$, we find that

$$\int_{x_0}^x |f(t)|\,dt \le K\,|x - x_0|\,.$$

Combining these last two inequalities, we obtain the desired inequality (0.29) if $x_0 \le x$. The other case, $x \le x_0$, can be reduced to that already proved by observing that

$$\left|\int_{x_0}^x f(t)\,dt\right| = \left|-\int_{x_0}^x f(t)\,dt\right| = \left|\int_{x_0}^x f(t)\,dt\right| \le K\,|x_0 - x| = K\,|x - x_0|\,.$$

Exercises

(1) In the complex sequences below, which ones converge, which do not converge but have at least one convergent subsequence, and which do neither? In all cases $n = 1, 2, 3, \ldots$.

  (a) $z_n = \frac{i}{n} + 3i - 4$
  (b) $z_n = 2i + (-1)^n$
  (c) $z_n = n - i$
  (d) $z_n = i^n$
  (e) $z_n = 1 + i\sqrt{3} - \frac{(-1)^n}{7n}$
  (f) $z_n = \frac{(4+6i)n-5}{1-2ni}$.

(2) Write the following complex-valued functions $f(x)$ of the real variable x as $f(x) = u(x) + iv(x)$, where u and v and real-valued.

  (a) $f(x) = i + 2(3 - 2i)x^2$,
  (b) $f(x) = (1 + 2ix)^2$
  (c) $f(x) = \cos 3x^2 - (3 + i)\sin x$
  (d) $f(x) = \frac{1}{1+2i-x}$

(3) (a) Use the definition of the derivative to compute $\frac{df}{dx}$ for the function in Exercise 2a above.

  (b) Find $\frac{df}{dx}$ for all the functions in Exercise 2 above.

(4) Evaluate

  (a) $\int_{-1}^3 (1 + 2ix)\,dx$
  (b) $\int_1^4 [x + (1 - i)\cos 2x]\,dx$

Chapter 1

Infinite Series

1.1 Introduction

In elementary calculus you have met the notion of the limit of a sequence of numbers (see also Chapter 0, sections 5 and 7). This concept of limit is just what essentially distinguishes calculus from algebra. It was crucial in the definition of the derivative as the limit of a difference quotient and the integral as the limit of a Riemann sum. We now propose to discuss another limiting process, *infinite* series, in detail.

An infinite series is a sum of the form

$$\sum_{k=1}^{\infty} a_k = a_1 + a_2 + a_3 + \cdots , \tag{1-1}$$

where the a_k's are real or complex numbers. Since there is no added difficulty we shall suppose the a_k's are complex numbers. One immediate trouble is that it would take us an infinite amount of time to add an infinite sum. For example, what is

(a) $\sum_{k=1}^{\infty} 1 = 1+1+1+1+\cdots \qquad = ?$
(b) $\sum_{k=1}^{\infty} 1 = 1-1+1-1+1-1\cdots \qquad = ?$
(c) $\sum_{k=1}^{\infty} \frac{1}{2^{k-1}} = 1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\cdots = ?$

Thus, we are faced with the realization that be sum (1) is not really well defined, even in cases where we feel it might make sense.

Our first task is to give a more adequate definition. Let S_n be the sum of the first n terms:

$$S_n := a_1 + a_2 + \cdots + a_n = \sum_{k=1}^{n} a_k.$$

Then for each n, we have a complex number S_n, called the n^{th} *partial sum* of the series (1).

Definition: If $\lim_{n\to\infty} S_n = S$, where S is a (finite) complex number, we say that the *infinite series converges* to S. If the sequence $S_1, S_2, S_3, \ldots$ has no limit, we say that the infinite series *diverges*.

For the examples given just above, we have

(a) $S_n = \sum_1^n 1 = n \to \infty$ so the infinite series diverges to ∞.

(b) $S_n = \sum_1^n (-1)^{n+1} = \left\{ \begin{array}{lll} 1 & n & \text{odd,} \\ 0 & n & \text{even,} \end{array} \right\}$. which does not have a limiting value since it oscillates between 1 and 0.

(c) $S_n = \sum_1^n \frac{1}{2^{k-1}} = 2(1 - \frac{1}{2^n}) \to 2$, so the infinite series converges to the number 2 (we found the sum of the series by realizing it is a simple geometric series:

$$1 + r + r^2 + \cdots + r^N = \frac{1 - r^{N+1}}{1 - r}) \quad \text{for } (r \neq 1).$$

With an adequate definition of convergence of infinite series, it is clear that we should develop some tests for determining if a given series converges. That will be done in the next section. In preparation, let us examine some simple types of series which occur often and prove a few useful theorems.

There are two types of series whose sums can always be found, and for which the question of convergence is exceedingly elementary.

Definition: An infinite *geometric series* is a series of the form

$$\sum_{k=0}^{\infty} ar^k = a + ar + ar^2 + \cdots .$$

The partial sums are

$$S_n = a + ar + \cdots + ar^n = a\frac{1 - r^{n+1}}{1 - r} \quad \text{for } (r \neq 1).$$

Theorem 1.1 *The infinite geometric series* $\sum_{k=0}^{\infty} ar^k$, $a \neq 0$, *converges if and only if* $|r| < 1$. *Then the sum is* $\frac{a}{1-r}$.

Proof: $\lim_{n\to\infty} r^{n+1}$ exists only if $|r| < 1$. Then the limit is zero so $\lim_{n\to\infty} S_n = \frac{a}{1-r}$ (the non-convergence when $|r| = 1$ follow from Theorem 6, p. ?)

Examples:

1. $\sum_{k=0}^{\infty} (1+i)^k$ diverges since $|1+i| = \sqrt{2} \geq 1$.

2. $\sum_{k=0}^{\infty} (\frac{1+i}{2})^k$ converges since $\left|\frac{1+i}{2}\right| = \frac{\sqrt{2}}{2} < 1$. The sum of this series is $1 + i$.

3. $\sum_{k=1}^{\infty} 1$ diverges since $|1| = 1$.

4. $\sum_{k=1}^{\infty} (-1)^k$ diverges since $|-1| = 1$.

Definition: An infinite *telescopic series* is one of the form

$$\sum_{k=1}^{\infty} (\alpha_k - \alpha_{k+1}) = (\alpha_1 - \alpha_2) + (\alpha_2 - \alpha_3) + (\alpha_3 - \alpha_4) + \cdots .$$

It is clear that most of the terms cancel each other.

Theorem 1.2 *If* $\alpha_k \to \alpha$, *then* $\sum_{k=1}^{\infty} (\alpha_k - \alpha_{k+1}) = \alpha_1 - \alpha$.

Proof: $S_n = (\alpha_1 - \alpha_2) + (\alpha_2 - \alpha_3) + \cdots + (\alpha_n - \alpha_{n+1}) = \alpha_1 - \alpha_{n+1} \to \alpha_1 - \alpha$.

Examples:

(a) $\frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \frac{1}{3\cdot 4} + \cdots = \sum_{k=1}^{\infty} \frac{1}{k(k+1)} = \sum_{k=1}^{\infty} (\frac{1}{k} - \frac{1}{k+1}) = 1$.

(b) $\frac{1}{4\cdot 1^2 - 1} + \frac{1}{4\cdot 2^2 - 1} + \frac{1}{4\cdot 3^2 - 1} + \cdots = \frac{1}{2} \sum_{k=1}^{\infty} (\frac{1}{2k-1} - \frac{1}{2k+1}) = \frac{1}{2}$

We close this section with some reasonable (and desirable) theorems. The proofs are immediate consequences of the definition of convergence of infinite series and the related theorems about limits of sequences of numbers.

Theorem 1.3 . *If* $\sum_{k=1}^{\infty} a_k \to a$, *and* c *is any number then* $\sum_{k=1}^{\infty} ca_k \to ca$.

Theorem 1.4 *If* $\sum_{k=1}^{n} a_k \to a$ *and* $\sum_{k=1}^{n} bk \to b$, *then* $\sum_{k=1}^{n} (a_k + b_k) \to a + b$.

Theorem 1.5 *Let* $a_k = \alpha_k + i\beta_k$, *where* α_k *and* β_k *are real. The infinite series* $\sum a_k$ *converges if and only if the two real series* $\sum \alpha_k$ *and* $\sum \beta_k$ *both converge. That is, an infinite complex series converges if and only if* both *its real and imaginary parts converge.*

Proof: We must look at the partial sums. Let $\sigma_n = \sum_{k=1}^{n} \alpha_k$, and $\tau_n = \sum_{k=1}^{n} \beta_k$. Then

$$S_n = \sum_{k=1}^{n} \alpha_k = \sum_{k=1}^{n} (\alpha_k + i\beta_k) = \sum_{k=1}^{n} \alpha_k + i \sum_{k=1}^{n} \beta_n = \sigma_n + i\tau_n.$$

But we know from Theorem 12 of Chapter 0 that the complex sequence S_n converges if and only if both its real part, σ_n, and imaginary part, τ_n, both converge—in other words, if the series $\sum a_k$ and $\sum \beta_k$ both converge.

Two remarks should be made in an attempt to mitigate some confusion. First, the index k of the series $\sum_{k=1}^{\infty} a_k$ could have been any other letter. Thus $\sum_{k=1}^{\infty} a_k = \sum_{j=1}^{\infty} a_j$. This is perhaps indicated most clearly if we left an empty box instead of using any letter at all: . The connecting line means that the *same* letter must be used in both boxes. Now you can fill in any letter that makes you happy. No matter w hat you write, it still means $a_1 + a_2 + a_3 + \cdots$. In a similar way, the index need not begin with 1. Thus, for example, $\sum_{k=1}^{\infty} a_k = \sum_{k=17}^{\infty} a_{k-16} = a_1 + a_2 + \cdots$. Although this manipulation looks like unwanted silliness here, it is sometimes quite useful. Later on this year you will need it. The related transformation for integrals is illustrated by

$$\int_2^3 \frac{1}{t}\, dt = \int_1^2 \frac{1}{t+1}\, dt.$$

Exercises

1. Find a closed form expression for the n^{th} partial sum of the following infinite series and determine if they converge.

   a) $\frac{2}{3} + \frac{2}{9} + \frac{2}{27} + \cdots + \frac{2}{3n} + \cdots = \sum_{k=1}^{\infty} \frac{2}{3^k}$.

   b) $1 + i + i^2 + i^3 + \cdots + i^n + \cdots$

   c) $\frac{1}{2!} + \frac{2}{3!} + \frac{3}{4!} + \cdots = \sum_{k=2}^{\infty} \frac{k-1}{k!} = \sum_{k=2}^{\infty}(\frac{1}{(k-1)!} - \frac{1}{k})$

   d) $\ln\frac{1}{2} + \ln\frac{2}{3} + \ln\frac{3}{4} + \cdots + \ln(\frac{n}{n+1}) + \cdots$

   e) $\sum_{m=0}^{\infty}(\frac{3-4i}{7})^m$

   f) $\sum_{n=1}^{\infty} \frac{2-3i}{n(n+1)}$

2. The repeating decimal $1.565656\cdots$ can be written as

$$1 + \frac{56}{10^2} + \frac{56}{10^4} + \frac{56}{10^6} + \cdots = 1 + 56\sum_{k=1}^{\infty}(\frac{1}{10^2})^k.$$

   Sum the geometric series and find what rational number the repeating decimal represents. In a similar way, every decimal which begins to repeat eventually is a rational number. What rational number is represented by 1.4723?

3. A ball is dropped from a height of 20 feet. Every time it bounces, it rebounds to $\frac{3}{4}$ of its height on the previous bounce. What is the total distance traveled by the ball?

4. If $\sum_{k=1}^{\infty} a_k \to a$ and $\sum_{k=1}^{\infty} b_k \to b$, and if α and β are any numbers, prove that $\sum_{k=1}^{\infty}(\alpha a_k + \beta b_k) \to \alpha a + \beta b$.

5. If $a_n > 0$ and $\sum a_n$ converges, prove that $\sum \frac{1}{a_n}$ diverges.

6. Does the convergence of $\sum_{n=1}^{\infty} a_n$ imply the convergence of $\sum_{n=1}^{\infty}(a_n + a_{n+1})$?

7. (a) If the partial sums of $\sum a_n$ are bounded, and $\{ b_n \}$ is a strictly decreasing sequence with limit 0, $b_n \searrow 0$, prove that $\sum a_n b_n$ converges.

   (b) Use (a) to prove that if $\sum_{n=1}^{\infty} n a_n$ converges then so does the series $\sum_{n=1}^{\infty} a_n$.

   (c) Use (a) to discuss the convergence of $\sum_{n=1}^{\infty} \frac{\sin nx}{n}$.

1.2 Tests for Convergence of Positive Series

Tests to determine convergence are of several types, i) those that give sufficient conditions, ii) those that give necessary conditions, and iii) those that give both necessary and sufficient conditions. Theorem 1 of the last section governing geometric series was of the last type; however it is more common to find convergence tests of the first two types since they are usually easier to come by. You should be careful to observe the nature of a test . A simple theorem should make the point clear.

Theorem 1.6 . *If the series $\sum_{k=1}^{\infty} a_k$ —where a_k may be complex—converges, then*

$$\lim_{k\to\infty} |a_k| = 0.$$

PROOF: Let $S_n = a_1 + a_2 + \cdots + a_n$. Then $|a_n| = |S_n - S_{n-1}|$. As $n \to \infty$ both S_n and S_{n-1} tend to the same limit, so $|a_n| \to 0$.

Returning to the point made before, this theorem states a necessary but not sufficient (as we shall see) condition for an infinite series to converge. We can apply it to see that $\sum \frac{k}{k+1}$ diverges—since $\frac{k}{k+1} \to 1 \neq 0$. Thus this theorem is useful as a quick crude test to weed out series which diverge badly. But all it tells us about the series $\sum_{k=1}^{\infty} \frac{1}{k}$—for which $\frac{1}{k} \to 0$ so the criterion of the theorem is satisfied—is that it *might* converge. In fact, this series diverges too, as we shall now prove.

$$\sum_{k=1}^{\infty}\frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \cdots + \frac{1}{8} + \frac{1}{9} + \cdots + \frac{1}{16} + \frac{1}{17} + \cdots + \frac{1}{32} + \cdots$$

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{8} + \frac{1}{16} + \cdots + \frac{1}{16} + \frac{1}{32} + \cdots + \frac{1}{32} + \cdots$$

$$= 1 + \frac{1}{2} + \frac{1}{2} \quad + \quad \frac{1}{2} \quad + \quad \frac{1}{2} \quad + \quad \frac{1}{2} \quad + \cdots ..$$

Thus $S_1 = 1$, $S_2 = 1+\frac{1}{2}$, $S_4 > 1+\frac{1}{2}+\frac{1}{2} = 1+2\cdot\frac{1}{2}$, $S_8 > 1+3\cdot\frac{1}{2}$, $S_{16} > 1+4\cdot\frac{1}{2}, \ldots, S_{2^n} > 1+n\cdot\frac{1}{2}$. We can easily see that as $n \to \infty$, $S_{2^n} \to \infty$, so the series $\sum \frac{1}{k}$, called the *harmonic series*, diverges.

For the many series which slip through the test of Theorem 6, more refined criteria are needed. The criteria we shall present in the remainder of this section are for series with *positive* terms, $a_n \geq 0$. Application of these criteria to series with complex terms will be made in the next section.

Theorem 1.7 . *If $a_k \geq 0$ for each k, then the series $\sum_{k=1}^{\infty} a_k$ converges if and only if the sequence of partial sums is bounded from above.*

PROOF: Since all the a_k's are non-negative, $S_{n+1} \geq S_n$. Thus the S_n's are a monotone increasing sequence of real numbers. By Theorems 6 and 8 of Chapter 0, this sequence S_n converge if and only if it is bounded.

EXAMPLE: The series $\sum_{k=1}^{\infty} \frac{1}{k!}$ of positive terms converges, since

$$\frac{1}{k!} = \frac{1}{1 \cdot 2 \cdot 3 \cdots .k} \leq \frac{1}{1 \cdot 2 \cdot 2 \cdot 2 \cdots .2} = \frac{1}{2^{k-1}}$$

so

$$S_n = \sum_{k=1}^{n} \frac{1}{k!} \leq \sum_{k=1}^{n} \frac{1}{2^{k-1}} \leq \sum_{k=0}^{\infty} \frac{1}{2^k} = 2.$$

The convergence now follows since S_n is bounded from above.

We can extract an exceedingly useful idea from these examples: check the convergence of a given series by comparing it with another series which we know to converge or diverge.

Theorem 1.8 . (COMPARISON TEST) *Let $\sum a_k$ and $\sum b_k$ be two* positive *series for which $a_k \leq b_k$ for $n > N$. Then*
i) if $\sum b_k$ converges, so does $\sum a_k$.
ii) if $\sum a_k$ diverges, so does $\sum b_k$.

PROOF: Let $s_n = \sum_{k+N+1}^{n} a_k$ and $t_n = \sum_{k+N+1}^{n} b_k$. Then $s_n \leq t_n$ for all $n > N$, so i) if $t_n \to t$, then s_n is bounded ($s_n \leq t$), ii) if $s_n \to \infty$, then $t_n \to \infty$ too.
REMARK: The "$n > N$" part of the hypothesis reflects the fact that it is only the infinite tail of an infinite series that we need to worry about. Any *finite* number of terms can always be added later on.

EXAMPLES:
(a) $\sum_{k=1}^{\infty} \frac{1}{2^k+1}$ converges since $\frac{1}{2^k+1} < \frac{1}{2^k}$ and $\sum \frac{1}{2^k}$ converges.
(b) $\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}}$ diverges since $\frac{1}{\sqrt{k}} \geq \frac{1}{k}$ (for $k \geq 1$) and $\sum \frac{1}{k}$ diverges.
Our next test is based upon comparison with a geometric series $\sum r^n$.

Theorem 1.9 . (RATIO TEST) *Let $\sum a_n$ be a series with positive terms such that the following limit exists*

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = L.$$

Then
i) if $L < 1$, the series converges
ii) if $L > 1$, the series diverges
iii) if $L = 1$, the test is inconclusive.

REMARK: If the assumed limit does not exist, a variant of the theorem is still true but we shall not discuss it.
PROOF: i) If $L < 1$, pick any r, $L < r < 1$. Then there is an N such that for all $n \geq N$, $\frac{a_{n+1}}{a_n} < r$. Therefore $a_n < ra_{n-1} < r^2 a_{n-2} < \ldots < r^{n-N} a_N$, so that $a_n < Kr^n$, $n \geq N$, where $K > \frac{a_N}{r^N}$. The series $\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{N-1} a_n + \sum_{n=N}^{\infty} a_n$ consists of a finite sum plus an infinite tail which is dominated by the geometric series $\sum Kr^n$. Since $r < 1$, the geometric series converges and by the comparison test, so does $\sum a_n$.

ii) If $L > 1$, then $a_{n+1} > a_n$ for all $n > N$; thus $\lim_{n\to\infty} a_n \neq 0$. By Theorem 6, the series $\sum a_n$ cannot converge.
iii) This is seen from the two examples.
(a) $\sum \frac{1}{n}$, with $\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \lim_{n\to\infty} \frac{n}{n+1} = 1$, which we know diverges.
(b) $\sum \frac{1}{n(n+1)}$, with $\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \lim_{n\to\infty} \frac{(n+1)(n+2)}{n(n+)} = 1$, which we know (Theorem 2, Example a) converges.
In both these cases $L = 1$. You should notice that the criterion uses the *limiting value* of a_{n+1}/a_n. The divergent harmonic series $\sum \frac{1}{n}$, whose ratio $n/n+1$ is less than one for finite n, but 1 in the limit shows the mistake you will make if you use the ratio before passing to the limit.

EXAMPLES:

1. $\sum \frac{1}{n!}$: Since $\lim_{n\to\infty}(\frac{a_{n+1}}{a_n}) = \lim_{n\to\infty}(\frac{n!}{(n+1)!}) = \lim_{n\to\infty} \frac{1}{n+1} = 0 < 1$, the ratio is less than one so the series converges.

2. $\sum \frac{10^n}{n!}$: *Since* $\lim_{n\to\infty}(\frac{a_{n+1}}{a_n}) = \lim_{n\to\infty}(\frac{10}{n+1}) = 0 < 1$, the series converges.

3. $\sum \frac{n!}{2^n}$: *Since* $\lim_{n\to\infty}(\frac{a_{n+1}}{a_n}) = \lim_{n\to\infty}(\frac{n+1}{2}) = \infty$, the series diverges.

Our last test for series with positive terms is associated with a picture. The crux of the matter is very simple and clever. We associate an area with the infinite series $\sum_{n=1}^{\infty} a_n$. For the term a_n we use a rectangle between $n \leq x \leq n+1$ of height a_n and base one. Then the sum of the infinite series is represented by total area under the rectangles. Now by Theorem 7, if all the a_n's are positive we know the series converges if the total area is finite. Thus, if we can find a function $f(x)$ whose graph lies above the rectangles, and whose total area is finite, then we know the area contained in the rectangles is finite and so the series converges.

Theorem 1.10 . (INTEGRAL TEST) *Let $\sum_{n=i}^{\infty} a_n$ be a series of positive decreasing terms: $0 < a_{n+1} \leq a_n$, and $f(x)$ a continuous decreasing function with $f(n) = a_n$. Then the sequence*

$$S_N = \sum_{n=1}^{N} a_n \text{ and } T_N = \int_1^N f(x)\,dx$$

either both converge or both diverge, in fact, $S_N - a_1 \leq T_N \leq S_{N-1}$.

PROOF: First of all,

$$\int_1^N f(x)\,dx = \int_1^2 + \int_2^3 + \cdots + \int_{N-1}^N = \sum_{n=1}^{N-1} \int_n^{n+1} f(x)\,dx.$$

Since in the interval $n \leq x \leq n+1$ we know that

$$a_n = f(n) \geq f(x) \geq f(n+1) = a_{n+1},$$

we see that

$$a_n = \int_n^{n+1} f(n)\,dx \geq \int_n^{n+1} f(x)\,dx \geq \int_n^{n+1} f(n+1)\,dx = a_{n+1}.$$

Adding these up, we find

$$\sum_{n=1}^{N-1} a_n \geq \sum_{n=1}^{N-1} \int_n^{n+1} f(x)\,dx \geq \sum_{n=1}^{N-1} a_{n+1}$$

or

$$\sum_{n=1}^{N-1} a_n \geq \int_1^{N} f(x)\,dx \geq \sum_{n=2}^{N} a_n.$$

Thus

$$S_{N-1} \geq T_N \geq S_N - a_1.$$

From this last inequality, we see that $\lim_n \to \infty T_N$ is finite if and only if $\lim_{n\to\infty} S_N$ is finite. Since the sequences S_N and T_N are both monotone increasing sequences, by Theorem 7 the sequences converge or diverge together. And we are done.

EXAMPLES:

(a). $\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges if $p > 1$, diverges if $p \leq 1$. We use the function $f(x) = \frac{1}{x^p}$, which satisfies the hypothesis of the theorem, and examine the integral

$$T_N = \int_1^N \frac{1}{x^p}\,dx = \left\{ \begin{array}{cc} \frac{N^{1-p}-1}{1-p} & , \quad p \neq 1. \\ \ln N & , \quad p = 1 \end{array} \right\}.$$

As $N \to \infty$, $\ln N \to \infty$, and so does N^{1-p} if $p < 1$, while $N^{1-p} \to 0$ if $p > 1$. Therefore T_N converges if and only if $p > 1$, so by our theorem $\sum_{n=1}^{\infty} \frac{1}{n^p}$converges if and only if $p > 1$. In the special case $p = 1$ we have again proven that the harmonic series $\sum \frac{1}{n}$ diverges. Another often seen special case is $p = 2$, $\sum \frac{1}{n^2}$, which converges. Sometime later we shall prove the amazing $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$.

(b) $\sum_{n=2}^{\infty} \frac{1}{n \ln n}$ diverges since as $N \to \infty$, $\int_2^N \frac{dx}{x \ln 2} = \ln(\ln N) - \ln(\ln 2) \to \infty$

Exercises

1. Determine if the following series converge or diverge.
   a) $\sum_{n=1}^{\infty} \frac{1}{n^2+1}$
   b) $\sum_{n=1}^{\infty} \frac{1}{2n-1}$
   c) $\sum_{n=1}^{\infty} \frac{1}{n(\ln n)^2}$
   d) $\sum_{n=1}^{\infty} \frac{1}{10n^2}$
   e) $\sum_{n=1}^{\infty} \frac{n}{n^2+1}$

   f) $\sum_{n=1}^{\infty} \frac{1}{2n+3}$
   g) $\sum_{n=1}^{\infty} \frac{\cos^2 n}{2^n}$
   h) $\sum_{n=1}^{\infty} \frac{\sqrt{n}}{n^3+1}$
   i) $\sum_{n=1}^{\infty} \frac{n^2}{2^n}$
   j) $\sum_{n=1}^{\infty} ne^{-n^2}$
   k) $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n(n+1)(n+2)}}$
   l) $\sum_{n=1}^{\infty} \frac{n!}{2^{2n}}$
   m) $\sum_{n=1}^{\infty} \frac{|a_n|}{10^n}$, $|a_n| < 10$
   n) $\sum_{n=1}^{\infty} n^p e^{-n}$, $p \in \mathbb{R}$

2. If $a_n \geq 0$ and $b_n \geq 0$ for all $n \geq 1$, and if there is a constant c such that $a_n \leq cb_n$, prove that the convergence of $\sum b_n$ implies the convergence of $\sum a_n$.

3. Use the geometric idea of the integral test to show $\lim_{n\to\infty}[1 + \frac{1}{2} + \cdots + \frac{1}{n} - \ln n]$ converges to a constant γ, and show that $\frac{1}{2} < \gamma < 1$. γ is called *Euler's constant.*

4. If $\sum a_n$ converges, where $a_n \geq 0$, prove that $\sum \frac{a_n}{1+a_n}$ also converges.

5. (a). If $\sum a_n$ converges, where $a_n \geq 0$, and c_n have the property $0 \leq c_n \leq K$, the same K for all n, then prove that $\sum c_n a_n$ converges.

   (b). Deduce the result of Exercise 4 from Exercise 5a.

6. Use the geometric idea behind the integral test to prove that

   (a). $\ln n! = \ln 1 + \ln 2 + \ln 3 + \cdots + \ln n > \int_1^n \ln x\,dx = n \ln n - n + 1$ when $n \geq 2$. From this deduce that

   (b). $n! > e(\frac{n}{e})^n$, when $n \geq 2$.

   (c). As an application of (b), prove that $\lim_{n\to\infty} \frac{x^n}{n!} = 0$.

7. a) Use the idea in the proof of the divergence of the harmonic series, $\sum \frac{1}{n}$, to prove the following test for convergence: Let $\{ a_n \}$ be a positive monotonically decreasing sequence. Then $\sum a_n$ converges or diverges respectively if and only if the "condensed" series $\sum 2^n a_{2^n}$ converges or diverges.

   b) Apply the test of part (a) to again prove that $\sum \frac{1}{n^p}$ converges if $p > 1$, and diverges if $p \leq 1$.

   c) Apply the test of part (a) to determine the values of p for which the series $\sum_{n=2}^{\infty} \frac{1}{n(\ln n)^p}$ converges and diverges.

1.3 Absolute and Conditional Convergence

The tests just given for series with positive terms can be applied to many series with complex terms a_n by utilizing the concept of absolute convergence.
Definition: The series $\sum_{k=1}^{\infty} a_k$, where the a_k may be complex numbers, *converges absolutely* if the series of positive numbers $\sum_{k=1}^{\infty} |a_k|$ converges. It is called *conditionally convergent* if $\sum_{k=1}^{\infty} a_k$ converges but $\sum_{k=1}^{\infty} |a_k|$ diverges.

Absolute convergence is stronger than ordinary convergence because

Theorem 1.11 . *If $\sum_{n=1}^{\infty} |a_n|$ converges, then $\sum_{n=1}^{\infty} a_n$ converges.*

Proof: Let $a_N = \alpha_n + i\beta_n$. We shall show that the real series $\sum \alpha_n$ and $\sum \beta_n$ both converge. Then by Theorem 5 $\sum a_n$ converges too. To show that $\sum \alpha_n$ converges, let $c_n = \alpha_n + |a_n|$. Since $|\alpha_n| \le \sqrt{(\alpha_n^2 + \beta_n^2)} = |a_n|$, we know that $0 \le c_n \le 2|a_n|$. Thus the positive series $\sum c_n$ is bounded, $\sum c_n \le 2\sum |a_n| < \inf ty$, and so converges by the comparison test (Theorem 8). Since $\sum \alpha_n = \sum (c_n - |a_n|)$, and both $\sum c_n$ and $\sum |a_n|$ converge, then $\sum \alpha_n$ also converges by Theorem 4. Similarly, by taking $d_n = \beta_n + |a_n|$, the series $\sum d_n$ converges, from which we can conclude that $\sum \beta_n$ converges.

Examples:

1. The complex series $\frac{1}{1^2} + \frac{i}{2^2} + \frac{i^2}{3^2} + \frac{i^3}{4^2} + \cdots = \sum_{n=1}^{\infty} \frac{i^n}{n^2}$ converges absolutely since $\left|\frac{i^{n-1}}{n^2}\right| = \frac{1}{n^2}$ and the positive series $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges.

2. $1 + \frac{1}{2^2} - \frac{1}{2^3} - \frac{1}{2^4} + \frac{1}{2^5} + \frac{1}{2^6} - \frac{1}{2^7} - \frac{1}{2^8} + \cdots$, which is the geometric series $\sum \frac{1}{2^n}$ with negative signs thrown in, converges absolutely since $\sum \frac{1}{2^n}$ converges.

3. $\sum r^n$, r complex, converges absolutely if $\sum |r|^n$ converges, that is, if $|r| < 1$.

4. $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} \ldots = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}$, the alternating harmonic series does not converge absolutely because $\sum \frac{1}{n}$ diverges. It does converge though, as we shall see shortly. Thus the alternating harmonic series is conditionally convergent.

On the basis of this last theorem, many complex series can be proved to converge by proving they converge absolutely. Since absolute convergence concerns itself with series having only positive terms, all the tests for convergence developed in the previous section may be used. This is the most common way of proving a complex series converges. If it does not converge absolutely, the proof of convergence will usually be more difficult and use special ingenuity based on the particular series at hand.

There is one case of conditional convergence which is easy to treat, that of alternating series.

Definition: A series of *real numbers* is called *alternating* if the positive and negative terms occur alternately. They have the form

$$\sum_{n=1}^{\infty} (-1)^{n-1} a_n = a_1 - a_2 + a_3 - a_4 + \cdots,$$

where the a_n's are all positive.

Theorem 1.12 . *The alternating series $\sum_{n=1}^{\infty} (-1)^{n-1} a_n$, $a_n > 0$, converges if i) the a_n are monotone decreasing ($a_n \searrow$), and ii) $\lim_{n\to\infty} a_n = 0$. If S is the sum of the series, the inequality*

$$0 < |S - S_N| < a_{N+1} \tag{1-2}$$

shows how much the N^{th} partial sum differs from the limit S. In words inequality (2) says that the error which results by using the first N terms is less than the first neglected term a_{N+1}.

Proof: The idea is quite simple. Observe that since $a_n \searrow$, $S_{2n} - S_{2n-2} = a_{2n-1} - a_{2n} > 0$, so the S_{2n}'s increase. Similarly the S_{2n+1}'s decrease. Also both sequences are bounded—from below by S_2 and from above by S_1 (you should check this). Therefore by Theorem 8 Chapter 0, the bounded monotonic sequences S_{2n} and S_{2n+1} converge to real numbers S and $\hat{S}$ respectively. Let us show that $S = \hat{S}$.

$$\hat{S} - S = \lim_{n\to\infty} S_{2n+1} - \lim_{n\to\infty} S_{2n} = \lim_{n\to\infty} (S_{2n+1} - S_{2n}) = \lim_{n\to\infty} a_{2n+1} = 0$$

Thus the alternating series converges to the unique limit S. All that is left to verify is inequality (2). Because S_{2n} is increasing and S_{2n+1} is decreasing, we know that

$$S_{2n} < S \text{ and } S < S_{2n+1}$$

Therefore

$$0 < S - S_{2n} < S_{2n+1} - S_{2n} = a_{2n+1} \qquad \text{and} \qquad 0 < S_{2n-1} - S < S_{2n-1} - S_{2n} = a_{2n}.$$

These two inequalities are the cases N even and N odd in (2).

Examples:

1. $\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}$ converges since it is an alternating sequence and $\frac{1}{n}$ decreases monotonically to zero. Later we shall show that its sum is $\ln 2$.

2. $\sum_{n=2}^{\infty} \frac{(-1)^n}{\ln n}$ converges since $\frac{1}{\ln n}$ decreases monotonically to zero.

3. $\sum_{n=1}^{\infty} (-1)^{n-1} \frac{n}{n+1}$ diverges by Theorem 6 since $\lim_{n\to\infty} \frac{(-1)^{n-1} n}{n+1}$ is not zero.

Exercises

1. Determine which of the following series converge absolutely, converge conditionally, or diverge.
   a) $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{\sqrt{n}}$
   b) $\sum_{n=1}^{\infty} \frac{(2-3i)^n}{n!}$
   c) $\sum_{k=2}^{\infty} \frac{(2k+i)^2}{e^k}$
   d) $\sum_{n=1}^{\infty} (-1)^{n-1} \frac{\ln n}{n}$
   e) $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{2^2} + \frac{1}{5} - \frac{1}{2^3} + \frac{1}{7} - \frac{1}{2^4} + \frac{1}{9} - \cdots$.
   f) $\sum_{n=1}^{\infty} \frac{1}{n^2+2i}$
   g) $\sum_{n=1}^{\infty} \frac{1}{n+2i}$
   h) $\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n+2i}$
   i) $\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^p}$, $p > 0$,
   j) $\sum_{n=1}^{\infty} (-1)^n \frac{(1+i)n^2}{2n^2+1}$
   k) $\sum_{n=1}^{\infty} \frac{\cos n\theta}{n^2}$, θ arbitrary.

2. If $\sum a_n$ and $\sum b_n$ are absolutely convergent, and α and β are any complex numbers, prove that $\sum(\alpha a_n + \beta b_n)$ also converges absolutely.

3. Show that $\sum_{n=1}^{\infty} nz^n$ converges absolutely if $|z| < 1$.

4. Show that for any $\theta \in \mathbb{R}$, then $\sum_{n=0}^{\infty} \cos n\theta$ diverges, and that if $\theta \neq 0, \pm\pi, \pm 2\pi, \ldots$, then $\sum_{n=0}^{\infty} \sin n\theta$ also diverges.

1.4 Power Series, Infinite Series of Functions

As you will all agree, the simplest functions are polynomials. With infinite series at hand, it is reasonable to consider an "infinite polynomial"

$$a_0 + a_1 z + a_2 z^2 + a_3 a^3 + \cdots = \sum_{n=0}^{\infty} a_n z^n.$$

Because of the appearance of the powers of z, this is called a *power series*. The question of convergence of a power series is trivial at $z = 0$, for then we have only the one term a_0. Does this series converge for any other values of z , and if so, for which ones?

The answer depends on the coefficients a_n, but in any case, the set of complex numbers, $z \in \mathbb{C}$, for which the series converges is always a disc $|z| < \rho$—with possibly some additional

points on the boundary $|z| = \rho$—in the complex pane bC. This number ρ is called the *radius of convergence* of the power series. We shall first prove that the set $z \in \mathbb{C}$ for which a power series converges is always a disc. Then we shall give a way of computing the radius ρ of that disc.

Theorem 1.13 . *The set $z \in \mathbb{C}$ for which the power series $\sum a_n z^n$ converges is always a disc $|z| < \rho$, inside of which it even converges absolutely. We do not exclude the two extreme possibilities that the radius of this disc is zero or infinity.*

The series might converge at some, none, or all of the points on the boundary of the disk $|z| = \rho$.

PROOF: We shall show that if the series converges for any $\zeta \in \mathbb{C}$, then it converges absolutely for all complex z with $|z| < |\zeta|$. If $\zeta = 0$, there is nothing to prove, so assume $\zeta \neq 0$. Because $\sum a_n \zeta^n$ converges, $\lim_{n\to\infty} |a_n \zeta^n| \to 0$. Thus all the terms are bounded in absolute value, that is, there is an M such that $|a_n \zeta^n| < M$ for all n. Then, since

$$|a_n z^n| = \left| a_n \zeta^n \frac{z^n}{\zeta^n} \right| < M \left| \frac{z}{\zeta} \right|^n \quad \text{for all } n,$$

the series $\sum |a_n z^n|$ is dominated by $M \sum \left| \frac{z}{\zeta} \right|^n$. But this last series is a geometric series which does converge since $|z| < |\zeta|$, so $\left| \frac{z}{\zeta} \right| < 1$. Thus by the comparison test $\sum a_n z^n$ converges absolutely for all $z \in \mathbb{C}$ with $|z| < |\zeta|$.

Therefore, if the power series $\sum a_n z^n$ converges for some complex number ζ, then it converges in the whole disc $|z| < |\zeta|$. The radius of convergence ρ is then the radius of the largest disc $|z| < \rho$ for which the series converges.

See Exercise 3 for examples concerning convergence on the boundary of the disk.

Let us now give a method of computing ρ which covers most cases arising in practices.

Theorem 1.14 . *If $\lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| = L$ exists, the power series $\sum a_n z^n$ has radius of convergence $\rho = \frac{1}{L}$ if $L \neq 0$, ∞ if $L = 0$. In other words, if $L \neq 0$ the series converges in the disc $|z| < \frac{1}{L}$ and diverges if $|z| > \frac{1}{L}$. On the circumference $|z| = 1/L$, anything may happen (see Exercise 3 at the end of this section). If $L = 0$, the series converges in the whole complex plane.*

PROOF: This is a simple application of the ratio test. The series converges if the limit of the ratio of successive terms $\lim_{n\to\infty} \left| \frac{a_{n+1} z^{n+1}}{a_n z^n} \right|$ is less than one and diverges if it is greater than one. Thus we have convergence if

$$\lim_{n\to\infty} \left| \frac{a_{n+1} z}{a_n} \right| = |z| L < 1, \text{ i.e. if } |z| < \frac{1}{L},$$

and divergence if

$$\lim_{n\to\infty} \left| \frac{a_{n+1} z}{a_n} \right| = |z| L > 1, \text{ i.e. if } |z| > \frac{1}{L}.$$

REMARK: In the one additional case $\left|\frac{a_{n+1}}{a_n}\right| \to \infty$ as $n \to \infty$, the series diverges for every $|z| \neq 0$, as can easily be seen again by the ratio test.

EXAMPLES:

1. $\sum_{n=0}^{\infty} z^n$ converges where $\lim_{n\to\infty} \left|z^{n+1}/z^n\right| < 1$ that is; for $|z| < 1$.

2. $\sum_{n=0}^{\infty} \frac{nz^n}{2^n}$ converges where $\lim_{n\to\infty} \left|\frac{(n+1)z^{n+1}}{2^{n+1}} / \frac{nz^n}{2^n}\right| < 1$. Since
$$\lim_{n\to\infty} \left|\frac{(n+1)z^{n+1}}{2^{n+1}} / \frac{nz^n}{2^n}\right| = \lim_{n\to\infty} \left|\frac{(n+1)z}{2n}\right| = \left|\frac{z}{2}\right|,$$
the series converges for all $|z| < 2$.

3. $\sum_{n=0}^{\infty} \frac{z^n}{n!}$ converges where $\lim_{n\to\infty} \left|\frac{z^{n+1}}{(n+1)!} / \frac{z^n}{n!}\right| < 1$. Since
$$\lim_{n\to\infty} \left|\frac{z^{n+1}}{(n+1)!} / \frac{z^n}{n!}\right| = \lim_{n\to\infty} \left|\frac{z}{n+1}\right| = 0,$$
the series converges for all $z \in \mathbb{C}$, that is, in the whole complex plane.

4. $\sum_{n=0}^{\infty} n!z^n$ converges where $\lim_{n\to\infty} \left|\frac{(n+1)!z^{n+1}}{n!z^n}\right| < 1$. But
$$\lim_{n\to\infty} \left|\frac{(n+1)!z^{n+1}}{n!z^n}\right| = \lim_{n\to\infty} |(n+1)z| = \infty$$
unless $z = 0$. Thus the ratio is less than one only at $z = 0$, so the series converges only at the origin.

Only minor modifications are needed for the more general power series
$$a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \cdots = \sum_{n=0}^{\infty} a_n(z - z_0)^n,$$
where $a_0 \in \mathbb{C}$. Again the series converges in a disc in the complex plane, only now the disc has its center at z_0 instead of the origin, so if the radius of convergence is ρ, the series converges for $|z - z_0| < \rho$. An example should make this clear.

EXAMPLE: $\sum_{n=1}^{\infty} \frac{(z-2i)^n}{n}$. By the ratio test, this converges when
$$\lim_{n\to\infty} \left|\frac{(z-2i)^{n+1}}{n+1} / \frac{(z-2i)^n}{n}\right| < 1,$$

that is, when $|z - 2i| < 1$. This is a disc with center at $2i$ and radius 1.

A few words should be said about *real power series* $\sum a_n(x - x_0)^n$ where both x and x_0 are real (some people only use this phrase if the a_n are also real). This is a special case of $\sum a_n(z - z_0)^n$ where z_0 is on the real axis and we only ask for what *real* z the series converges. However we know that $\sum a_n(z - z_0)^n$ converges only for those z in the disc of convergence $|z - z_0| < \rho$—and possibly some boundary points. Thus the *real* values of z for which the series $\sum a_n(z - z_0)^n$ converges are exactly those points on the real axis which are also inside the disc of convergence of the complex power series. In particular the series $\sum a_n(x - x_0)^n$, with both x and x_0 real converges for $|x - x_0| < \rho$, i.e., in the interval $x_0 - \rho \leq x \leq x_0 + \rho$.

EXAMPLE: For what $x \in \mathbb{R}$ does $\sum_{n=0}^{\infty} \frac{1}{2^n}(x-1)^n$ converge? The related complex series $\sum_{n=0}^{\infty} \frac{1}{2^n}(z-1)^n$ converges in the disc $|z - 1| < 2$. The points on the real axis which are in this disc are $|x - 1| < 2$, which is $-1 < x < 3$. A direct check shows the series diverges at both end points $x = -1$ and $x = 3$. If $\sum a_n$ and $\sum b_n$ both converge, can we define their product in a meaningful way
$$(\sum_{n=0}^{\infty} a_n)(\sum_{n=0}^{\infty} b_n) = \sum_{n=0}^{\infty} c_n?$$
and if so, does the resulting series converge? The most simple-minded approach is to insert powers of z (a bookkeeping device), giving $(\sum a_n z^n)(\sum b_n z^n)$, try long multiplication and see what happens. A computation shows that
$$(a_0 + a_1 z + a_x z^2 + \cdots)(b_0 + b_1 z + b_2 z^2 + \cdots) = a_0 b_0 + (a_0 b_1 + a_1 b_0)z$$
$$+(a_0 b_2 + a_1 b_1 + a_2 b_0)z^2 + \cdots + (a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0)z^n + \cdots .$$

Motivated by this, we make the following

DEFINITION: The formal *product*, called the *Cauchy product*, of the series $\sum a_n$ and $\sum b_n$ is defined to be
$$(\sum_{n=0}^{\infty} a_n)(\sum_{n=0}^{\infty} b_n) \equiv \sum_{n=0}^{\infty} c_n,$$
where
$$c_n = a_0 b_n + a_z b_{n-1} + \cdots + a_n b_0 = \sum_{k=0}^{n} a_k b_{n-k}.$$

With this definition we shall answer the question we raised about multiplication of power series.

Theorem 1.15 . *If $\sum_{n=0}^{\infty} a_n = A$ and $\sum_{n=0}^{\infty} b_n = B$ both converge absolutely, then the Cauchy product series*
$$(\sum_{n=0}^{\infty} a_n)(\sum_{n=0}^{\infty} b_n) \equiv (\sum_{n=0}^{\infty} c_n),$$

where

$$c_n = \sum_{k=0}^{\infty} a_k b_{n-k},$$

also converges absolutely, and to $C = AB$.

PROOF: Let $A_N = \sum_{n=0}^{N} a_n$, $B_N = \sum_{n=0}^{N} b_n$, and $C_N = \sum_{n=0}^{N} c_n$. We shall show that by picking N large enough, $|A_N B_N - C_N|$ can be made arbitrarily small. Since $A_N B_N \to AB$, this will complete the proof. Observe that

$$C_N = a_0 b_0 + (a_0 b_1 + a_1 b_0) + \cdots + (a_0 b_N + \cdots + a_N b_0) = \sum\sum a_j b_k,$$

while

$$A_N B_N = (a_0 + \cdots + a_N)(b_0 + \cdots + b_N) = \sum_{j=0}^{N}\sum_{k=0}^{N} a_j b_k.$$

Therefore

$$|A_N B_N - C_N| = \left|\sum_{j=0}^{N}\sum_{k=0}^{N} a_j b_k\right| \le \sum_{j=0}^{N}\sum_{k=0}^{N} |a_j|\ |b_k|\,.$$

Since $j + k > N$, either $j > N/2$ or $k > N/2$, so

$$|A_N B_N - C_N| \le \sum_{j>\frac{N}{2}}^{N}\sum_{k=0}^{N} |a_j|\ |b_k| + \sum_{j=0}^{N}\sum_{k>\frac{N}{2}}^{N} |a_j|\ |b_k|\,.$$

Because the original series both converge absolutely, they are bounded,

$$\sum_{j=0}^{\infty} |a_j| < M \text{ and } \sum_{k=0}^{\infty} |b_k| < M.$$

Consequently,

$$|A_N B_N - C_N| \le M(\sum_{j>\frac{N}{2}}^{\infty} |a_j| + \sum_{k>\frac{N}{2}}^{\infty} |b_k|).$$

Again using the absolute convergence of the original series, we see that for N large, the right side can be made arbitrarily small.

Since we shall need the ideas later on, let us digress briefly and examine the convergence of infinite series of functions, $\sum u_n(z)$. In the special case where $u_n(z) = a_n(z - z_0)^n$, this is a power series. Generally, there is little one can say about the convergence of such series except to apply our general tests and hope for the best. We shall only illustrate the situation with two

EXAMPLES:

1. $\sum_{n=1}^{\infty} \frac{\cos n\theta}{n^2}$, where θ is any real number. This converges for all θ since it converges absolutely, that is $\sum + \left|\frac{\cos n\theta}{n^2}\right|$ converges. We can see this last statement is true by comparing $\sum + \left|\frac{\cos n\theta}{n^2}\right|$ with the larger convergent series (since $|\cos n\theta| \le 1$) $\sum_{n=1}^{\infty} \frac{1}{n^2}$.

2. $\sum_{n=1}^{\infty} ne^{nx}$. By the ratio test, converges if $\lim_{n\to\infty} \left|(n+1)e^{(n+1)x}/ne^{nx}\right| < 1$. Since
$$\lim_{n\to\infty} \left|(u+1)e^{(n+1)x}/ne^{nx}\right| = \lim_{n\to\infty} \left|\frac{n+1}{n}\right| e^x = e^x,$$
the series converges if $e^x < 1$, which happens only when $x < 0$.

Exercises

1. Find the disc of convergence of the following power series by finding the center and radius of the disc.
   a) $\sum_{n=0}^{\infty} \frac{z^n}{n+1}$
   b) $\sum_{n=0}^{\infty} \frac{(z-2)^n}{n}$
   c) $\sum_{n=0}^{\infty} \frac{in}{2n-1} z^{n-1}$
   d) $\sum_{n=0}^{\infty} (n+1)[z - 2 + 3i]^n$
   e) $\sum_{n=0}^{\infty} \frac{(2z+3)^n}{n^2+2i}$
   f) $\sum_{n=0}^{\infty} \frac{1}{\ln n} z^{n-2}$
   g) $\sum_{n=0}^{\infty} \frac{(2n-i)}{3^n} z^n$
   h) $\sum_{n=0}^{\infty} \frac{2^n z^n}{n!} (0! \equiv 1)$
   i) $\sum_{n=0}^{\infty} (\frac{1}{2^n} + \frac{i}{3^n} z^n$
   j) $\sum_{n=0}^{\infty} \frac{(z+i)^n}{2^{2n}}$
   k) $\sum_{n=0}^{\infty} \frac{z^2 n}{(2n)!}$
   l) $\sum_{n=0}^{\infty} n^n (z-1)^n$
   m) $\sum_{n=0}^{\infty} \frac{z^{2n}}{4^n}$
   n) $\sum_{n=0}^{\infty} (\frac{1}{n} + \frac{i}{n^2+1})(z - \sqrt{2}i)^n$

2. Find the set $x \in \mathbb{R}$ for which the following series converge.
   a) $\sum_{n=0}^{\infty} \frac{(x-1)^n}{n2^n}$
   b) $\sum_{n=0}^{\infty} \frac{\cos nx}{2^n}$
   c) $\sum_{n=0}^{\infty} \frac{1}{n} (\frac{x-1}{x})^n$

d) $\sum_{n=0}^{\infty} e^{-n(x+1)}$

e) $\sum_{n=0}^{\infty} \frac{2^n (\sin x)^n}{n}$

f) $\sum_{n=0}^{\infty} (1+e^x)^n$

g) $\sum_{n=0}^{\infty} (1-e^x)^n$

3. The point of this exercise is to show that a power series might converge at some, none, or all of the points on the boundary of the disk of convergence.
   a) Show that $\sum_{n=0}^{\infty} z^n$ diverges at every point on the boundary of its disc of convergence.
   b) Show that $\sum_{n=0}^{\infty} \frac{z^n}{n+1}$ diverges for $z = 1$ but converges for $z = -1$ (in fact, it converges everywhere on $|z| = 1$ except at $z = 1$).
   c) Show that $\sum_{n=0}^{\infty} \frac{x^n}{(n+1)^2}$ converges at every point on the boundary of its disc of convergence.

4. If $\sum a_n z^n$ diverges for $z = \zeta \in \mathbb{C}$, prove that it diverges for all $z \in \mathbb{C}$ with $|z| > |\zeta|$.

5. For what $z \in \mathbb{C}$ does $\sum_{n=0}^{\infty} \frac{z^2}{(1+z^2)^n}$ converge? Find a formula for the nth partial sum $S_n(z)$. Evaluate $\lim_{n\to\infty} S_n(z)$. Is the limit function continuous?

6. Let $\sum_{n=0}^{\infty} P(n) a_n z^n$ have radius of convergence rho, and let $P(n)$ be any polynomial. Prove that $\sum_{n=0}^{\infty} P(n) a_n z^n$ converges and also has ρ as its radius of convergence. (By $P(n)$ w e mean $P(n) = A_k n^k + A_{k-1} n^{k-1} + \cdots + A_1 n + A_0$).

1.5 Properties of Functions Represented by Power Series

Having found that a power series $\sum a_n (z-z_0)^n$ converges in some disc, $|z-z_0| < \rho$, it is interesting to study the function $f(z)$ defined by the power series for z in the disc of convergence

$$f(z) = \sum_{n=0}^{\infty} a_n (z-z_0)^n, \qquad |z-z_0| < \rho.$$

It turns out that functions $f(z)$ defined by a convergent power series are delightful, as nicely behaved as functions can be. In particular, they are not only continuous, but also automatically have an infinite number of continuous derivatives and many other amazing properties.

This section will be devoted to proving the more elementary properties of functions represented by power series, while in the next section we will begin with given functions, like $\sin x$, and see if there is a convergent power series associated with the m, as well as showing a way of obtaining the coefficients a_n of that power series. The profound theory of

functions represented by convergent power series is called *analytic functions of a complex variable.*

DEFINITION: A function $f(z)$ of the complex variable z is said to be *analytic* in the disc $|z-z_0| < \rho$ if $f(z)$ can be represented by a convergent power series in that disc:

$$f(z) = \sum_{n=0}^{\infty} a_n (z-z_0)^n, \qquad |z-z_0| < \rho.$$

Since we have not developed the notion of the derivative, $\frac{df}{dz}$, of a complex valued function $f(z)$ of the complex variable z, nor have we considered the corresponding theory of integration, $\int f(z)dz$, the scope of our treatment will regrettably have to be narrowed. However our proofs will have the property that as soon as an adequate theory of differentiation and integration is given, the theorems and proofs remain unchanged.

Instead of considering power series in the complex variable z, we shall restrict our attention to series in the real variable x

$$f(x) = \sum_{n=0}^{\infty} a_n (x-x_0)^n, \ |x-x_0| < \rho, \tag{1-3}$$

still allowing the coefficients a_n to be complex. Thus, $f(x)$ is a complex-valued function of the real variable x. The definitions of derivative and integral for such functions were given in Section 7 of Chapter 0. We shall use that material here . Our aim is the following:

Theorem 1.16 . *Suppose that $\sum_{n=0}^{\infty} a_n x^n$ has radius of convergence $\rho > 0$ (possibly ∞). Then*

(a) *the function $f(x)$ defined by*

$$f(x) = \sum_{n=0}^{\infty} a_n x^n, \ |x| < \rho,$$

has an infinite number of derivatives;

(b) *the series $\sum_{n=0}^{\infty} n a_n x^{n-1}$ has the same radius of convergence ρ and*

$$f'(x) = \sum_{n=0}^{\infty} n a_n x^{n-1}, \ |x| < \rho,$$

and

(c) *the series $\sum_{n=0}^{\infty} \frac{a_n}{n+1} x^{n+1}$ has the same radius of convergence ρ, and*

$$\int_0^x f(t)\, dt = \sum_{n=0}^{\infty} \frac{a_n}{n+1} x^{n+1}, \ |x| < \rho.$$

REMARK: If we omit $f(x)$ from the picture and write (b) and (c) directly in terms of the infinite sum, we find

$$\text{(b)}'\ \frac{d}{dx}[\sum_{n=0}^{\infty} a_n x^n] = \sum_{n=0}^{\infty} n a_n x^{n-1}$$

and

$$\text{(c)}'\ \int_0^x [\sum_{n=0}^{\infty} a_n t^n]\,dt = \sum_{n=0}^{\infty} \frac{a_n}{n+1} x^{n-1}.$$

These two statements are usually abbreviated "a power series may be differentiated term by term" and "a power series may be integrated term by term" within their domain of convergence (these statements are *not* generally true for an arbitrary infinite series of functions $\sum u_n(x)$, see Exercise 4 below). The generalization to $\sum_{n=0}^{\infty} a_n(x-x_0)^n$ is obvious.

Our proof will be given in several parts. We begin with the *Lemma 1.* Under the hypothesis of the theorem, $f(x)$ is continuous for all $\tilde{x}$ with $|\tilde{x}| < \rho$. PROOF: (This is a little dull). Given any $\epsilon > 0$, we must find a $\delta > 0$ such that

$$|f(x) - f(\tilde{x})| < \epsilon \text{ when } |x - \tilde{x}| < \delta.$$

Let us write $f_N(x) = \sum_{n=0}^{N} a_n x^n$ and $R_N(x) = \sum_{N+1}^{\infty} a_n x^n$, so that $f(x) = f_N(x) + R_N(x)$.

Observe that $|f(x) - f(\tilde{x})| = |f_N(x) - f_N(\tilde{x}) + R_N(x) - R_N(\tilde{x})| \le |f_N(x) - f_N(\tilde{x})| + |R_N(x)| + |R_N(\tilde{x})|$.

We shall show that each of these three terms can be made $< \frac{\epsilon}{3}$ by picking x close enough to $\tilde{x}$ and N -which is entirely at our disposal- large enough.

First work with $R_N(x)$ and $R_N(\tilde{x})$. Choose r such that $|\tilde{x}| < r < \rho$. This is to insure that we stay away from the boundary $|x| = \rho$ where the series may diverge. Then $\sum_{n=0}^{\infty} |a_n r^n|$ converges absolutely, say to the number S. If we let $S_N = \sum_0^N |a_n r^n|$, we know that by picking N large enough, $\sum_{N+1}^{N} |a_n r^n| = S - S_N < \frac{\epsilon}{3}$. But $|R_N(x)| = \left|\sum_{N+1}^{\infty} a_n x^n \le \sum_{N+1}^{\infty} |a_n x^n|\right|$, so that *if* $|x| + \le r$, by using the same N found above, we have

$$|R_N(x)| \le \sum_{N+1}^{\infty} |a_n r^n| = S - S_N < \frac{\epsilon}{3}.$$

Since by the definition of r we know $|\tilde{x}| \le r$, this also proves that for this same N $|R_N(\tilde{x})| < \frac{\epsilon}{3}$. Thus by restricting $|x| \le r$, we have seen that both $|R_N(x)|$ and $|R_N(\tilde{x})|$ can be made less than $\frac{\epsilon}{3}$.

Having fixed N, $f_N(x)$ is a polynomial -which we know is continuous. Thus there is a $\delta, > 0$ such that

$$|f_N(x) - f_N(\tilde{x})| < \frac{\epsilon}{3} \text{ when } |x - \tilde{x}| < \delta_1.$$

This shows that $|f(x) - f(\tilde{x})| < \epsilon$ if x is in the intersection of the intervals $|x| \le |\tilde{x}| < r < \rho)$ and $|x - \tilde{x}| < \delta_1$. That there is some interval contained in both of these intervals is easy to see since both contain all points sufficiently close to $\tilde{x}$. And the proof is completed. As you have observed, the proof involves no new ideas but is rather technical.

With this lemma proved, we know that $f(x)$ is continuous -and hence integrable. Thus we can work with $\int_0^x f(t)\,dt$. Our next task is to prove a portion of Part (c) of Theorem 16.

Lemma 1.17 *If $\sum_{n=0}^{\infty} a_n x^n$ has radius of convergence $\rho > 0$, then*

$$\sum_{n=0}^{\infty} \frac{a_n}{n+1} x^{n+1} = \int_0^x f(t)\,dt \text{ for all } |x| < \rho.$$

PROOF: We shall show that

$$\left| \int_0^x f(t)\,dt - \sum_{n=0}^{N} \frac{a_n}{n+1} x^{n+1} \right| \tag{1-4}$$

can be made arbitrarily small by choosing N large enough. Write

$$f(t) = \sum_{n=0}^{N} a_n t^n + \sum_{n=N+1}^{\infty} a_n t^n.$$

Then since we can integrate any *finite* sum term by term, we have

$$\int_0^x f(t)\,dt = \sum_{n=0}^{N} a_n \int_0^x t^n\,dt + \int_0^x [\sum_{n=N+1}^{\infty} a_n t^n]\,dt = \sum_{n=1}^{N} \frac{a_n}{n+1} x^{n+1} = \int_0^x [\sum_{n=N+1}^{\infty} a_n t^n]\,dt,$$

so that (4) reduces to showing that

$$\left| \int_0^x \sum_{n=N+1}^{\infty} a_n t^n\,dt \right|$$

can be made small by choosing N large. The idea here is to apply Theorem 16 of Chapter 0. This means we need to estimate the size of the above integrand. By now you should recognize the method. Because $|x| < \rho$, we can choose an r such that $|x| < r < \rho$. Then $\sum a_n r^n$ is convergent so its terms are bounded, say $M \ge |a_n r^n|$ for all n, that is, $|a_n| \le \frac{M}{r^n}$. Therefore, since $|t| < |x|$, we find the inequality

$$\left| \sum_{N+1}^{\infty} a_n t^n \right| \le \sum_{N+1}^{\infty} |a_n|\,|t|^n \le \sum_{N+1}^{\infty} \frac{M}{r^n} |x|^n.$$

But the last series is a geometric series whose sum is $\left|\frac{x}{r}\right|^N \frac{M|x|}{r-x}$. Thus

$$\left| \sum_{N+1}^{\infty} a_n t^n \right| \le \left|\frac{x}{r}\right|^N \frac{M\,|x|}{r-x}.$$

Applying Theorem 16 of Chapter 0, we find that

$$\left|\int_0^x (\sum_{N+1}^{\infty} a_n t^n)\, dt\right| \le \left|\frac{x}{r}\right|^N \frac{M\,|x|^2}{r-x}.$$

that is,

$$\left|\int_0^x f(t)\, dt - \sum_0^N \frac{a_n}{n+1} x^{n+1}\right| \le \left|\frac{x}{4}\right|^N \frac{M\,|x|^2}{r-x}.$$

Since $\left|\frac{x}{r}\right| < 1$, we know that $\left|\frac{x}{r}\right|^N \to 0$ as $N \to \infty$, which completes the proof of the lemma.

Incidentally, all we have left to prove of part c of the theorem is that the radius of convergence of the integrated series is no larger than ρ (since the lemma shows it is at least ρ). But this will have to wait until after

Lemma 1.18 *If $\sum a_n x^n$ has radius of convergence ρ, the series obtained by formally differentiating term by term, $\sum n a_n x^{n-1}$, has the same radius of convergence.*

REMARK: This lemma does *not* say that the derived series is equal to the derivative of the function defined by the original series. It only discusses the radius of convergence, not the relationship of the functions represented b y the two series.
PROOF: Let ρ_1 be the radius of convergence of $\sum n a_n x^{n-1}$. First we show that $\rho_1 \le \rho$. If $\sum n a_n x^{n-1}$ converges for some fixed x, then so does $\sum n a_n x^n$. But the terms of this last sequence are larger than those of $\sum a_n x^n$ since $|n a_n x^n| \ge |a_n x^n|$. Thus by the comparison test $\sum a_n x^n$ also converges for that x, which shows $\rho_1 \le \rho$.

To show that $\rho \le \rho_1$, assume $\sum a_n x^n$ converges for some x and choose r between $|x|$ and ρ, $|x| < r < \rho$. As in the proof of Lemma 2 we find that $|a_n| < M r^{-n}$. Then the terms in the series $\sum \left|n a_n x^{n-1}\right|$ are smaller than the corresponding terms in $\sum n \frac{M}{r} \frac{|x|}{r}^{n-1}$. By the ratio test this last series converges, since $|x| < r$. Thus the derived series $\sum n a_n x^{n-1}$ also converges, showing that $\rho \le \rho_1$ and completing the proof of the lemma.

Now we can complete the proof of part c of Theorem 16.

Corollary 1.19 *If $\sum a_n x^n$ has radius of convergence ρ, then the series obtained by formally integrating term by term, $\sum \frac{a_n}{n+1} x^{n+1}$ also has radius of convergence ρ.*

PROOF: The series $\sum a_n x^n$ is the formal derivative of the series $\sum \frac{a_n}{n+1} x^{n+1}$, and we have just seen that these two series have the same radius of convergence.

We shall next prove part (b) of Theorem 16 as

Lemma 1.20 *$f(x) \equiv \sum_{n=0}^{\infty} a_n x^n$ has radius of convergence $\rho > 0$ then*

$$\frac{df}{dx} = \frac{d}{dx}[\sum_{n=0}^{\infty} a_n x^n] = \sum_{n=0}^{\infty} n a_n x^{n-1},$$

and this series also has radius of convergence ρ.

PROOF: In Lemma 3 we proved that the radii of convergence are the same. What we must prove here is that the derivative of the function is given by the derivative of the series. This is a more or less immediate consequence of Lemma 2, for let us apply this integration lemma to the function $g(x)$ defined by

$$g(x) \equiv \sum_{n=1}^{\infty} n a_n x^{n-1},\ |x| < \rho.$$

Then we find that

$$\int_0^x g(t)\, dt = \sum_{n=1}^{\infty} a_n x^n = f(x) - a_0,\ |x| < \rho.$$

By the fundamental theorem of calculus, we can take the derivative of the left side, and it is $g(x)$. Thus

$$g(x) = f'(x),$$

that is,

$$\sum_{n=1}^{\infty} n a_n x^{n-1} = \frac{d}{dx} f(x).$$

This incidentally also proves the otherwise not obvious fact that $f(x)$, only known to be continuous so far (Lemma 1) is also differentiable.

To complete the proof of Theorem 16, we must prove Lemma 5. If the power series $\sum a_n x^n$ converges for $|x| < \rho$, then the function $f(x)$ defined by $f(x) \equiv \sum_{n=0}^{\infty} a_n x^n$ has an infinite number of derivatives. The derivatives are represented by the formal series obtained by term-by-term differentiation.
PROOF: By induction, Lemma 4 shows us that $f(x)$ has one derivative. Assume $f(x)$ has k derivatives. We shall show that is has $k+1$. Let $f^{(k)}(x) = \sum b_n x^n$ be the series for the k^{th} derivative of f. Applying Lemma 4 to this series we find that $f^{(k)}(x)$ is differentiable. This proves that f has $k+1$ derivatives and completes the induction proof.

EXAMPLES: (a) We know that

$$\frac{1}{1+t} = \sum_{n=0}^{\infty} (-t)^n = 1 - t + t^2 - t^3 + \cdots .$$

where the geometric series converges for $|t| < 1$. Applying the theorem, we integrate term by term to find that

$$\ln(1+x) = \int_0^x \frac{1}{1+t}\, dt = \sum_{n=0}^{\infty} \frac{(-1)^n \cdot x^{n+1}}{n+1},\ |x| < 1,$$

or

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} + \cdots .$$

Thus the function $\ln(1+x)$ is equal to the power series on the right. With a little more work we can prove that the series, which converges at $x = 1$, converges to $\ln(1 + 1)$ and obtain the following interesting formula.

$$\ln 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots .$$

The power series for $\ln(1 + x)$ can be used to illustrate the possibilities of computing with infinite series. If $0 < x < 1$ the series for $\ln(1 + x)$ is a strictly alternating series to which we can apply inequality (2) of Theorem 12. For this series it reads

$$0 < \left| \ln(1+x) - \sum_{n=0}^{k} \frac{(-1)^n x^{n+1}}{n+1} \right| < \frac{x^{k+2}}{k+2}, \qquad x > 0.$$

This inequality states that if only the first k terms of the infinite series are used to compute $\ln(1 + x)$, the error will be less than $\frac{x^{k+2}}{k+2}$. Say we want to compute $\ln(1 + \frac{1}{4}) = ln\frac{5}{4}$ to 5 decimal places. Then we want t o choose k so that

$$\frac{\frac{1}{4}^{k+2}}{k+2} < \frac{1}{1,000,000} = 10^{-6}$$

Cross-multiplying, writing $4 = 2^2$, we want k such that $10^6 < (k + 2)2^{2k+4}$, since $k + 2 \geq 2$, $2^{2k+5} \leq (k + 2)2^{2k+4}$. Thus, we are done if we can find k such that

$$10^6 \leq 2^{2k+5}.$$

But since $2^{10} = 1024 > 10^3$, we know $2^{20} > 10^6$. Thus if $2k + 5 \geq 20$, or $k = 8$ we will have the desired accuracy. This means that

$$\ln\frac{5}{4} = \frac{1}{4} - \frac{1}{2}(\frac{1}{4})^2 + \cdots + \frac{1}{9}(\frac{1}{4})^{8+1} + \text{ error}$$

where the error is less than 10^{-6}.

From the form of the error estimate, it is clear that the series converges faster if x is smaller. This power series, valid only if $|x| < 1$ can be used to compute $\ln(1+x)$ if $|x| > 1$ by utilizing the observation illustrated by

$$\ln 6 = 3\ln(\frac{3}{2}) + 2\ \ln(\frac{4}{3}) = 3\ \ln(1 + \frac{1}{2}) + 2\ \ln(1 + \frac{1}{3}),$$

where both $\ln(1 + \frac{1}{2})$ and $\ln(1 + \frac{1}{3})$ can be computed using the power series. We should confess that this series converges too slowly to be of much value for that purpose in real life.

(b) Since $\frac{1}{1+t^2}$ is also the sum of a geometric series

$$\frac{1}{1+t^2} = 1 - t^2 + t^4 - t^6 + t^8 + \cdots = \sum_{0}^{\infty} (-1)^n t^{2n},\ |t| < 1,$$

if we integrate term by term, we find

$$\tan^{-1} x = \int_0^2 \frac{dt}{1+t^2} + \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{2n+1} = x - \frac{x^3}{3} + \frac{x^5}{5} + \cdots ,$$

which converges if $|x| < 1$. Further investigation shows that the series also converges at $x = 1$ and represents the function at that point. This yields the wonderful formula (obtained by letting $x = 1$)

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$$

from which we can compute π to any desired accuracy.

Exercises

1. Write down an infinite series whose sum is $\frac{1}{1-t}$ and integrate the series term by term to obtain a power series for $\ln(1 - x)$. For what x does the series converge?

2. Find a power series which converges about $x = 0$ for the function $\frac{x}{(1-x)^2}$ by recognizing $\frac{1}{(1-x)^2}$ as the derivative of a function whose power series in known. For what x does the series converge?

3. Compute $\ln\frac{9}{8}$ to 4 decimal places, proving the error in your approximation is correct.

4. Show that $\sum_{n=1}^{\infty} \frac{\sin n^2 x}{n^2}$ converges for all x but the series obtained by differentiating term-by-term does not converge, say at $x = 0$.

5. Exercise your ingenuity and apply the theorems of this section to find the function whose power series is
   (a) $a + 2x^2 + 4x^4 + 6x^6 + 8x^8 + \cdots + (2n)x^{2n} + \cdots$.
   (b) $2 + 3 \cdot 2x + 4 \cdot 3x^2 + 5 \cdot 4x^3 + \cdots + (k + 2)(k + 1)x^k + \cdots$

6. Taylor's Theorem. Representation of a Given Function in a Power Series. The Binomial Theorem.

In this section we prove Taylor's Theorem, an important generalization of the mean value theorem, and use it to investigate the questions i) when does a given function $f(x)$ have a power series? and ii) if $f(x)$ has a power series about x_0, $f(x) = \sum_{n=0}^{\infty} a_n(x - x_0)^n$, how can we find the coefficients a_n? As a partial answer to i) we know from Theorem 16 of the last section that if $f(x)$ has a power series about x_0, it must necessarily have an infinite number of derivatives at x_0. It turns out that this is not enough.

Perhaps it is easiest to begin with question ii).

Assume $f(x)$ has a power series about x_0,

$$f(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n,$$

which converges for $|x - x_0| < \rho$. How can we find the coefficients a_n? By Theorem 16 we know that f has an infinite number of derivatives at x_0. Moreover these derivatives can be calculated by differentiating the power series term-by-term. F or convenience we let $x_0 = 0$.

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n + \cdots,$$

$$f'(x) = a_1 + 2a_2 x + 3a_3 x^2 + \cdots + n a_n x^{n-1} + \cdots,$$

$$f''x(x) = 2a_2 + 2 \cdot 3a_3 x + 3 \cdot 4 \cdot a_4 x^2 + \cdots + n(n-1) a_n x^{n-2} + \cdots,$$

$$f^{(3)}(x) = 2 \cdot 3a_3 + 2 \cdot 3 \cdot 4a_4 x + 3 \cdot 4 \cdot 5a_5 x^2 + \cdots +,$$

$$f^{(n)}(x) = n! a_n + (n+1)! a_{n+1} x + \frac{(n+2)}{x}! a_{n+2} x^2 + \cdots.$$

By letting $x = 0$ in each line, we find

$$a_0 = f(0),\ a_1 = f'(0),\ a_2 = \frac{f''(0)}{2}, \ldots, a_n = \frac{f^{(n)}(0)}{n!}.$$

This proves

Theorem 1.21 *If $f(x) = \sum a_n (x - x_0)^n$ has a convergent power series representation about x_0, then the coefficients a_n are equal to $f^{(n)}(x_0)/n!$, so in fact*

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!} (x - x_0)^n. \tag{1-5}$$

This formula (1.21) completely solves the problem of finding the coefficients a_n of a function if that function has a power series. A simple consequence is the

Corollary 1.22 *A function $f(x)$ has at most one convergent Taylor series about a point x_0.*

PROOF: By the above theorem, if $f(x) = \sum a_n (x - x_0)^n$ and $f(x) = \sum b_n (x - x_0)^n$, then $a_n = \frac{f^{(n)}(x_0)}{n!} = b_n$, so the power series are identical.
REMARK: When f has a power series expansion about x_0, the series is usually called the *Taylor series* of f at x_0. In the special case $x_0 = 0$, the series is sometimes called the *Maclaurin series* for f.

EXAMPLES:

1. If $f(x) = e^x$ has a power series about $x = 0$, what is it? Since $f^{(n)}(0) = \frac{d^n}{dx^n} e^x \big|_{x=0} = e^0 = 1$, we know that $a_n - 1/n!$ so that the power series is $\sum_{n=0}^{\infty} \frac{1}{n!} x^n$. We cannot yet write $e^x = \sum_{n=0}^{\infty} \frac{1}{n!} x^n$ since we have not proved that e^x does have a power series.

2. If $f(x) = \cos x$ has a power series about $x = 0$, what is it? $f(0) = 1$, $f'(0) = -\sin 0 = 0$, $f''(0) = -\cos 0 = -1$, $f'''(0) = \sin 0 = 0$, $f^{(4)}(0) = \cos 0 = 1, \ldots$. All the odd derivatives at 0 are zero while the even derivatives alternate between $+1$ and -1. Therefore the series is

$$a - \frac{1}{2!} x^2 + \frac{1}{4!} x^4 - \frac{1}{6!} x^6 + \cdots \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!}.$$

Again we cannot yet claim that this is $\cos x$.

3. If $f(x) = \left\{ \begin{array}{lcl} e^{-\frac{1}{x^2}} & , & x \neq 0 \\ 0 & , & x = 0 \end{array} \right\}$. has a power series about $x = 0$ what is it?

The computation is somewhat more difficult here. $f'(x) = \frac{2}{x^3} e^{-\frac{1}{x^2}}$, $f''(x) = (-\frac{6}{x^4} - \frac{4}{x^6}) e^{-\frac{1}{x^2}}$, and generally $f^{(n)}(x) = (\frac{\alpha_{3n}}{x^{3n}} + \cdots + \frac{\alpha_{2n-2}}{x^{n+2}}) e^{-\frac{1}{x^2}}$ where the α_k are real numbers we don't need to find. If we let $x = 0$ in $f^{(n)}(x)$, the resulting expression has the indeterminate form $\infty \cdot 0$. Thus l'Hôspital's rule must be invoked. Now $f^{(n)}(x)$ is the sum of terms of the form $\frac{e^{-1/x^2}}{x^k}$, $k > 0$. What is $\lim_{x \to 0} \frac{e^{-1/x^2}}{x^k}$? Let $t = \frac{1}{x^2}$, and we must evaluate $\lim_{t \to 0} t^{k/2} e^{-t} = \lim_{t \to \infty} \frac{t^{k/2}}{e^t}$. If k is an even integer, applying l'Hôspital's rule $k/2$ times leaves a constant in the numerator and e^t in the denominator, so the limit is $\lim_{t \to \infty} \frac{\text{const}}{e^t} = 0$. If k is odd, applying l'Hôspital's rule $(k+1)/2$ times leaves a function of the form $\frac{\text{const}}{\sqrt{t} e^t}$, which also tends to 0 as $t \to \infty$.

What we have just shown is that $f^{(n)}(0) = 0$. The power series associated with e^{-1/x^2} is

$$0 + 0 \cdot x + \frac{0}{2!} x^2 + \cdots \frac{0}{n!} x^n + \cdots \equiv 0.$$

This function e^{-1/x^2}, whose power series about $x = 0$ is zero, is an example of a function which is clearly not equal to the power series, 0, associated with it.

To find if a given function has a power series expansion about x_0 we turn to Taylor's Theorem (also known as the extended mean value theorem). Now if a function f defined in a neighborhood of x_0 has a power series expansion there, we know the series is given by (5). Thus we should investigate

$$R_N(x) \equiv f(x) - \sum_{n=0}^{N} \frac{f^{(n)}(x_0)}{n!} (x - x_0)^n.$$

To say that f is equal to its series expansion is the same as saying that the remainder, $R_N(x)$, becomes arbitrarily small as $N \to \infty$. We must now seek an estimate of this remainder $R_N(x)$. Taylor's theorem is one way of finding an estimate.

Theorem 1.23 . *(Taylor's Theorem). Let f be a real-valued function with $N+1$ continuous derivatives defined on an interval containing x_0 and x. There exists a number ζ between x_0 and x such that*

$$f(x) = f(x) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \frac{f'''(x_0)}{3!}(x-x_0)^3 + \cdots$$

$$+\frac{f^{(N)}(x_0)}{N!}(x-x_0)^N + \frac{f^(N+1)(\zeta)}{(N+1)!}(x-x_0)^{N+1}. \tag{1-6}$$

In other words,

$$R_N(x) = \frac{f^{(N+1)}(\zeta)}{(N+1)!}(x-x_0)^{N+1}. \tag{1-7}$$

Remark: 1 The proof will only tell us that such a ζ exists but will give us no way to find it. In practice we often try to find some upper bound M for $f^(N+1)(\zeta)$, so $\left|f^(N+1)(\zeta)\right| \le M$, for *all* N, and only use the crude resulting estimate

$$|R_N(x)| \le \frac{M}{(N+1)!}\left|x-x_0\right|^{N+1}. \tag{1-8}$$

An example of this is the series for $\cos x$. Assuming the proof of the theorem, we know that (see Example b above) about $x_0 = 0$,

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots + \frac{(-1)^N}{(2N)!}x^{2N} + R_N(x),$$

where

$$R_N(x) = \frac{1}{(2N+2)!}[\frac{d^{2N+2}}{dx^{2N+2}}\cos x]_{x=\zeta}x^{2N+2},\ \zeta\epsilon(0,x).$$

Since

$$\left|\frac{d^{2N+2}}{dx^{2N+2}}\cos x\right|_{x=\zeta} \le 1,$$

we find that

$$|R_N(x)| \le \frac{1}{(2N+2)!}\left|x\right|^{2N+2}$$

Because, for fixed x, this remainder tends to 0 as $n \to \infty$, we have proved that the power series for $\cos x$ at $x_0 = 0$ does converge to $\cos x$, so in the limit

$$\cos x = \sum_{n=0}^{\infty}\frac{(-1)^n}{(2n)!}x^{2n}.$$

We can apply Theorem 16 and differentiate both sides of this to find the series for $\sin x$.
Remark: 2 Observe that Taylor's Theorem is only proved for *real-valued* functions f. It is not true if f is complex-valued. However using it we will be able to prove the inequality (7) for complex-valued f.
Proof: (Taylor's Theorem). Our proof is short—perhaps a little too slick. The trick is to appeal to the mean value theorem (really only Rolle's theorem is used).

Fix x and define the real number A by

$$f(x) = \sum_{n=0}^{N}\frac{f^(n)(x_0)}{n!}(x-x_0)^n + A\frac{(x-x_0)^{N+1}}{(N+1)!}. \tag{1-9}$$

Now let

$$H(t) := f(x) - \Big[f(t) + f'(t)(x-t) + \frac{f''(t)(x-t)^2}{2!} + \cdots + \frac{f^{(N)}(t)}{N!}(x-t)^N\Big] - A\frac{(x-t)^{N+1}}{(N+1)!}.$$

Thus we are letting x_0 vary, *not* x. Observe that $H(x) = 0$ (obviously) and $H(x_0) = 0$ (by definition of A). Since $H(t)$ satisfies the hypotheses of the mean value theorem, we conclude that there is some ζ between x_0 and x such that $H'(\zeta) = 0$. But

$$H'(t) = -f'(t) - \big[f''(t)(x-t) - f'(t)\big] - \cdots - \Big[\frac{f^{(N+1)}(t)}{N!}(x-t)^N - \frac{f^{(N)}(t)}{(N-1)!}(x-t)^{N-1}\Big]$$

$$-A\frac{(x-t)^N}{N!} = \frac{(x-t)^N}{N!}\big[A - f^{(N+1)}(t)\big].$$

Amazingly, almost all the terms canceled. Since $H'(\zeta) = 0$ and $\zeta \ne x$, we now know that $A = f^{(N+1)}(\zeta)$. Substitution of this value of A into (8) gives us exactly (6), which is just what we wanted to prove.

As an application let us prove the Binomial Theorem. That is the name given to the Maclaurin series for $(1+x)^\alpha$, where $\alpha \in \mathbb{R}$. The derivatives are easy to compute.

$$\begin{aligned} f(x) &= (1+x)^\alpha \\ f'(x) &= \alpha(1+x)^{\alpha-1} \\ f'''(x) &= \alpha(\alpha-1)(1+x)^{\alpha-2} \\ &\cdots \\ f^{(n)} &= \alpha(\alpha-1)\cdots.(\alpha-n+1)(1+x)^{\alpha-n}. \end{aligned}$$

Thus the power series about 0 associated formally with $(1+x)^\alpha$ is

$$\sum_{n=0}^{\infty}\frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!}x^n.$$

By the ratio test this series converges for $|x| < 1$. Does it converge to $(1+x)^\alpha$ when $|x| < 1$?

If α is a positive integer, $\alpha = N$, the terms in the power series from $n = N+1$ on all are zero since they contain the factor $(N-N)$. In this case we have only a finite series so convergence is trivial. The resulting polynomial is the familiar Binomial Theorem of high school algebra.

Let us therefore assume α is not a positive integer (or 0). Then we have an honest infinite series. In order to prove that $(1+x)^\alpha$ is equal to the infinite series, we must show that the remainder

$$R_N(x) \equiv (a+x)^\alpha - \sum_{n=0}^{N} \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!} x^n$$

tends to zero as $N \to \infty$. By Taylor's Theorem

$$R_N(x) = \frac{\alpha(\alpha-1)\cdots(\alpha-N)}{(N+1)!}(1+\zeta)^{\alpha-N-1} x^{N+1},$$

where ζ is between 0 and x. We shall prove that this tends to 0 as $N \to \infty$ only when $0 \le x < 1$. It is also true for $-1 < x \le 0$, but the proof is much longer so we will not give it [however a different attack yields the proof easily].

Now if $0 \le x < 1$, since $0 < \zeta < x$, then $1 < 1+\zeta$. Therefore for $N \ge \alpha$, we have $(z+\zeta)^{\alpha-N-1} < 1$. Thus

$$|R_N(x)| < \left| \frac{\alpha(\alpha-1)\cdots(\alpha-N)}{(N+1)!} x^{N+1} \right|$$

which does tend to zero as $N \to \infty$ (since it is the $N+1$st term of the convergent series $\sum^\infty \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!} x^n$, $|x| < 1$).

Although we have proved it only if $0 \le x < 1$, we shall state the complete

Theorem 1.24 *(Binomial Theorem). The function $(1+x)^\alpha$ is equal to a power series which converges for $|x| < 1$. It is*

$$(1+x)^\alpha = \sum_{n=0}^{\infty} \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!}. \tag{1-10}$$

In practice it is silly to memorize this formula since it is easier to expand $(1+x)^\alpha$ directly in a Maclaurin series, which we have just shown (partly anyway) is equal to the function.

We close this section with the generalization of Taylor's Theorem to complex-valued function $f(x)$.

Theorem 1.25 . *Let $f(x) = u(x) + iv(x)$ be a complex-valued function with $N+1$ continuous derivatives defined on an interval containing x_0 and x. There exists a real number M_N depending on N such that*

$$\left| f(x) - \sum_{n=0}^{N} \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n \right| \le \frac{M_N}{(N+1)!} |x-x_0|^{N+1} \tag{1-11}$$

Proof: Since f has $N+1$ continuous derivatives, so do the real-valued functions $u(x)$ and $v(x)$. Applying Taylor's Theorem to u and v, we find numbers ζ_1 and ζ_2, both between x_0 and x, such that

$$u(x) - \sum_{n=0}^{N} \frac{u^{(n)}(x_0)}{n!}(x-x_0)^n = \frac{u^{(N+1)}(\zeta_1)}{(N+1)!}(x-x_0)^{N+1},$$

and

$$v(x) - \sum_{n=0}^{N} \frac{v^{(n)}(x_0)}{n!}(x-x_0)^n = \frac{v^{(N+1)}(\zeta_2)}{(N+1)!}(x-x_0)^{N+1}.$$

Thus, by addition, since $f^{(n)} = u^{(n)} = iv^{(n)}$, we find

$$f(x) - \sum_{n=0}^{N} \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n = \frac{u^{(N+1)}(\zeta_1) + iv^{(N+1)}(\zeta_2)}{(N+1)!}(x-x_0)^{N+1}.$$

However since $u^{(N+1)}$ and $v^{(N+1)}$ are assumed continuous in an interval containing x_0 and x, they are bounded there, say by $\hat{M}_N$ and $\tilde{M}_N$. Taking absolute values of the last equation, we obtain equation (10) where $M_N = \sqrt{\hat{M}_N^2 + \tilde{M}_N^2}$.

Exercises

1. Find the Taylor series about the specified point x_0 and determine the interval of convergence for the following functions. You need not prove that the series do converge to the functions.

   a) $\sin x$, $x_0 = 0$,
   b) $\ln x$, $x_0 = 1$,
   c) $\frac{1}{x}$, $x_0 = -1$,
   d) $\sqrt{x}$, $x_0 = 6$,
   e) $\frac{1}{2}(e^x + e^{-x})$, $x_0 = 0$
   f) $\frac{x+i}{1+x}$, $x_0 = 0$,
   g) $\cos x$, $x_0 = \frac{\pi}{4}$,
   h) $\frac{1}{i+x}$, $x_0 = 0$
   i) e^{-x^2}, $x_0 = 0$,
   j) $(1+x+x^2)^{-1}$, $x_0 = 0$,
   k) $\cos x + i \sin x$, $x_0 = 0$,
   l) $\frac{1}{\sqrt{1+2x}}$, $x_0 = 0$.

2. Prove that in their interval of convergence about 0 the following power series associated with the given functions converge to the functions. Do this by proving that the remainder $|R_N(x)| \to 0$ as $N \to \infty$.
   a) $\sin x$,
   b) $\frac{1}{1+x^4}$,
   c) e^{-x}
   d) $\cosh x$ [Recall the definition: $\cosh x = \frac{e^x+e^{-x}}{2}$].

3. One often approximates $\frac{1}{\sqrt{1+x^2}}$ by $1-\frac{x^2}{2}$ when $|x|$ is small. Give some estimate of the error if a) $|x|+<10^{-1}$, b) $|x|<10^{-2}$, c) $|x|<10^{-4}$.

4. Use the Taylor series

$$e^{-x^2} = 1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \cdots + \frac{(-1)^n x^{2n}}{n!} + \cdots .$$

   to evaluate $\int_0^1 e^{-x^2}\,dx$ to three decimal places. I suggest using Theorem 16 and the error estimate of Theorem 12.

5. Assume the ordinary differential equation $y'-y=0$, with $y(0)=1$ has a power series solution $y(x)=\sum_{n=0}^{\infty} a_n x^n$ about $x=0$. a). Substitute this series directly into the differential equation and solve for the coefficients a_n. b). Find when the series converges; c). justify (a posteriori) the fact that the function defined by the convergent series does satisfy the differential equation. [We do not yet know that this is the only solution. All we know is that it is the only solution which has a power series].

6. In this exercise you will prove that e is irrational. It all hinges on the series for 3.

$$e = 1 + 1 + \frac{1}{2} + \frac{1}{3!} + \cdots + \frac{1}{n!} + \cdots .$$

   a) Prove that $2<e<3$, so e is not an integer (cf. page 58, bottom).
   b) Assume e is rational, $e=\frac{p}{q}$, where p and q are integers with no common factor and $q \ge 2$. Then use the Taylor series with q terms and the remainder R_q to show that $e \cdot q! = N + \frac{e^{\zeta}}{q+1}$, where $0<\zeta<1$, and N is an integer.
   c) From this deduce that $\frac{e^{\zeta}}{q+1}$ must be an integer, and show that this contradicts $e^{\zeta} < e' < 3$, and $q+1 \ge 3$.

7. This exercise generalizes the form of the remainder (6') in Taylor's Theorem. Fix x and define the number B by

$$f(x) = \sum_{n=0}^{N} \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + B(x-x_0)^{\alpha},\ \alpha \ge 1.$$

   Then consider the function $H(t)$ defined by

$$H(t) \equiv f(x) - \sum_{n=0}^{N} \frac{f^{(n)}(t)}{n!}(x-t)^n - B(x-t)^{\alpha}.$$

   Show that there is a ζ between x_0 and x such that

$$B = \frac{f^{(N+1)}(\zeta)}{\alpha N!}(x-\zeta)^{N+1-\alpha},$$

   so that

$$R_N = \frac{f^{(N+1)}(\zeta)}{\alpha N!}(x-x_0)^{\alpha}(x-\zeta)^{N+1-\alpha}.$$

   This is Schlomilch's form of the remainder. In the special case $\alpha = N+1$, we obtain Lagrange's form of the remainder, (6) found previously, while for $\alpha = 1$ we obtain Cauchy's form of the remainder

$$R_N = \frac{f^{(N+1)}(\zeta)}{N!}(x-x_0)(x-\zeta)^N.$$

Here are two applications of Taylor's Theorem to problems other than infinite series. The first one deals with max-min. Let $f(x)$ be a sufficiently smooth function (by which we mean f has plenty of derivatives—we'll specify the number later). Now we know that if f has a local maximum or minimum at x_0, then $f'(x_0)=0$, and it is a maximum if $f''(x_0)<0$, a minimum if $f''(x_0)>0$. But what if $f''(x_0)=0$? Consider the examples $f_1(x)=x^4$, $f_2(x)=-x^4$, $f_3(x)=x^3$, the first of which has a minimum at $x=0$, the second a maximum at $x=0$, while the third has neither. These three examples suggest the criterion will depend upon the lowest non-zero derivative being an even or odd derivative, and on its sign.

A FIGURE GOES HERE

By the definition of local maximum and minimum, the issue is the behavior of $f(x)$ in a neighborhood of x_0, that is, the nature of $f(x_0+h)$ for $|h|$ small. We remind you that f has a local max at x_0 if $f(x_0+h)-f(x_0) \le 0$ for all $|h|$ sufficiently small, and a local min at x_0 if $f(x_0+h)-f(x_0) \ge 0$ for all $|h|$ sufficiently small. Since the behavior of $f(x)$ near x_0 is determined by the Taylor polynomial

$$f(x_0+h) = f(x_0) + f'(x_0)h + \frac{f''(x_0)h^2}{2!} + \cdots + \frac{f^{(n)}}{n!}(x_0)h^n + \frac{f^{(n+1)}(\zeta)h^{n+1}}{(n+1)!}$$

where ζ is between x_0 and x_0+h, it is natural to look at this polynomial to answer our question.

Theorem 1.26 *Assume f has (at least) $n+1$ continuous derivatives in some interval containing x_0. Say $f'(x_0) = f''(x_0) = \ldots = f^n(x_0) = 0$ but $f^{(n+1)}(x_0) \neq 0$, then*

(a) if n is even, then f has neither a max nor min at x_0.

(b) if n is odd, then

i) f has a max at x_0 if $f^{(n+1)}(x_0) < 0$.

ii) f has a min at x_0 if $f^{(n+1)}(x_0) > 0$.

PROOF: We shall use Taylor's polynomial with $n+1$ terms.

Since the first n derivatives vanish at x_0, we have $f(x_0+h) - f(x_0) = \frac{f^{(n+1)}(\zeta)}{(n+1)!}h^{n+1}$, ζ between x_0 and x_0+h. Because $f^{(n+1)}(x)$ is assumed continuous at x_0, $f^{(n+1)}(\zeta)$ must have the same sign as $f^{(n+1)}(x_0)$ in some neighborhood of x_0. Restrict your attention to the neighborhood. *If n is even*, $n+1$ is odd, so that h^{n+1} is positive if $h > 0$, negative if $h < 0$. Thus $f(x_0+h) - f(x_0)$ changes sign in any neighborhood of x_0. However *if n is odd*, h^{n+1} is positive no matter if $h > 0$ or $h < 0$. Therefore $f(x_0+h) - f(x_0)$ has the same sign as $f^{(n+1)}(x_0)$ throughout some neighborhood about x_0. The precise conditions are easy to verify now.

EXAMPLES:

1. $f(x) = x^5 + 1$ has neither a max nor min at $x = 0$, since $f'(0) = \ldots = f^{(4)}(0) = 0$, but $f^{(5)}(0) = 5! \neq 0$.

2. $f(x) = (x-1)^6 - 7$ has a min at $x = 1$ since $f'(1) = \ldots = f^{(5)}(1) = 0$, but $f^{(6)}(1) = 6! > 0$.

Our second application is a geometrical interpretation of the Taylor polynomial. Given the function $f(x)$, consider the polynomial

$$P_n(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n,$$

whose first n derivatives agree with those of f at $x = x_0$. $P_1(x) = f(x_0) + f'(x_0)(x-x_0)$ is the equation of the tangent to the curve $y = f(x)$ at x_0. It is the straight line which most closely approximates the curve at x_0. Similarly $P_2(x)$ is the parabola which most closely approximates the curve at x_0. Generally, $P_n(x)$ is the polynomial of degree n which most closely approximates the curve $y = f(x)$ at the point x_0. Using this Taylor polynomial, we can define the order of contact of two curves at a point.

DEFINITION: The two curves $y = f(x)$ and $y = g(x)$ have *order of contact n at the point x_0* if their Taylor polynomials of degree n at x_0 are identical, but their $n+1$st Taylor polynomials differ.

An equivalent definition is that $f(x_0) = g(x_0)$, $f'(x_0) = g'(x_0)$, $\ldots$, $f^{(n)}(x_0) = g^{(n)}(x_0)$, but $f^{(n+1)}(x_0) \neq g^{(n+1)}(x_0)$. We have assumed that f and g have $n+1$ continuous derivatives. If f and g have contact n at x_0, then

$$f(x_0+h) - g(x_0+h) = \frac{f^{(n+1)}(\zeta_1) - g^{(n+1)}(\zeta_2)}{(n+1)!}h^{n+1}.$$

One interesting consequence of this formula is that if f and g have contact of even order, then the curves will cross at x_0, while if the contact is of odd order, the curves will *not* cross in some neighborhood of x_0.

We can define the curvature of a curve in the plane by using the concept of contact. First we define the curvature of a circle (whose curvature had better be constant).

DEFINITION: The *curvature* k of a *circle* of radius R is defined to be $\frac{1}{R}$, $k = \frac{1}{R}$.

Thus the smaller the circle, the larger the curvature—a natural outcome. Furthermore, a straight line—which may be thought of as a circle with infinite radius—has curvature zero. How can we define the curvature of a given curve? For all non-circles, the curvature will clearly vary from point to point of the curve. Thus, the concept we want is the curvature of a given curve $y = f(x)$ *at a point* x_0. Our definition should appear reasonable.

DEFINITION: The *curvature k of a plane curve* $y = f(x)$ at the point x_0 is the curvature of the circle which has contact of order two at x_0.

This circle which has contact of order two is called the *osculating circle* to the curve at x_0 (osculate: Latin, to kiss). Let us convince ourselves that there is only one osculating circle (for if there were two, the curvature would not b e well defined.) Consider all circles of contact one to $f(x)$ at x_0. These are all circles tangent to $f(x)$ at x_0. Their centers lie on the line l normal to the curve at x_0 ("normal" means perpendicular to the tangent line). It is geometrically clear that of these circles with contact 1, there will be exactly one with contact 2.

EXAMPLE: Find the curvature of $y = e^x$ at $x = 0$. The slope of the curve at $(0,1)$ is 1. Therefore the equation of the normal is $y - 1 = -x$. Since the center (x_0, y_0) of the osculating circle must lie on this line, and the circle contains the point $(0,1)$, subject to $y_0 = 1 - x_0$, the value of x_0 must be determined from the fact that the second derivative of the circle $(0,1)$ must equal the second derivative of $y = e^x$ at $x = 0$, that is, it must equal 1. But for any circle, $(y - y_0)y'' + y'^2 + 1 = 0$. In our case $y' = 1$ at $(0,1)$ (recall the circle is tangent to e^x at $(0,1)$), so that $(1 - y_0) \cdot 1 + 1 + 1 = 0$, or $y_0 = 3$. The equation $y_0 = 1 - x_0$ implies that $x_0 = -2$. Thus the equation of the osculating circle is $(y-3)^2 + (x+2)^2 = 8$, and the curvature of $y = e^x$ at $x = 0$ is $k = \frac{1}{\sqrt{8}}$. Later on we will give another definition of curvature which is applicable not only to plane curves, but also to curves in space.

Exercises

1. What is the order of contact of the curves $y = e^{-x}$ and $y = \frac{1}{1+x} + \frac{1}{2}\sin^2 x$ at $x = 0$?

2. Find the osculating circle and curvature for the curve $y = x^2$ at $x = 1$.

3. Show that at $x = a$, the curve $y = f(x)$ has curvature $k = \frac{f''(a)}{[1+f'(a)^2]^{\frac{3}{2}}}$ and the center of the osculating circle is at the point $(a - \frac{f'(a)}{f''(a)}[1 + f'(a)^2],\ f(a) + \frac{1+f'(a)^2}{f''(a)})$. What is the messy equation of the osculating circle?

4. At the given points, the following curves have slope zero. Determine if the curve has a max, min, or neither there.

   (a). $y = (x+1)^4$, $x = -1$,

   (b). $y = x^2 \sin x$, $x = 0$.

5. Let P_1, P, and P_2 be three distinct points on the curve $y = f(x)$, and consider the circle passing through those three points. Show that in the limit as both P_1 and P_2 approach P, this circle becomes the osculating circle. (Hint: Taylor's Theorem will be needed here).

6. In this problem we outline another derivation of Taylor's Theorem. Whereas the one in the notes did not use the fact the $f^{(n+1)}$ was continuous, this proof relies upon that fact.

   a) Show that

$$\int_{x_0}^{x} \frac{(x-t)^{k-1}}{(k-1)!} f^{(k)}(t)\,dt = f^{(k)}x_0 \frac{(x-x_0)^k}{k!} + \int_{x_0}^{x} \frac{(x-t)^k}{k!} f^{(k+1)}(t)\,dt.$$

   b) Prove by induction that

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \int_{x_0}^{x} \frac{(x-t)^n}{n!} f^{(n+1)}(t)\,dt.$$

   The remainder is expressed as an integral here. It is because $f^{(n+1)}$ is to be integrated that we require its continuity.

7. a) Let $g(x)$ have contact of order n with the function 0 at the point $x = a$, and assume that $f(x)$ has contact of order at least n with the function 0 at $x = a$. Use Taylor's Theorem to prove that

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \frac{f^{(n+1)}(a)}{g^{(n+1)}(a)}$$

   This is *l'Hôspital's Rule.*

   b) Apply l'Hôspital's rule to evaluate

$$\text{i) } \lim_{x \to 0} \frac{x - \sin x}{x^3}, \quad \text{ii) } \lim_{\theta \to \frac{\pi}{4}} \frac{1 - \tan\theta}{\theta - \frac{\pi}{v}}$$

8. Assume f has two derivatives in the interval $[a, b]$, and assume that $f'' \geq 0$ throughout the interval. Prove that if ζ is any point in $[a, b]$, then the curve $y = f(x)$ never falls below its tangent at the point $x = \zeta$, $y = f(\zeta)$. [Hint: Use Taylor's Theorem with three terms].

9. Use Cauchy's form of the remainder (p. 103-4, no. 7) for Taylor's Theorem to prove that the binomial series converges to $(1+x)^\alpha$ for $-1 < x \leq 0$. This will complete the proof of the binomial theorem.

10. The n^{th} Legendre polynomial $P_n(x)$ is defined by $P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}[(x^2-1)^n]$. Prove that $P_n(x)$ is a polynomial of degree n and has n distinct real zeros in the interval $(-1, 1)$.

11. Verify that e^{ax} is a solution of $y' = ay$. Prove that every solution has the form Ae^{ax}, where A is a constant.

12. Assume that $f(x)$ has plenty of derivatives in the interval $[a, b]$, and that f has $n+1$ distinct zeros in the interval. Prove that there is at least one $c \in (a, b)$ such that $f^{(n)}(c) = 0$.

1.6 Complex-Valued Functions, e^z, $\cos z$, $\sin z$.

The task of this section is to answer the following question. Say $f(x)$ is a real or complex valued function of the *real* variable x. How can we define $f(z)$ where z is *complex*? For example, if $P(x) = a_0 + a_1 x + \cdots + a_n x^n$ is a polynomial, the answer is easily given: just define $P(z) = a_0 + a_1 z + \cdots + a_n z^n$. Since this function only involves addition and multiplication of complex numbers, for any complex z the number $P(z)$ can be computed. Similarly any rational function, $\frac{P(x)}{Q(x)}$, where $P(x)$ and $Q(x)$ are both polynomials, can be defined for complex z as $\frac{P(z)}{Q(z)}$ since both $P(z)$ and $Q(z)$ are defined separately and we can then take their quotient.

But how do we define e^z, or $\cos z$, or $(1+z)^\alpha$, where $\alpha \epsilon \mathbb{R}$ is not a positive integer? As might have been suspected, the trick is to use infinite series.

Definition: If $f(x)$, $x \epsilon \mathbb{R}$, has a convergent Taylor series,

$$f(x) = \sum_{n=0}^{\infty} a_n x^n, \qquad |x| < \rho,$$

then we define $f(z)$, $z \epsilon \mathbb{C}$, by the infinite series

$$f(z) = \sum_{n=0}^{\infty} a_n z^n,$$

and the infinite series converges throughout the disc $|z| < \rho$.

The assertion that the complex series converges throughout the disc $|z| < \rho$ is an immediate consequence of Theorem 13 on page ?.

Thus, for example, we *define*.

$$E(z) = \sum_{n=0}^{\infty} \frac{1}{n!} z^n,$$

$$C(z) = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n}}{(2n)!},$$

$$S(z) = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{(2n+1)!},$$

and

$$(1+z)^\alpha = \sum_{n=0}^{\infty} \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!} z^n, \quad \alpha\epsilon\mathbb{R}$$

where the first three series converge for all $z\epsilon\mathbb{C}$, while the last converge for $|z| < 1$. We have temporarily used the notation $E(z)$ in place of e^z, $C(z)$ for $\cos z$, and $S(z)$ for $\sin z$ so that you do not jump to hasty conclusion s about these functions by merely extrapolating your knowledge of e^x etc. For example it is *not* true that $|\sin z| \le 1$ for all $z\epsilon\mathbb{C}$, even though $|\sin x| \le 1$ for all $x\epsilon\mathbb{R}$. All properties of these function s for $z\epsilon\mathbb{C}$ must be proved again beginning with the power series definitions. Known properties of e^x, $x\epsilon\mathbb{R}$ and wishful thinking don't prove properties of e^z, $z\epsilon\mathbb{C}$. Let us begin by proving

Theorem 1.27 .

(a) $E(iz) = C(z) + iS(z)$, *for all* $z \in \mathbb{C}$.

(b) $E(-iz) = C(z) - iS(z)$, *for all* $z \in \mathbb{C}$.

(c) $C(z) = \frac{1}{2}[E(iz) + E(-iz)]$, *for all* $z \in \mathbb{C}$.

(d) $S(z) = \frac{1}{2i}[E(iz) - E(-iz)]$, *for all* $z \in \mathbb{C}$.

PROOF: a). b). Just substitute and rearrange the series. For example

$$C(z) = 1 - \frac{z^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots.$$

$$iS(z) = i[z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots]$$

so

$$C(z) + iS(z) = 1 + iz - \frac{z^2}{2!} - i\frac{z^3}{3!} + \frac{z^4}{4!} + i\frac{z^5}{5!} - \cdots,$$

where the adding of the two series is justified by Theorem 5(page ?). We must compare the last series with that for $E(iz)$:

$$E(iz) = 1 + iz + \frac{(iz)^2}{2!} + \frac{(iz)^3}{3!} + \frac{(iz)^4}{4!} + \cdots = 1 + iz - \frac{z^2}{2!} - i\frac{z^3}{3!} + \frac{z^4}{4!} + \cdots,$$

which is identical to the series for $C(z) + iS(z)$.

c)-d). These follow by elementary algebra from a) and b).

The formulas a)-d) of Theorem 21 show there is a close connection between the four functions $E(iz)$, $E(-iz)$, $C(z)$, and $S(z)$. Our next theorem shows that the formula $e^x e^y = e^{x+y}$, x, $y \in \mathbb{R}$, extends to the function $E(z)$.

Theorem 1.28 . *$E(z)E(w) = E(z+w)$, for all z, $w \in \mathbb{C}$.*

PROOF: We must show that

$$(\sum_{n=0}^{\infty} \frac{z^n}{n!})(\sum_{n=0}^{\infty} \frac{w^n}{n!}) = \sum_{n=0}^{\infty} \frac{(z+w)^n}{n!},$$

The product of the two series is defined in Theorem 15. Using that definition, we find that

$$(\sum_{n=0}^{\infty} \frac{z^n}{n!})(\sum_{n=0}^{\infty} \frac{w^n}{n!}) = \sum_{n=0}^{\infty} (\sum_{k=0}^{n} \frac{z^k}{k!} \frac{w^{n-k}}{(n-k)!}).$$

However, the binomial theorem for *positive integer* exponents (which only uses the *algebraic* rules for complex numbers) states that

$$(z+w)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} z^k w^{n-k}.$$

Upon substituting this into the last equation, we obtain the desired formula.

The formula of this theorem is the key to many results, like the following generalization of $\sin^2 x + \cos^2 x = 1$.

Corollary 1.29 *$C(z)^2 + S(z)^2 = 1$ for all $z \in \mathbb{C}$.*

PROOF: We use equations a) and b) of Theorem 21 to reduce the question to one of exponentials.

$$E(iz)E(-iz) = [C(z) + iS(z)][C(z) - iS(z)] = C^2(z) + S^2(z).$$

But by Theorem 22, $E(iz)E(-iz) = E(iz - iz) = E(0)$. Directly from the power series we see that $E(0) = 1$. This proves the formula.

Our next corollary states that the addition formulas for $\sin x$ and $\cos x$ are still valid for $C(z)$ and $S(z)$.

Corollary 1.30 *$C(z+w) = C(z)C(w) - S(z)S(w)$ and $S(z+w) = S(z)C(w) - S(w)C(z)$ for all z, $w \in \mathbb{C}$*

PROOF: A direct algebraic computation does the job.

$$C(z+w)+iS(z+w)=E(iz+iw)=E(iz)E(iw)=[C(z)+iS(z)][C(w)+iS(w)]$$

$$=[C(z)C(w)-S(z)S(w)]+i[S(z)C(w)+S(w)C(z)].$$

Similarly we find that

$$C(z+w)-iS(z+w)=[C(z)C(w)-S(z)S(w)]-i[S(z)C(w)+S(w)C(z)].$$

Addition of these two equations gives the formula for $C(z+w)$, while subtraction gives the formula for $S(z+w)$.

Had we but world enough, and time, we would linger a while. A lovely result we have not proved is that $E(z+2\pi i)=E(z)$, the periodicity of $E(z)$, which is a consequence of the formulas $C(z+2\pi)=C(z)$, and $S(z+2\pi)=S(z)$, the periodicity of $C(z)$ and $S(z)$, by using Theorem 21 (but see pp. ??).

We shall close this chapter by restating the results proved above in the usual language of e^z etc. instead of the temporary notation $E(z)$ etc. we have been using.

$$e^{iz}=\cos z+i\sin z \tag{1-12}$$
$$e^{-iz}=\cos z-i\sin z \tag{1-13}$$
$$\cos z=\tfrac{1}{2}(e^{iz}+e^{-iz}) \tag{1-14}$$
$$\sin z=\tfrac{1}{2i}(e^{iz}-e^{-iz}) \tag{1-15}$$
$$e^z e^w=e^{z+w} \tag{1-16}$$
$$\sin^2 z+\cos^2 z=1 \tag{1-17}$$
$$\cos(z+w)=\cos z\cos w-\sin z\sin w \tag{1-18}$$
$$\sin(z+w)=\sin z\cos w+\sin w\cos z \tag{1-19}$$

Generally, *all algebraic* formulas for $\sin x$, $\cos x$, and e^x remain valid for $\sin z$, $\cos z$, and e^z. In fact any *algebraic* relationship between any combination of analytic functions remains valid as we change the in dependent variable from a real x to the complex z. Inequalities almost always fall apart in the transition from $x\in\mathbb{R}$ to $z\in\mathbb{C}$. Exercise 2e below illustrates this.

One formula which we will use frequently later on is a specialization of (1-12) to the case when z is real. Then writing the real z as θ we have Euler's beautiful formula

$$e^{i\theta}=\cos\theta+i\sin\theta,\ \theta\in\mathbb{R}. \tag{1-20}$$

We cannot resist stating this formula down again for $\theta=\pi$:

$$e^{i\pi}=-1,$$

an almost mystical identity connecting the four numbers e, $i\,\pi$, and -1. Notice that (1-20) also implies $\left|e^{i\theta}\right|=1$.

If we write $z=x+iy$, then using (1-16) and (1-20) we find

$$e^z=e^{x+iy}=e^xe^{iy}=e^x(\cos y+i\sin y). \tag{1-21}$$

A consequence of this is

$$|e^z|=e^x \tag{1-22}$$

Exercises

1. Observe that (directly from the power series)

$$\cos(-z)=\cos z,\ \text{ and }\ \sin(-z)=-\sin z.$$

Use this and the addition formula for $\cos(z+w)$ to prove that $\sin^2 z+\cos^2 z=1$.

2. If we define $\sin hx=\frac{1}{2}(e^x-e^{-x})$ and $\cos hx=\frac{1}{2}(e^x+e^{-x})$, $x\in\mathbb{R}$, we prove that
   a) $\cos ix=\cos h\,x$, $\sin ix=i\sin h\,x$
   b) $\cos z=\cos h\,y-i\sin x\sin h\,y$, $(z=x+iy)$
      $\sin z=\sin x\cos h\,y+i\cos x\sin h\,y$
   c) $|\cos z|^2=\cos^2 x+\sin h^2 y$
      $|\cos z|^2=\cos h^2 y-\sin^2 x$
      $|\sin z|^2=\sin^2 x+\sin h^2 y$
      $|\sin z|^2=\cos h^2 y-\cos^2 x$
   d) Use the identities of part c) to deduce that

$$|\sin h\,y|\le|\cos z|\le\cos h\,y$$

$$|\sin h\,y|\le|\sin z|\le\cos h\,y$$

   e) Prove that there is some $z\in\mathbb{C}$ such that

$$|\sin z|>1,\ \text{ and }\ |\cos z|>1.$$

3. Define the derivative of $f(z)$ at z_0, where z, $z_0\in\mathbb{C}$, as

$$\lim_{z\to z_0}\frac{f(z)-f(z_0)}{z-z_0},$$

if the limit exists.

a) By working directly with the power series, show that e^z is differentiable for all z, and that
$$\frac{d}{dz}e^{az} = ae^{az},\ a,\ z \in \mathbb{C},$$

b) Apply this to (1-12) and (1-13) to deduce that
$$\frac{d}{dz}\cos z = -\sin z,\ \frac{d}{dz}\sin z = \cos z$$
(We cannot appeal to Theorem 16 and differentiate term-by-term since that theorem assumed the independent variable, x, was real).

4. Use the results of Exercise 2c to show that the only complex roots $z = x + iy$ of $\sin z$ and $\cos z$ are at the points on the real axis $y = 0$ where $\sin x = 0$ and $\cos x = 0$, respectively.

5. Use the results of this section to prove *DeMoirve's Theorem*
$$(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta,\ \theta \in \mathbb{R},$$
where n is a positive integer.

6. a) Show that the sum of the finite geometric series $\sum e^{inx}$ is
$$\sum_{n=1}^{N} e^{inx} = \frac{e^{i(N+1/2)x} - e^{ix/2}}{e^{ix/2} - e^{-ix/2}}.$$

b) Take the real and imaginary parts of the above formula and prove that for all $x \neq 0,\ x \in (0, 2\pi)$,
$$\sum_{n=1}^{N}\cos x = \frac{\sin(N+1/2)x - \sin 1/2x}{2\sin\frac{1}{2}x}$$
$$\sum_{n=1}^{N}\sin nx = \frac{\cos\frac{1}{2}x - \cos(N+\frac{1}{2})x}{2\sin\frac{1}{2}}.$$

1.7 Appendix to Chapter 1, Section 7.

As a special dessert let us take some time out and prove some interesting results you would probably never see otherwise. We have in mind to *define* a specific number $\alpha \in \mathbb{R}$ as the smallest positive zero of $\cos x,\ x \in \mathbb{R}$—so α had better turn out as $\pi/2$. Then we prove that 1) $\sin(x + 4a) = \sin x$ etc., 2) the ratio of the circumference to diameter of a circle is

2α so that 2α does equal the π of public school fame. Furthermore, we also present a way of computing α.

In this section we take $\sin z$ and $\cos z$, $z \in \mathbb{C}$ to be defined by their power series, and use *only* the properties of these functions which were obtained from the power series definition.

Lemma 1.31 *The set $A = \{x \in \mathbb{R}\colon \cos x = 0,\ 0 < x < 2\}$ is not empty, that is, the equation $\cos x = 0$ has at least one real root for $x \in (0, 2)$.*

PROOF: Since $\cos x$ is defined by a convergent power series, it is continuous (even infinitely differentiable); furthermore because $x \in \mathbb{R}$ and the power series has real coefficients, we know that $\cos x,\ x \in \mathbb{R}$ is real-valued. Observe that $\cos 0 = 1 > 0$, and the following crude inequality
$$\begin{aligned}\cos 2 &= 1 - \frac{2^2}{1\cdot 2} + \sum_{n=2}^{\infty}\frac{(-1)^n 2^{2n}}{(2n)!} < -1 + \sum_{n=2}^{\infty}\frac{2^{2n}}{(2n)!} \\ &< -1 + \frac{2^4}{4!}\sum_{k=0}^{\infty}(\frac{2}{5})^{2k} = -1 + \frac{50}{63} < 0.\end{aligned} \tag{1-23}$$
Thus $\cos 0 > 0$ and $\cos 2 < 0$, so there is at least one point in $(0, 2)$ where the real-valued continuous function $\cos x$ vanishes. This proves the lemma.

Denote the g.l.b of A (which does exist since A is bounded—say by 0 and 2) by α. We shall show that $\alpha \in A$. Since α is the g.l.b. of A, there exists a sequence of points $\alpha_k \in A$ (the α_k may just be the same point repeated over and over) such that $\alpha_k \to \alpha$ and $\cos\alpha_k = 0$. But since $\cos x$ is continuous,
$$0 = \lim_{k\to\infty}\cos\alpha_k = \cos\alpha,$$
so in fact $\cos\alpha = 0$ too $\Rightarrow \alpha \in A$.

Now $\cos x$ must be positive throughout the interval $[0, \alpha)$, since it is positive at $x = 0$ and α is the first place it vanishes. Therefore the formula $\frac{d}{dx}\sin x = \cos x$—obtained by differentiating the *real* power series for $\sin x$ term by term—shows that $\sin x$ is increasing for $x \in [0, \alpha)$. Since $\sin 0 = 0$, we see that $\sin x \geq 0$ *for* $x \in [0, \alpha)$. Thus the formula $\frac{d}{dx}\cos x = -\sin x$ tells us that $\cos x$ *is decreasing in the interval* $[0, \alpha]$. From the formula
$$1 = \sin^2\alpha + \cos^2\alpha = \sin^2\alpha,$$
and the fact that $\sin\alpha > 0$, we find that $\sin\alpha = 1$. We can thus conclude from the addition formulas for $\sin x$ and $\cos x$ the:

Theorem 1.32 *Let α denote the smallest zero of $\cos x$ for $x > 0$. Then*
$$\cos\alpha = 0,\ \cos 2\alpha = -1,\ \cos 3\alpha = 0,\ \cos 4\alpha = 1$$
$$\sin\alpha = 1,\ \sin 2\alpha = 0,\ \sin 3\alpha = -1,\ \sin 4\alpha = 0,$$

or more generally

$$\cos(z+\alpha) = -\sin z,\ \sin(z+\alpha) = \cos z$$

$$\cos(z+4\alpha) = \cos z,\ \sin(z+4\alpha) = \sin z$$

This proves that the $\sin z$ *and* $\cos z$ *are periodic with period* 4α.

As you have guessed, α is another name for $\pi/2$—and serves as our *definition of* π. This is based upon power series and is independent of circles or triangles—or even the entire concept of angle. A simple consequence is the

Corollary 1.33 *The function* e^z *is periodic with period* $4\alpha i$,

$$e^{z+4\alpha i} = e^z e^{4\alpha i} = e^z.$$

PROOF: $e^{z+4\alpha} = e^z e^{4i\alpha} = e^z(\cos 4\alpha + i\sin 4\alpha) + e^z(1+i0) = e^z$.

Two issues remain to be settled before closing up. We should 1) prove that the ratio of the circumference C of a circle to its diameter D is π, i.e., $C = 2\alpha D$, and 2) find some way of approximating α numerically (for all we know of $alpha$ so far is that it is the smallest element in a set and $0 < \alpha < 2$). The two problems are closely related.

The circle of radius R has the equation $x^2+y^2 = R^2$. Consider the portion in the first quadrant. Then using the familiar formulas for arc length, we find that

$$\frac{C}{4} = R\int_0^R \frac{dx}{\sqrt{R^2-x^2}} = R\int_0^1 \frac{dt}{\sqrt{1-t^2}},$$

where the change of variable $x = Rt$ has been used to obtain the last integral [this is legal since the mapping "multiply by R" is a bijection and hence an invertible function]. Thus, the desired result, $C = 2\alpha D = 4\alpha R$ will be proved if we can p rove

Theorem 1.34 $\int_0^1 \frac{dt}{\sqrt{1-t^2}} = \alpha(=\frac{\pi}{2})$

Corollary 1.35 *If* C *denotes the arc length of the circumference of a circle of radius* R, *then* $C = 4\alpha R$.

PROOF: of Theorem. We want to make the change of variable $t = \sin\zeta$, where $t \in [0,1]$. In order to do this we must only check that the function $\sin\zeta$ is differentiable and invertible function there. We know it is differentiable . Since $\sin x$ is continuous and monotone increasing for $x \in [0,\alpha]$, and since the end points are mapped into 0 and 1 respectively ($\sin 0 = 0$, $\sin\alpha = 1$), the function $f(\zeta) = \sin\zeta$ is invertible for $x \in [0,\alpha] \iff t \in [0,1]$. The usual formulas are applicable and yield

$$\int_0^1 \frac{1}{\sqrt{1-t^2}}\,dt = \int_0^\alpha d\zeta = \alpha$$

Q.E.D.

To compute $\pi = 2\alpha$, it is convenient to introduce $\tan z = \sin z/\cos z$, for all z where $\cos z \neq 0$. In particular $\tan x$ is defined for all real x in the interval $0 \le x < \alpha/2$. From the behavior of $\sin x$ and $\cos x$ in the interval $x \in [0,\alpha/2)$, it is easy to show that $\tan x$ has infinitely many derivatives and is increasing for $x \in [0,\alpha/2)$, assuming the values from $0 = \tan 0$ to $1 = \tan\frac{\alpha}{2}$. The function $\tan x$ is therefore invertible in that interval, so we can make the natural change of variable $t = \tan x$ and obtain

$$\int_0^1 \frac{dt}{1+t^2} = \int_0^{\alpha/2} \frac{1}{1+\tan^2 x}\left(\frac{d}{dx}\tan x\right)dx = \int_0^{\alpha/2} dx = \frac{\alpha}{2}.$$

But the integral on the left can be approximated readily because of the algebraic identity

$$\frac{1}{1+t^2} = \sum_0^N (-1)^n t^{2n} + \frac{(-1)^{N+1}t^{2N+2}}{1+t^2}, \quad \text{all } t \neq i.$$

Thus

$$\frac{\pi}{4} = \frac{\alpha}{2} = \int_0^1 \frac{dt}{1+t^2} = \sum_0^N (-1)^n \int_0^1 t^{2n}\,dt + (-1)^{N+1}\int_0^1 \frac{t^{2N+2}}{1+t^2}\,dt,$$

or

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots + \frac{(-1)^N}{2N+1} + R_N,$$

where since $2t \le 1+t^2$ the remainder R_N can be estimated by

$$|R_N| = \int_0^1 \frac{t^{2N+2}}{1+t^2}\,dt < \int_0^1 \frac{t^{2N+2}}{2t}\,dt = \frac{1}{4N+4}$$

If the first 250 terms in the series are used, $N = 250$, we find

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \cdots + \frac{1}{251} + R_{250},$$

where $|R_{250}| < \frac{1}{1004} < \frac{1}{1000}$, so three decimal accuracy is obtained. This is quite slow—but it does work. For practical computations, a series which converges much faster is needed. See exercise 2 below; it is neat.

Since $R_N \to 0$ as $N \to \infty$, the following formula is a consequence of our effort:

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots..$$

Exercises

1. Use the method illustrated here to slow that

$$\ln 2 = \int_0^1 \frac{1}{1+x}\,dx = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots + \frac{(-1)^{N+1}}{N} + R_N,$$

where $\lim_{N\to\infty} R_N = 0$. Find an N such that $|R_N| < 10^{-3}$.

[Hint: Write $\frac{1}{1+x} = \sum_0^N (-1)^n x^n + \frac{(-1)^{N+1}x^{N+1}}{1+x}$, $x \neq -1$].

2. To approximate $\frac{\pi}{4}$ with fewer terms, the following clever device works. Write

$$\frac{1}{1+t^2} = \sum_{0}^{N-1}(-1)^n t^{2n} + \frac{(-1)^N t^{2N}}{2} + (\frac{(-1)^N t^{2N}}{2} + \frac{(-1)^{N+1} t^{2N+2}}{1+t^2})$$

and show that

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} + \cdots + \frac{(-1)^{N-1}}{2N-1} + \frac{(-1)^N}{2(2N-1)} + \tilde{R}_N,$$

where $\tilde{R}_N + \frac{(-1)^N}{2}\int_0^1 \frac{t^{2N}-t^{2N+2}}{1+t^2}\,dt$.

a) Prove that $\left|\tilde{R}_N\right| < \frac{1}{8N^2+8N}$.

b) What should N be to make $\left|\tilde{R}_N\right| < 10^{-3}$? Amazing saving, isn't it? The technique does generalize to other series and can be refined to yield even better results.

c) Apply the method given here to problem 1 above to show that $\ln 2 = 1 - \frac{1}{2} + \frac{1}{3} + \cdots + \frac{(-1)^N}{N-1} + \frac{1}{2}\frac{(-1)^{N+1}}{N} + \tilde{R}_N$, where $\left|\tilde{R}_N\right| < \frac{1}{(2N+1)(2N+3)}$. Pick N so that $\left|\tilde{R}_N\right| < 10^{-3}$.

Chapter 2

Linear Vector Spaces: Algebraic Structure

2.1 Examples and Definition

In order to develop intuition for linear vector spaces, a slew of standard examples are needed. From them we shall abstract the needed properties which will then be stated as a set of axioms.

a) The Space $\mathbb{R}^2$.

We begin by informally examining a space of two dimensions (whatever that means). It is constructed by taking the Cartesian Product of $\mathbb{R}$ with itself. We are thus looking at $\mathbb{R}\times\mathbb{R}$, which is denoted by $\mathbb{R}^2$. A point X in this space is an ordered pair, $X = (x_1, x_2)$, where $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$. x_1 and x_2 are called the *coordinates* or *components* of the point x. Let us propose a reasonable algebraic structure on $\mathbb{R}\times\mathbb{R}$. If $X = (x_1, x_2)$, and $Y = (y_1, y_2)$ are any two points, and α is any real number, we *define*

addition: $X + Y = (x_1 + y_1, x_2 + y_2)$.
multiplication by scalars: $\alpha \cdot X = (\alpha x_1, \alpha x_2)$, $\alpha \in \mathbb{R}$.
equality: $X = Y \Longleftrightarrow x_1 = y_1$, $x_2 = y_2$

The addition formula states that the parallelogram rule is used to add points, whereas the second formula states that a point X is "stretched" by α by stretching each coordinate by α.

Some immediate consequences of the above definitions are, for all X, Y, Z in $\mathbb{R}\times\mathbb{R}$,

1. addition is *associative* $(X + Y) + Z = X + (Y + Z)$
2. addition is *commutative* $X + Y = Y + X$
3. There is an *additive identity*, $0 = (0, 0)$ with the property that $X + 0 = X$ for any X.

4. Every $X = (x_1, x_2) \in \mathbb{R} \times \mathbb{R}$ has an *additive inverse* $(-x_1, -x_2)$, which we denote by $-X$. Thus $X + (-X) = 0$. Thus the set of points in $\mathbb{R} \times \mathbb{R}$ forms an additive abelian group.

The following additional properties are also obvious, where α and β are arbitrary real numbers.

5. $\alpha(\beta X) = (\alpha\beta)X$
6. $1 \cdot X = X$.

And the two *distributive* laws:

7. $(\alpha + \beta)X = \alpha X + \beta X$
8. $\alpha(X + Y) = \alpha X + \alpha Y$.

To insure that you too feel these properties are obvious, let us prove, one, say 7.

$$\begin{aligned}(\alpha + \beta) \cdot X &= (\alpha + \beta) \cdot (x_1, x_2) = ((\alpha + \beta)x_1, (\alpha + \beta)x_2) \\ &= (\alpha x_1 + \beta x_1, \alpha x_2 + \beta x_2) = (\alpha x_1, \alpha x_2) + (\beta x_1, \beta x_2) \\ &= \alpha \cdot (x_1, x_2) + \beta \cdot (x_1, x_2) = \alpha \cdot X + \beta \cdot X\end{aligned} \tag{2-1}$$

EXAMPLE: If $X = (2, 1)$, then $3X = (6, 3)$ and $-2X = (-4, -2)$.

Instead of thinking of the elements (x_1, x_2) in $\mathbb{R}^2$ as points, it is sometimes useful to think of them as directed line segments, from the origin $(0, 0)$ directed to the point (x_1, x_2). The figure at the right illustrates this.

Note that the axes need not be perpendicular to each other in the space $\mathbb{R}^2$. They could just as well veer off at some outrageous angle, as in the diagram. This is because we have yet to place a metric (distance) structure on $\mathbb{R}^2$ or introduce any concept of angle measurement. When we do that, we will have Euclidean 2-space $\mathbf{E}^2$. But right now all we have is $\mathbb{R}^2$, which might be thought of as a floppy Euclidean space.

b) The Space $\mathbb{R}^n$

.

This is a simple-minded generalization of $\mathbb{R}^2$. A point X in $\mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$ is an ordered n tuple, $X = (x_1, x_2, \ldots, x_n)$ of real numbers, $x_k \in \mathbb{R}$. If $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ are any two points in $\mathbb{R}^n$, and α is any real number, we *define*

ADDITION: $\lambda + Y = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$

MULTIPLICATION BY SCALARS: $\alpha \cdot X = (\alpha x_1, \alpha x_2, \ldots, \alpha x_n),\ \alpha \in \mathbb{R}$.

EQUALITY: $X = Y \Longleftrightarrow x_j = y_j$ for all j.

EXAMPLE: The point $X = (1, 2, 3)$, and $\frac{1}{2}X = (\frac{1}{2}, 1, \frac{3}{2})$ in $\mathbb{R}^3$ are indicated in the figure. Again the coordinate axes need not be mutually perpendicular.

Properties 1-8 listed earlier remain valid - and with the proofs essentially unchanged (just add dots inside the parentheses).

REMARK: . At this stage, you probably are anxiously waiting for us to define multiplication in $\mathbb{R}^n$, that is, the product of two points in $\mathbb{R}^n$, $X \cdot Y = Z \in \mathbb{R}^n$, possibly using the multiplication of complex numbers (points in $\mathbb{R}^2$) as a guide. Well, we would if we could. It turns out that it is possible to define such a multiplication *only* in $\mathbb{R}^1$, $\mathbb{R}^2$, $\mathbb{R}^4$, and in $\mathbb{R}^8$ *–but in no others*. This is a famous theorem. In $\mathbb{R}^2$ ordinary complex multiplication does the job. To do it in $\mathbb{R}^4$, we have to abandon the commutative law for multiplication. The result is called *quaternions*. In $\mathbb{R}^8$, the multiplication is neither commutative nor associative. The result there is the *Cayley numbers*.

Here we shall not have time to treat this issue. All we shall do (later) is introduce a "pseudo multiplication" in $\mathbb{R}^3$—the so called cross product - obtained from the quaternion algebra in $\mathbb{R}^4$. The major importance of this pseudo multiplication which holds only in $\mathbb{R}^3$ is the fact of life that our world has three space dimensions. This multiplication is extremely valuable in physics.

c) The Space $C[a, b]$.

Our next example is of an entirely different nature, it is a space of functions, a *function space*. The space $C[a, b]$ is the set of all real-valued functions of a real variable x which are continuous for $x \in [a, b]$. If f and g are continuous for $x \in [a, b]$, that is if f and $g \in C[a, b]$, and if α is any real number, we define, in the usual way,

addition: $(f + g)(x) = f(x) + g(x)$,
multiplication by scalars: $(\alpha f)(x) = \alpha[f(x)]$. $\alpha \in \mathbb{R}$
equality: $f = g \Longleftrightarrow f(x) = g(x)$ for all $x \in [a, b]$.

Notice that the sum of two functions in $C[a, b]$ is again in $C[a, b]$, and the product of a continuous function - in $C[a, b]$—by a constant α is also an element of $C[a, b]$. We shall *ignore* the fact that the product of two continuous functions is also a continuous function.

Properties 1-8 listed earlier are also valid here, that is, if f, g, and h are any elements in $C[a, b]$, then

1. $f + (g + h) = (f + g) + h$
2. $f + g = g + f$
3. $f + 0 = f$
4. $f + (-1)f = 0$
5. $\alpha(\beta f) = (\alpha\beta)f$
6. $1 \cdot f = f$. Here $1 \in \mathbb{R}$.
7. $(\alpha + \beta)f = \alpha f + \beta f$

8. $\alpha(f+g) = \alpha f + \alpha g$.

Again, 1-4 state that the elements of $C[a,b]$ form an abelian group with the group operation being addition. When we define the dimension of a vector space, it will turn out that the space $C[a,b]$ is *infinite* dimensional, but don't let that bother you. This nice space, $C[a,b]$, and $\mathbb{R}^n$ are the two most useful examples of a vector space.

d) D. The Space $C^k[a,b]$.

The space $C^k[a,b]$ consists of all real-valued functions $f(x)$ which have k continuous derivatives for x in the interval $[a,b] \subset \mathbb{R}$. When $k=0$, this reduces to the space $C[a,b]$. Addition and scalar multiplication are defined just as in $C[a,b]$. The key property is that the sum of two functions with k continuous derivatives of $x \in [a,b]$ is also a function with k continuous derivatives. All of properties 1-8 are valid in $C^k[a,b]$.

Every function $f(x)$ which has one continuous derivative is necessarily continuous. This is a basic result from elementary calculus; it may be written as $C^1[a,b] \subset C[a,b]$. Since the function $|x|$, $x \in [-1,1]$ is in $C[-1,1]$ but not in $C^1[-1,1]$, we see that C^1 and C are not the same, that is C^1 is a proper subset of C. Similarly, $C^{k+1}[a,b] \subset C^k[a,b]$ (see Exercise 7).

The space $C^\infty[a,b]$ consists of all functions with an infinite number of continuous derivatives for $x \in [a,b]$. All functions which have a convergent Taylor series for $x \in [a,b]$ are in $C^\infty[a,b]$. In addition, $C^\infty[a,b]$ contains functions like $f(x) = e^{-1/x^2}$, $x \neq 0$, $f(0) = 0$, which have an infinite number of continuous derivatives (see p. ??) but do not have convergent Taylor series.

Another example of a function space is the set of *analytic functions* $A(z_0, R)$, functions which have a convergent Taylor series in the disc with center at $z_0 \in \mathbb{C}$ and radius at least R.

e) E. The Space ℓ_1.

The space ℓ_1 consists of all infinite sequences $X = (x_1, x_2, x_3, \ldots)$ which satisfy the condition $\sum_{n=1}^{\infty} |x_n| < \infty$. Addition and multiplication by scalars are defined in a natural way. If X and Y are in ℓ_1, then

$$X + Y + (x_1 + y_1, x_2 + y_2, x_3 + y_3, \ldots)$$

and, if α is any complex number

$$\alpha \cdot X = (\alpha x_1, \alpha x_2, \ldots).$$

Equality is defined by

$$X = Y \iff x_j = y_j \text{ for all } j.$$

We should show that if X and Y are in ℓ_1, then so is $X+Y$ and $x \cdot X$. To prove that $X + Y \in \ell_1$, we must show that $\sum |x_n + y_n| < \infty$. But since $|x_n + y_n| \leq |x_n| + |y_n|$, we have for any $N \in \mathbb{Z}_+$

$$\sum_{n=1}^{N} |x_n + y_n| \leq \sum_{n=1}^{N} |x_n| + \sum_{n=1}^{N} |y_n| \leq \sum_{n=1}^{\infty} |x_n| + \sum_{n=1}^{\infty} |y_n| < \infty.$$

Now letting $N \to \infty$ on the left, we see that $\sum_{n=1}^{\infty} |x_n + y_n| < \infty$. If $X \in \ell_1$, it is obvious that $\alpha \cdot X$ is also in ℓ_1 since

$$\sum_{n=1}^{\infty} |\alpha x_n| = \sum_{n=1}^{\infty} |\alpha|\, |x_n| = |\alpha| \sum_{n=1}^{\infty} |x_n| < \infty.$$

f) F. The Space $L_1[a,b]$.

Yes, the space $L_1[a,b]$ does consist of all functions $f(x)$ (possibly complex-valued) with the property that $\int_a^b |f(x)|\, dx < \infty$. It is the integral analogue of ℓ_1. Addition and scalar multiplication are defined as in $C[a,b]$, that is, as usual. If f and g are in $L_1[a,b]$, then so are $f+g$ and αf, where $\alpha \in \mathbb{C}$, since

$$\int_a^b |f(x) + g(x)|\, dx \leq \int_a^b |f(x)|\, dx + \int_a^b |g(x)|\, dx < \infty,$$

and

$$\int_a^b |\alpha f(x)|\, dx = |\alpha| \int_a^b |f(x)|\, dx < \infty.$$

For example, $f(x) = x$ is in $L_1[0,1]$ but $f(x) = \frac{1}{x^2}$ is *not* in $L_1[0,1]$. It is simple to check that properties 1-8 are satisfied in $L_1[a,b]$.

g) G. The Space $\mathcal{P}_n$.

If $p(x) = a_0 + a_1 x + \cdots + a_n x^n$ is any polynomial of degree atmost n with real coefficients and $q(x) = b_0 + b_1 x + \cdots + b_n x^n$ is another one, then with ordinary addition, multiplication by real scalars and equality the set $\mathcal{P}_n$ of all polynomials of degree n satisfy conditions 1-8. Since

$$a_0 + a_1 x + \cdots + a_{n-1} x^{n-1} = a_0 + a_1 x + \cdots + a_{n-1} x^{n-1} + 0 x^n,$$

it is clear that $\mathcal{P}_{n-1} \subset \mathcal{P}_n$.

Enough examples for now. You must have gotten the point. We shall meet more later on. Let us give the abstract definition of a linear vector space.
DEFINITION: . Let S be a set with elements $X, Y, Z, \ldots$ and F be a field with elements $\alpha, \beta, \ldots$. The set S is a *linear vector space*(*linear space, vector space*)*over the field* F if the following conditions are satisfied.

For any two elements $X, Y \in S$, there is a unique third element $X+Y \in S$, such that

1. $(X+Y)+Z = X+(Y+Z)$;
2. $X+Y = Y+X$;
3. There exists an element $0 \in S$ having the property that $0+X = X$ for all $X \in S$;
4. for every $X \in S$, there is an element $-X \in S$; such that $X+(-X) = 0$.

Furthermore, if α is any element of the field F, there is a unique element $\alpha X \in S$ such that, for any $\alpha, \beta \in F$,

5. $\alpha(\beta X) = (\alpha\beta)X$;
6. $1 \cdot X = X$.

The additive and field multiplicative structures are related by the following distributive rules:

7. $(\alpha+\beta)X = \alpha X + \beta X$
8. $\alpha(X+Y) = \alpha X + \alpha Y$.

Elements of the field F are called *scalars*, whereas elements of S are called *vectors*. We shall usually take the real numbers $\mathbb{R}$ for our field F, although the complex numbers $\mathbb{C}$ will be used at times, particularly when we discuss Fourier series. Exercise 4 shows the need for Axiom 6 (in case you thought it was superfluous).

All of the examples of this section are linear spaces. For most purposes the simple example $\mathbb{R}^2$ will serve you well as a guide to further expectations. The pictures there are simple. In fact, with a certain degree of cleverness, the "right" proof for $\mathbb{R}^2$ immediately generalizes to all other linear spaces - even "infinite dimensional" ones.

Since you probably think that everything is a linear space, here is an example to dispel the delusion. Let S be the subset of all functions $f(x)$ in $C[0,1]$ which have the property $f(0) = 1$. Then if f and g are in S, we are immediately stuck since $f(0)+g(0) = 2$, so that $f+g$ is *not* in S. Also, $0 \notin S$.

Both here, and before (p.?) when defining a field, axioms "0" have been used. They all express roughly the same concept. We have some set S and an operation * defined on the set. These axioms all stated that for any $x, y \in S$, we also have $x * y \in S$. In other words, the set S is *closed* under the operation * in the sense that performing that operation does not take us out of the set. We shall find this concept useful.

h) Appendix. Free Vectors

One more example is needed, an exceedingly important example. There are "physicists' vectors" or *free vectors*. I always thought they were easy to define - until today. Twelve hours and fifty pages later, I begin again on the fifth attempt. The essential idea is easy to imagine but difficult to convey in a clear and precise exposition.

Say you are given two elements X and Y of $\mathbb{R}^n$, which we represent by directed line segments from the origin. Somehow we want to find a directed line segment V *from* the tip of X *to* the tip of Y. Now V "looks" like a vector. The problem is that all of the vectors we have met so far have been directed line segments in $\mathbb{R}^n$ beginning at the origin.

In order to find a way out, it is best to examine the problem for the most simple case $-\mathbb{R}^1$, the ordinary line. Watch closely since we will be so shrewd that all the formalism will be adequate without change for the general case of $\mathbb{R}^n$.

We are given two points, X and Y of $\mathbb{R}^1$ which we shall represent by directed line segments from the origin. To make the picture clear, we will draw them slightly above the line.

A FIGURE GOES HERE

We want a directed line segment V from the tip of X to the tip of Y. Of course you recognize this as the problem of solving

$$X + V = Y$$

The solution, $V = Y - X$, is the difference of the two real numbers Y and X. But where should we draw V? If we are stubborn and demand that all real numbers must be represented by line segments beginning at the origin, we have the picture

A FIGURE GOES HERE

but what we really want to do is place the tail of V at the tip of X and add the line segments. Why not relent and allow ourselves this added flexibility.

A FIGURE GOES HERE

There! Now we have solved our problem. But we have made an important generalization in doing so. You see, this V has been released from its bondage to the origin and is now free to move along the whole of $\mathbb{R}$.

Although we were led to this V from the pair X and Y, the same V could have been generated by a different pair $\tilde{X}$ and $\tilde{Y}$, as the diagram below indicates,

A FIGURE GOES HERE

for we still have $\tilde{X}+V=\tilde{Y}$.

In the first case we might have had $X=2$ and $Y=3$, so that $V=1$, while in the second, we might have had $X=-4$ and $Y=-3$, and again $V=1$. Even though we have let this V go free, sliding from place to place along $\mathbb{R}$, we still want to say that this is only one V, and in fact, we want to identify this V with the V tied to the origin in (2). In other words, we would like to say that all three V's used above are equivalent to each other.

More formally, the element V is *generated* by an ordered pair, $V=[X,Y]$, which we read as the vector *from* X *to* Y, for $X,\ Y\in\mathbb{R}$. If some $\tilde{V}$ is generated by another ordered pair, $\tilde{V}=[\tilde{X},\tilde{Y}]$, $\tilde{X},\ \tilde{Y}\in\mathbb{R}$, then we want equality $V=\tilde{V}$ to mean that $\tilde{Y}-\tilde{X}=Y-X$. Moreover, we want to *represent* $V=[X,Y]$, the vector from X to Y, by the vector from the origin 0 to $Y-X$, $V=[0,Y-X]$. *This representation of* V *is unique*, since if any other pair also generates V, $V=[\tilde{X},\tilde{Y}]$, the representative $V=[0,\tilde{Y}-\tilde{X}]=[0,Y-X]$ since $V=V$ implies that $\tilde{Y}-\tilde{X}=Y-X$. Therefore much as each rational number is an equivalence class, represented by a single rational number - as $\frac{1}{2}$ represents the equivalence class $\frac{1}{2},\frac{2}{4},\frac{3}{6},\ldots$, each V is an equivalence class of ordered pairs $V=[X,Y]$, where $X,Y\in\mathbb{R}$. It is uniquely represented by an element of $\mathbb{R}$, viz. $V=Y-X$, the representation being independent of the particular ordered pair $[X,Y]$ which generates V. It is possible to think of V either as an ordered pair with an equivalence relation, or just as the representative $V=[0,Y-X]$ of the whole equivalence class, the representation being written more simply as an element of $\mathbb{R}: V=Y-X$, where here equality is between elements of $\mathbb{R}$.

The generalization is now easily made

DEFINITION: . (Free vectors). Let X and Y be any elements of $\mathbb{R}^n$. An element $V\in V^n$, "physicists' n-space", is defined as an equivalence class of ordered pairs of elements in $\mathbb{R}^n$,

$$V=[X,Y],\qquad X,\ Y\in\mathbb{R}^n,$$

with the following equivalence relation: If $V=[X,Y]$ and $\tilde{V}=[\tilde{X},\tilde{Y}]$, then

$$V=\tilde{V}\Longleftrightarrow\tilde{Y}-\tilde{X}=Y-X,$$

where the second equality is that of elements in $\mathbb{R}^n$. If we are given X and Y in $\mathbb{R}^n$, we speak of $V=[X,Y]$ as the free vector going *from* X *to* Y.

Previous reasoning also shows that *each* $V\in\mathcal{V}n$ *is uniquely represented by the ordered pair* $V=[0,Y-X]$. This representation is independent of the elements $[X,Y]$ which generated V.

We were led to this definition of $\mathcal{V}^n$ by examining the situation in the special case of $\mathcal{V}^1$. Since our formal reasoning there was quite algebraic and general, we know that the definition works algebraically. The geometry works too. An example in $\mathcal{V}^2$ should make the general case clear.

Let $X=(1,3)$ and $Y=(2,1)$. These two points in $\mathbb{R}^2$ generate the ordered pair $V=[(1,3),(2,1)]$ in $\mathcal{V}^2$. V is the vector going *from* $X=(1,3)$ *to* $Y=(2,1)$. Of all equivalent V's, the unique representative which begins at the origin is $V=[(0,0),(1,-2)]$, which we simply write as $V=(1,-2)$ and represent as an ordinary element of $\mathbb{R}^2$. On the same diagram we exhibit the vector from $\tilde{X}=(-2,2)$ to $\tilde{Y}=(-1,0)$, which is $\tilde{V}=[(-2,2),(-1,0)]$. The unique representative (of all $\tilde{V}$'s equivalent of $\tilde{V}$) which begins from (0,0) is $\tilde{V}=[(0,0),(1,-2)]$, which we write simply as $\tilde{V}=(1,-2)$. Comparison of V and $\tilde{V}$ reveals that they are equal, $V=\tilde{V}$. Thus, from the diagram, we see that a free vector is an equivalence class of directed line segments, with two directed line segments $V,\tilde{V}$ being equivalent as vectors in $\mathcal{V}^2$ if they are equivalent to the same directed line segment which begins at the origin. In more geometrical language, $V=\tilde{V}$ if by sliding them "parallel to themselves", they can be made to coincide with their representer which begins at the origin. (We shall not define "parallel" here. It is not needed because we already have a satisfactory algebraic definition of equivalence.)

Notice that $X=(1,3)$ and $Y=(2,1)$ also generates a second ordered pair $\hat{V}=[(2,1),(1,3)]$, the vector *from* $Y=(2,1)$ *to* $X=(1,3)$. Its unique representation which begins at the origin is $\hat{V}=[(0,0),(-1,2)]$, or more simply $\hat{V}=(-1,2)$. Comparison with the previous example shows that $\hat{V}=-V$: *the vector from* Y *to* X *is the negative of the vector from* X *to* Y. We need the little arrow on our picture of $V=[X,Y]$ to distinguish it from $-V=[Y,X]$ which is also between the same points but headed in the opposite direction.

From now on we shall denote a vector $V\in\mathcal{V}^n$ *from* X *to* Y by its representative $Y-X$ in $\mathbb{R}^n$, so $V=Y-X$. Hence the vector from (1,3) to (2,1) will be immediately written as $V=(1,-2)$. As we have said many times, the representation $V=Y-X$ as an element on $\mathbb{R}^n$ is independent of which particular pair $[X,Y]$ happened to generate V. The following diagram shows a whole bunch of equivalent vectors $V_j\in\mathcal{V}^2$,

A FIGURE GOES HERE

$V_j=V_k$, and their particular representative V chained to the origin.

In order to justify calling the elements of $\mathcal{V}^n$ vectors, we should prove that the *elements of* $\mathcal{V}^n$ *do form a vector space.* Addition and scalar multiplication must first be defined, an easy task. Since every $V\in\mathcal{V}^n$ is uniquely represented as an element of $\mathbb{R}^n$, $V=Y-X\in\mathbb{R}^n$, we use addition and scalar multiplication for elements of $\mathbb{R}^n$—which has already been defined. Because $\mathbb{R}^n$ is known to be a vector space, it is a tedious triviality to prove.

Theorem 2.1 . *$\mathcal{V}^n$ is a linear vector space.*

PROOF: . Only a smattering.

1. $\mathcal{V}^n$ is closed under addition. Say V_1 and V_2 are in $\mathcal{V}^n$. Then they are represented as the difference of two elements of $\mathbb{R}^n$, say $V_1=Y_1-X_1$ and $V_2=Y_2-X_2$. Thus

$$V_1+V_2=(Y_1-X_1)+(Y_2-X_2)=(Y_1+Y_2)-(X_1+X_2),$$

so that their sum is generated by $[X_1+X_2,Y_1+Y_2]$. In other words, there is at least one pair of elements, $[X_3,Y_3]$, $X_3=X_1+X_2$ and $Y_3=Y_1+Y_2$, in $\mathbb{R}^n$ which generate V_1+V_2, so that $V_3=V_1+V_2\in\mathcal{V}^n$. Of course $[0,Y_3-X_3]$ and many other pairs also generate V_3.

2. *Commutativity.*

$$V_1+V_2=(Y_1-X_1)+(Y_2-X_2)=(Y_2-X_2)+(Y_1-X_1)=V_2+V_1.$$

3. $(\alpha+\beta)V_1=(\alpha+\beta)(Y_1-X_1)=\alpha(Y_1-X_1)+\beta(Y_1-X_1)=\alpha V_1+\beta V_1$

EXAMPLE: If $A=(4,2,-3)$, $B=(0,1,-2)$, $C=(-1,0,\frac{1}{2})$ and $D=(4,-\frac{1}{2},1)$, find the vector V_1 from A to B and the vector from C to D. Then compute V_1+2V_2 and V_1-V_2.
SOLUTION: $V_1=B-A=(0,1,-2)-(4,2,-3)=(-4,-1,1)$

$$V_2=D-C=(4,-\frac{1}{2},1)-(-1,0,\frac{1}{2})=(5,-\frac{1}{2},\frac{1}{2})$$

$$V_1+2V_2=(-4,-1,1)+2(5,-\frac{1}{2},\frac{1}{2})=(-4,-1,1)+(10,-1,1)=(6,-2,2)$$

$$V_1-V_2=(-4,-1,1)-(5,-\frac{1}{2},\frac{1}{2})=(-4,-1,1)+(-5,\frac{1}{2},-\frac{1}{2})=(-9,-\frac{1}{2},\frac{1}{2})$$

Exercises

1. a) Find the vector representing the free vectors from the given $A\in\mathbb{R}^n$ to $B\in\mathbb{R}^n$.

i) $A=(3,1),\ B=(2,2)$	iv) $A=(0,0,0)\ B=(9,8,-3)$
ii) $A=(-3,3),\ B=(0,4)$	v) $A=(1,2,3),\ B=(0,0,-1)$
iii) $A=(2,2,3),\ B=(5,2,17)$	vi) $A=(0,0,-1),\ B=(1,2,3)$

   b) Let V_1 and V_2 be the respective vectors of iii) and v) above. Compute V_1+V_2, V_1-V_2, and $2V_1-3V_2$.

   c) Draw a diagram on which you indicate the vector going from $A=(3,1)$ to $B=(2,2)$, and indicate the representer of that vector which begins at the origin. Do the same with the vector from B to A.

2. Which of the following subsets of $C[-1,1]$ are linear spaces:
   a) The set of all even functions in $C[-1,1]$, that is, functions $f(x)$ with the additional property $f(-x)=f(x)$, like x^2 and $\cos x$.
   b) The set of all functions f in $C[-1,1]$ with the additional property that $|f(x)|\le 1$.
   c) The set of all functions f in $C[-1,1]$ with the property that $f(0)=0$.

3. In $\mathbb{R}^3$, let $X=(1,-1,2)$ and $Y=(0,4,-3)$. Find $X+2Y$, $Y-X$, and $7X-4Y$.

4. a) Show that for every $X\in\mathbb{R}^3$ you can find scalars $\alpha_j\in\mathbb{R}$ such that X can be written as
$$X=\alpha_1e_1+\alpha_2e_2+\alpha_3e_3,$$
where $e_1=(1,0,0),\ e_2=(0,1,0),\ e_3=(0,0,1)$.

   b) If $X\in\mathbb{R}^3$, can you find scalars $\alpha_j\in\mathbb{R}$ such that
$$X=\alpha_1v_1+\alpha_2v_2+\alpha_3v_3,$$
where $v_1=(1,-1,0)$, $v_2=(-1,1,0)$ and $v_3=(0,0,1)$, and $\alpha_j\in\mathbb{R}$? Proof or counter-example.

   c) Find two polynomials $p_1(x)$ and $p_2(x)$ in $\mathcal{P}_1$ such that for every polynomial $p(x)\in\mathcal{P}_1$ you can find scalars $\alpha_1,\ \alpha_2\in\mathbb{R}$ such that p can be written in the form
$$p(x)=\alpha_1p_1(x)+\alpha_2p_2(x).$$

5. Let $V=\mathbb{R}\times\mathbb{R}$ with the following definition of addition and scalar multiplication
$$X+Y=(x_1+x_2,\ y_1+y_2),\qquad \alpha X=(\alpha x_1,0),$$
$$0=(0,0),\qquad -X=(-x_1,-x_2).$$
Is V a vector space? Why?

6. Show that any field can be considered to be a vector space over itself.

7. Consider the set
$$S=\{\,u(x)\in C^2[0,1]\colon a_2(x)u''+a_1(x)u'+a_0(x)u=0\,\},$$
where the $a_j(x)\in C[0,1]$. Is S a linear space? Note that we do not yet know that S has any elements at all. The proof that S is not empty is the existence theorem for ordinary differential equations.

8. By integrating $|x|$ the "right" number of times, find a function which is in $C^1[-1,1]$ but is not in $C^2[-1,1]$.

2.2 Subspaces. Cosets.

With this section we begin the process of assigning names to the various concepts surrounding the idea of a linear vector space. This name calling will take us the balance of the chapter. Although the ideas are elementary and theorems simple, do not deceive yourselves into thinking this must be some grotesque joke that mathematicians have perpetrated. You see, we are in the process of building a machine. Most of its constituent parts are very easy to grasp. But when combined, the machine will be equipped successfully to assault a diversity of problems which appear off hand to be unrelated.

The value of this abstract formalism is that many seemingly distinct complicated specific problems are just one single problem in a variety of fancy dresses. By ignoring the extraneous paraphernalia we can concentrate on the essential issues.

A FIGURE GOES HERE

We begin by defining what is meant by a subspace of a vector space W. While reading the definition, think of a plane through the origin, which is a subspace of ordinary three dimensional space.

DEFINITION: . A set A is a *linear subspace* (linear variety, linear manifold) of the linear space W if i) A is a subset of W, and ii) A is also a linear space under the operations of vector addition and multiplication by scalars already defined on V.

EXAMPLES:

1. Let $A = \{ X \in \mathbb{R}^3 \colon X = (x_1, x_2, 0) \}$, that is, the points in $\mathbb{R}^3$ whose last coordinate is zero. Since $A \subset \mathbb{R}^3$, and a simple check shows that A is also a linear space, we see that A is a linear subspace of $\mathbb{R}^3$. Intuitively, this set A certainly "looks like" $\mathbb{R}^2$. You are right, and recall that the fancy word for this equivalence - of $\mathbb{R}^2 = (x_1, x_2)$ and the points in $\mathbb{R}^3$ of the form $(x_1, x_2, 0)$—is *isomorphic*. Similarly, the set $B = \{ X \in \mathbb{R}^3 \colon X = (x_1, 0, x_3) \}$ is also a subspace of $\mathbb{R}^3$. B is also isomorphic to $\mathbb{R}^2$.

2. Let $A = \{ X \in \mathbb{R}^n \colon X = (x_1, x_2, \ldots, x_k, 0, 0, \ldots, 0) \}$, that is, the points in A are those points in $\mathbb{R}^n$ whose last $n-k$ coordinates are zero. It is easy to see that A is a linear subspace of $\mathbb{R}^n$, and that A is isomorphic to $\mathbb{R}^k$.

3. Let $A = \{ f \in C[0,1] \colon f(0) = 0 \}$. A is a subset of the linear space $C[0,1]$, and is also a linear space (check this). Thus A is a linear subspace of $C[0,1]$.

4. Let $A = \{ f \in C[0,1] \colon f(0) = 1 \}$. A is a subset of $C[0,1]$, but it is *not a linear subspace* since - as we saw in the last section (p. ?)—A is itself not a linear space.

The following lemma supplies a convenient criterion for checking if a given subset A of a linear space W is a subspace.

Theorem 2.2 . *If A is a non-empty subset of the linear space W, then A is a linear subspace of $W \iff A$ is closed under addition of vectors in A and multiplication by all scalars.*

PROOF: . $\Rightarrow$. Since A is a subspace, it is itself a linear space. But all linear spaces are, by definition, closed under addition and multiplication by scalars.

$\Leftarrow$. Because A is a subset of W, and properties 1,2,5,6,7, and 8 hold in W, they also hold for the particular elements in W which happened to be in A. Notice that here we use the fact that A is closed under addition. Therefore only the existential axioms 3 and 4 need be checked. Since A is not empty, it contains at least one element, say $X \in A$. Because A is closed under multiplication by scalars we see that $0 = 0 \cdot X \in A$. Furthermore, for every $X \in A$, also $-X = (-1) \cdot X \in A$.

EXAMPLE: Let $A = \{ f \in C^1[0,1] \colon f'(0) = 0 \}$. Since A is a subset of the linear space $C^1[0,1]$, all we need show is that A is closed under addition and multiplication by scalars in order to prove A a linear subspace of $C^1[0,1]$. If $f, g \in A$, then $(f+g)'(0) = (f'+g')(0) = f'(0) + g'(0) = 0$, so $f + g \in A$. Also, for any $\alpha \in \mathbb{R}$, $(\alpha f)'(0) = \alpha(f')(0) = \alpha \cdot 0 = 0$, so $\alpha f \in A$.

Theorem 2.3 . *The intersection of two subspaces is also a subspace, but the union of two subspaces is* not necessarily *a subspace. More generally, the intersection of any collection of subspaces is also a subspace.*

PROOF: . Let A, B be subspaces of W. We show that $A \cap B$ is a subspace. Since $A \cap B \subset W$, all we need show is the closure properties of $A \cap B$. If $X, Y \in A \cap B$, then X and Y are both in A and B, so $X + Y \in A$ and $X + Y \in B \Rightarrow X + Y \in A \cap B$ too. Similarly for scalar multiples. The proof that $A \cap B \cap C \cap \cdots$ is a subspace is identical except for a notational mess.

For the second part of the theorem we merely exhibit an example of two subspaces A, B for which $A \cup B$ is not a subspace. In $\mathbb{R}^2$ let A be the linear subspace "horizontal axis", that is, $A = \{ X \in \mathbb{R}^2 \colon X = (x_1, 0) \}$, while B is "the vertical axis", $B = \{ X \in \mathbb{R}^2 \colon X = (0, x_2) \}$. Then $A \cup B$ is the "cross" of all points on either the horizontal axis or the vertical axis. This is not a linear space because points like $(1,0) \in A$, $(0,1) \in B$ do not have their sum $(1,0) + (0,1) = (1,1)$ in $A \cup B$. Precisely for this reason $\mathbb{R}^2 = \mathbb{R}^1 \times \mathbb{R}^1$ was constructed as the Cartesian product of $\mathbb{R}^1$ with itself; for if it had been constructed as $\mathbb{R}^1 \times \mathbb{R}^1$, then only the points situated on the axes themselves would get caught. More generally - and for the same reason - the Cartesian product is the process always used to "glue" together a larger space from several linear spaces. Only when $A \subset B$ (or $B \subset A$) is $A \cup B$ also a subspace (Exercise 4).

Your image of a linear space should be $\mathbb{R}^3$, and a subspace $\mathcal{S}$ is a plane or line in $\mathbb{R}^3$. Note that since every subspace must contain 0, these planes or lines *must pass through the origin.*

EXAMPLE: Let $\mathcal{S}_c = \{ X \in \mathbb{R}^2 \colon x_1 + 2x_2 = c, c \text{ real} \}$. Thus, the set $\mathcal{S}_c$ is all points $S = (s_1, s_2) \in \mathbb{R}^2$ on the straight line $s_1 + 2s_2 = c$. For what value(s) of c is $\mathcal{S}_c$ a subspace? If $\mathcal{S}_c$ is a subspace, then we must have $aS \in \mathcal{S}_c$ for all scalars a, that is $aS = (as_1, as_2) \in \mathcal{S}_c \Rightarrow as_1 + 2as_2 = c$. But for $a = 0$ this states that $c = 0$. Therefore the only possible subspace is $\mathcal{S}_0 = \{ X \in \mathbb{R}^2 \colon x_1 + 2x_2 = 0 \}$. It is easy to check that if S_1

and S_2 are in $\mathcal{S}_0$, then so are S_1+S_2 and aS_1. Thus $\mathcal{S}_0$ is a subspace. Similarly, every straight line through the origin is a subspace.

Our question now is, how can we talk about the other straight lines or planes which do not happen to pass through the origin? First we answer the question for our example above. There we have the linear space $\mathbb{R}^2$ and the subspace $\mathcal{S}_0$ which will be simply written as $\mathcal{S}$. $\mathcal{S}$ is a line through the origin. Let X_1 be any element in $\mathbb{R}^2$ (think of X_1 as a point). Then the set of all elements of $\mathbb{R}^2$ which can be written in the form $S+X_1$, where $S\in\mathcal{S}$, is the line "parallel" to $\mathcal{S}$ which passes through X_1. This line is written as $\mathcal{S}+X_1$. More explicitly, say $X_1=(1,\frac{3}{2})$. The set $\mathcal{S}+X_1$ is the set of all points $X=(x_1,x_2)\in\mathbb{R}^2$ of the form

$$X=S+X_1,\ \text{ which }\ S\in\mathcal{S},$$

or

$$(x_1,x_2)=(s_1,s_2)+(1,\tfrac{3}{2}),\quad \text{where}\quad s_1+2s_2=0.$$

Consequently $x_1=s_1+1$, and $x_2=s_2+\frac{3}{2}$. Using the relation $s_1+2s_2=0$, we find that $x_1+2x_2=4$— exactly the equation of the straight line through $X_1=(1,\frac{3}{2})$ and "parallel" to the subspace $\mathcal{S}$. This subset, $\mathcal{S}+X_1=\{\,X\in\mathbb{R}^2\colon X=S+X_1,\text{ where } S\in\mathcal{S}\,\}$, is called the X_1 *coset of* $\mathcal{S}$. Thus, *cosets* are the names given to "linear objects" which are not subspaces. They are subspaces translated to pass through X_1. You might prefer to call them *affine subspaces* instead of cosets.

Please observe that the cosets $\mathcal{S}+X_1$ and $\mathcal{S}+X_2$, where $X_1,X_2\in W$, are not necessarily distinct. In our example, these cosets coincide if and only if X_2 is on the line $\mathcal{S}+X_1$, that is, if $X_2\in\mathcal{S}+X_1$. The easiest way to test this is to see if $X_2-X_1\in\mathcal{S}$. Say $X_1=(1,\frac{3}{2})$ as before, and that $X_2=(2,1)$. Then the cosets $\mathcal{S}+X_1$ and $\mathcal{S}+X_2$ are the same since the point $X_2-X_1=(1,-\frac{1}{2})$ is in $\mathcal{S}$. It should be geometrically clear that the relation of equality among these cosets is an equivalence relation (and so deserving of the title "equality"). We shall state these ideas formally as we turn from this special - but characteristic - example to the general situation.

The general problem of describing lines or planes or "higher dimensional linear objects" which do not pass through the origin - so are not subspaces - is solved similarly.

DEFINITION: . Let W be a linear space, $\mathcal{S}$ a subspace of $\mathcal{V}$, and X_1 any element of W. All elements in W which can be written in the form $S+X_1$, where $S\in\mathcal{S}$, is called the X_1 *coset* of $\mathcal{S}$, and written as $\mathcal{S}+X_1$.

Our first theorem states that if X_2 is in the X_1 coset of $\mathcal{S}$, then X_1 is in the X_2 coset of $\mathcal{S}$:

Theorem 2.4 . $X_2\in\mathcal{S}+X_1 \iff X_1\in\mathcal{S}+X_2$.

PROOF: Since $X_2\in\mathcal{S}+X_1$, there is an $S\in\mathcal{S}$ such that $X_2=S+X_1$. Therefore $X_1=(-S)+X_2$. Because $\mathcal{S}$ is a linear space, $(-S)\in\mathcal{S}$. Thus X_1 has been written as the sum of X_2 and an element of $\mathcal{S}$, which means that $X_1\in\mathcal{S}+X_2$.

By the same argument, one sees that *any two cosets* $\mathcal{S}+X_1$ *and* $\mathcal{S}+X_2$ *are either identical or are disjoint* (have no element in common). Thus the cosets of $\mathcal{S}$ *partition* W in the sense that every element of W is in exactly one coset, just as for our example, every point in the plane $\mathbb{R}^2$ was in exactly one straight line parallel to the subspace determined by $x_1+2x_2=0$.

Although we were motivated by geometrical considerations, the ideas apply without alteration to any linear space. This is illustrated by again examining the set

$$A=\{\,f\in C[-1,1]\colon f(0)=1\,\},$$

which is not a subspace. It is a coset of a subspace $\mathcal{S}$ of $C[-1,1]$ which is constructed as follows. Consider the subspace $\mathcal{S}$ which is "naturally" associated with A, viz.

$$\mathcal{S}=\{\,g\in C[-1,1]\colon g(0)=0\,\}.$$

Then A is the coset $\mathcal{S}+1$, $A=\mathcal{S}+1$. This is true since clearly $A\supset\mathcal{S}+1$. Also $A\subset\mathcal{S}+1$ because for every $f\in A$,

$$f(x)=[f(x)-1]+1=g(x)+1,\ \text{ where }\ g\in\mathcal{S}.$$

Therefore $A=\mathcal{S}+1$. Similarly, we could have written A as $\mathcal{S}+\hat{f}$, where $\hat{f}$ is *any* function in A, for example $A=\mathcal{S}+\cos x$.

Exercises

1. Find which of the following subsets of $\mathbb{R}^n$ are subspaces.
   a) $\{\,X\in\mathbb{R}^n\colon x_1=0\,\}$,
   b) $\{\,X\in\mathbb{R}^n\colon x_1\ge 0\,\}$,
   c) $\{\,X\in\mathbb{R}^n\colon x_1-x_2=0\,\}$,
   d) $\{\,X\in\mathbb{R}^n\colon x_1-x_2=1\,\}$,
   e) $\{\,X\in\mathbb{R}^n\colon x_1^2-x_2=0\,\}$,

2. In $\mathcal{P}_3$, the linear space of all polynomials of degree at most 3, let $A=\{\,p(x)\in\mathcal{P}_3\colon p(0)=0\,\}$, and let $B=\{\,p(x)\in\mathcal{P}_3\colon p(1)=0\,\}$.
   a) Show that A and B are subspaces of $\mathcal{P}_3$.
   b) Find $A\cap B$ and $A\cup B$. Give an example which shows that $A\cup B$ is not a subspace of $\mathcal{P}_3$.

3. a) If X_1 and X_2 are given fixed vectors in $\mathbb{R}^2$, is
   $$A=\{\,X\in\mathbb{R}^2\colon X=a_1X_1+a_2X_2,\ a_1 \text{ and } a_2 \text{ any scalars }\}$$
   a subspace of $\mathbb{R}^2$?

b) Same as (a) but replace $\mathbb{R}^2$ by an arbitrary linear space W.

c) If $X_1, X_2, \ldots, X_k \in W$, is

$$A = \{ X \in W : X = \sum_1^k a_j X_j, \text{ for any scalars } a_j \},$$

a subspace of W?

4. Let A and B be subspaces of a linear space W. Prove that $A \cup B$ is also a subspace if and only if either $A \subset B$ or $B \subset A$, that is, if one of the subspaces contains the other.

5. Let $\mathcal{S}$ and $\mathcal{T}$ be subspaces of a linear space W, and suppose that A is a coset of $\mathcal{S}$ and B is a coset of $\mathcal{T}$. Prove that

   a) $A \subset B \Rightarrow \mathcal{S} \subset \mathcal{T}$.

   b) $A = B \Rightarrow \mathcal{S} = \mathcal{T}$.

6. a) Write the plane $2x_1 - 3x_2 + x_3 = 7$ as a coset of some suitable subspace $\mathcal{S} \subset \mathbb{R}^{36}$.

   b) Write the set $A = \{ f \in C[0,4] \colon f(0) = 1,\ f(1) = 3 \}$, as a coset of some suitable subspace $\mathcal{S} \subset C[0,4]$.

   c) Write the set $A = \{ f \in C^1[0,4] \colon f(1) = 1,\ f'(1) = 2 \}$ as a coset of some suitable subspace $\mathcal{S} \subset C^1[0,4]$.

2.3 Linear Dependence and Independence. Span.

If W is a linear space and $X_1, X_2, \ldots, X_k \in W$, then we know that, for any scalars a_j,

$$Y = \sum_{j=1}^k a_j X_j = a_1 X_1 + a_2 X_2 + \cdots + a_k X_k$$

is also in $\mathcal{V}$. Y is a *linear combination* of the X_j's. Now if 0 can be expressed as a linear combination of the X_j's, where at least one of the a_j's is not zero we expect that there is something degenerate around. In fact, if $0 = a_1 X_1 + \cdots + a_k X_k$ where say $a_1 \neq 0$, then we can solve for X_1 as a linear combination of $X_2, X_3, \ldots, X_k$,

$$X_1 = \frac{-1}{a_1}(a_2 X_2 + \cdots + a_{,} X_k).$$

This leads us to make a definition and state a theorem.
DEFINITION: . A finite set of elements $X_j \in W,\ j = 1, \ldots, k$ is called *linearly dependent* if there exists a set of scalars $a_j,\ j = 1, \ldots, k$, *not* all zero such that $0 = \sum_1^k a_j X_j$. If the X_j are not linearly dependent, we say they are *linearly independent.*

Theorem 2.5 . *A set of vectors $X_j \in W,\ j = 1, \ldots, k$ is linearly dependent if and only if at least one of the X_j's can be written as a linear combination of the other X_j's.*

To test if a given set of vectors is linearly independent, an equivalent form of Theorem 5 is useful.

Corollary 2.6 *A set of vectors $X_j \in W,\ j = 1, \ldots, k$ is linearly independent if and only if $\sum_{j=1}^k a_j X_j = 0$ implies that $a_1 = a_2 = \cdots = a_k = 0$.*

EXAMPLES:

1. The vectors $X_1 = (2,0)$, $X_2 = (0,1)$, $X_3 = (1,1)$ in $\mathbb{R}$ are linearly dependent since $0 = X_1 + 2X_2 - 2X_3$. Equivalently, we could have applied the theorem since X_3 can be written as a linear combination of X_1 and X_2

$$X_3 = \frac{1}{2} X_1 + X_2.$$

2. The functions $f_1(x) = e^x$, $f_2(x) = e^{-x}$, $f_3(x) = \frac{e^x + e^{-x}}{2}$ in $C[0,1]$ are linearly dependent since

$$0 = f_1 + f_2 - 2f_3$$

3. The vectors $X_1 = (2,0,1)$, $X_2 = (-1,0,0)$ in $\mathbb{R}^3$ are linearly independent, since if for some a_1, a_2,

$$0 = a_1 X_1 + a + 2X_2 = (2a_1, 0, a_1) + (-a_2, 0, 0),$$

then

$$0 = (0,0,0) = (2a_1 - a_2, 0, a_1),$$

which implies that $2a_1 - a_2 = 0$, and $a_1 = 0 \Longrightarrow a_1 = a_2 = 0$.

A FIGURE GOES HERE

A simple consequence of these ideas is the following

Theorem 2.7 . *If A and B are any subsets of the linear space W and if $A \subset B$, then i) A is linearly dependent $\Rightarrow$ B is linearly dependent; and the contrapositive: ii) B is linearly independent $\Rightarrow$ A is linearly independent.*

We now prove the transitivity of linear dependence.

Theorem 2.8 . *If Z is linearly dependent on the set $\{ Y_j \}$, $j = 1, \ldots, n$ and each Y_j is linearly dependent on the set $\{ X_l \}$, $l = 1, \ldots, m$ then Z is linearly dependent on the $\{ X_l \}$.*

Proof: . This is trivial arithmetic. We know that

$$Z = a_1Y_1 + \ldots + a_nY_n \quad \text{and that} \quad Y_j = c_{j1}X_1 + c_{j2}X_2 + \cdots + c_{jm}X_m$$

By substitution then

$$\begin{aligned} Z &= a_1(c_{11}X_1 + \cdots + c_{1m}X_m) + a_2(c_{21}X_1 + \cdots + c_{2m}X_m) \\ &\quad + \cdots + a_n(c_{n2}X_1 + \cdots + c_{nm}X_m) \\ &= (a_1c_{11} + a_2c_{21} + \cdots + a_nc_{n1})X_1 + (a_1c_{12} + \cdots + a_nc_{n2})X_2 \\ &\quad + \cdots + (a_1c_{1m} + \cdots + c_{nm})X_m \\ &= \gamma_1X_1 + \cdots + \gamma_mX_m, \text{ where } \gamma_\ell = \sum_{j=1}^{n} a_jc_{j\ell}. \end{aligned}$$

More concisely:

$$Z = \sum_{j=1}^{n} a_jY_j = \sum_{j=1}^{n} a_j\left(\sum_{\ell=1}^{m} c_{j\ell}X_\ell\right) = \sum_{\ell=1}^{m}\left(\sum_{j=1}^{n} a_jc_{j\ell}\right)X_\ell = \sum_{\ell=1}^{m}\gamma_l X_\ell.$$

Let X_1 and X_2 be any elements of a linear space W. Is there a smallest subspace A of W which contains X_1 and X_2? There are two possible ways of answering this, constructively and non-constructively.

First, constructively. We observe that the desired subspace must contain X_1 and X_2, and all linear combinations of X_1 and X_2, that is, A must contain all $X \in W$ of the form $X = a_1X_1 + a_2X_2$ for all scalars a_1 and a_2. But observe that the set $B = \{ X \in \mathbf{V} \colon X = a_1X_1 + a_2X_2 \}$ is a linear space, since if X and $Y \in B$, then $aX \in B$ for any scalar a, and also $X + Y \in B$. Thus the desired subspace A is just B itself.

The constructive proof goes as follows: just let A be the intersection of all subspaces containing X_1 and X_2. By Theorem 2.3 the intersection of these subspaces is also a subspace. It is clearly the smallest one. Do you feel cheated? This type of reasoning is often used in modern mathematics. Although it reveals little more than the existence of the sought-after object, it is an extremely valuable procedure when you really don't want anything more than to know the object exists. More important, procedures like this are vital when there is no constructive proof available.

More generally, if $S = \{ X_j \}$, $j = 1, \ldots, k$, is any *finite* subset of a linear space W, we ask for the smallest subspace A of W which contains S. There are two proofs - exactly as in the simple case above (where $k = 2$). From the constructive proof we find that

$$A = \{ X \in W \colon X = \sum_{1}^{k} a_jX_j,\ a_j \text{ scalars} \},$$

so A *is the set of all linear combinations of the* X_j's. This set A is called the *span* of S, and denoted by $A = \operatorname{span}(S)$. We also say that S *spans* A, or that A is *generated* by S.

Examples:

1. In $\mathbb{R}^3$ let $X_1 = (1, 0, 0)$ and $S_2 = (0, 1, 0)$. Then the span of $S = \{ X_j,\ j = 1, 2 \}$ is all $X \in \mathbb{R}^3$ of the form $X = a_1X_1 + a_2X_2 = (a_1, a_2, 0)$. If we imagine $\mathbb{R}^3$ as ordinary 3-space, then the span of X_1 and X_2 is the entire x_1, x_2 plane.

2. In $\mathbb{R}^3$, let $X_1 = (1, 0, 0)$, $X_2 = (0, 1, 0)$, and $X_3 = (0, 0, 1)$. Then the span of $T = \{ X_j,\ j = 1, 2, 3 \}$ is all $X \in \mathbb{R}^3$ of the form $X = a_1X_1 + a_2X_2 + a_3X_3(a_1, a_2, a_3)$. Since all of $\mathbb{R}^3$ can be so represented, we have $\operatorname{span}(T) = \mathbb{R}^3$, that is, the set B spans $\mathbb{R}^3$. Comparing these two examples, we see that $S \subset T$ and $\operatorname{span}(S) \subset \operatorname{span}(T)$.

3. In $\mathbb{R}^3$, let $X_1 = (1, 0)$ and $X_2 = (0, 1)$. Then the span of $S = \{ X_1, X_2 \}$ is all of $\mathbb{R}^2$, since every $X \in \mathbb{R}^2$ can be written as $X = a_1X_1 + a_2X_2$, where a_1 and a_2 are scalars. Many other sets also span $\mathbb{R}^2$. In fact almost every set of two vectors X_1 and X_2 in $\mathbb{R}^2$ span $\mathbb{R}^2$. This can be seen from the diagram, where we have drawn a net parallel to X_1 and X_2. Then $X = a_1X_1 + a_2X_2$. Any vectors X_1 and X_2 would do equally well, as long as they do *not* point in the same (or opposite) direction.

   We collect some properties of the span

Theorem 2.9 . *Let R, S, and T be subsets of a linear space W. Then*

(a) $R \subset \operatorname{span}(R)$.

(b) $R \subset S \Longrightarrow \operatorname{span}(R) \subset \operatorname{span}(S)$.

(c) $R \subset \operatorname{span}(S)$ *and* $S \subset \operatorname{span}(T) \Longrightarrow R \subset \operatorname{span}(T)$.

(d) $S \subset \operatorname{span}(T) \Longrightarrow \operatorname{span}(S) \subset \operatorname{span}(T)$.

(e) $\operatorname{span}(\operatorname{span}(T)) = \operatorname{span}(T)$.

(f) *A vector $X_j \in S$ is linearly dependent on the other elements of S* $\Longleftrightarrow \operatorname{span}(S) = \operatorname{span}(S - \{ X_j \})$. *(Here $S - \{ X_j \}$ means the set A with the one vector X_j deleted).*

Proof: These all depend on the representation of $\operatorname{span}(S)$ as a linear combination of the elements of S.

a) and b) —Obvious. They really should be if you understand the definitions.

c) A direct translation of Theorem 2.8.

d) This is the special case $R = \operatorname{span}(S)$ of part c.

e) By part (a) $\operatorname{span}(\operatorname{span}(T)) \supset \operatorname{span}(T)$. The opposite inclusion $\operatorname{span}(\operatorname{span}(T)) \subset \operatorname{span}(T)$ is the special case $S = \operatorname{span}(T)$ of part (d).

f) X_j linearly dependent on $S - \{ X_j \} \Longrightarrow S \subset \operatorname{span}(S - \{ X_j \})$. Thus by part (d), $\operatorname{span}(S) \subset \operatorname{span}(S - \{ X_j \})$. Inclusion in the opposite direction $\operatorname{span}(S - \{ X_j \}) \subset \operatorname{span}(S)$

follows from part (b). Therefore $\operatorname{span}(S) = \operatorname{span}(S - \{X_j\})$ means that $X_j \in \operatorname{span}(S)$ can be expressed as a linear combination $S - \{X_j\}$, i.e., the other X_k's.

Now most likely this proof was your first taste of abstract juggling and you find it difficult. Relax and don't be impressed with how formidable it appears. Except for parts a and b, the whole business hinges on the explicit construction of Theorem 2.8. Since (d) is a special case of (c), a good exercise is to write out the proof of (d) without relying on (c).

In $\mathbb{R}^2$, let $X_1 = (1,0)$, $X_2 = (0,1)$, and X_3 be any vector in $\mathbb{R}^2$. Observe that X_1 and X_2 together span $\mathbb{R}^2$. Thus X_3 can be expressed as a linear combination of X_1 and X_2, so that X_1, X_2, and X_3 are linearly dependent. The next theorem is a generalization of this idea.

Theorem 2.10 . *If a finite set $A = \{X_j,\ j = 1,\dots,n\}$ spans a linear space W, then every set $\tilde{S} = \{Y_j \in V,\ j = 1,\dots,m > n\}$ with more than n elements is linearly dependent. In other words, every linearly independent set has at most n elements.*

PROOF: Pick any $n+1$ elements $Y_1,\dots Y_{n+1}$ from $\tilde{S}$ and throw the rest away. Call the new set S. We shall show that these $n+1$ elements are linearly dependent. Then, since $S \subset \tilde{S}$, Theorem 2.7 tells us that $\tilde{S}$ is also linearly dependent. The only problem is how to carry out the proof without getting involved in a mess of algebra. By the principle of conservation of effort, this means that there will be some fancy footwork.

Reasoning by contradiction, assume S is linearly independent. If we can show that $\operatorname{span}(A) = \operatorname{span}(S - \{Y_{n+1}\})$, then $\operatorname{span}(S) \subset \operatorname{span}(S - \{Y_{n+1}\})$ because $\operatorname{span}(S) \subset V = \operatorname{span}(A) = \operatorname{span}(S - \{Y_{n+1}\}))$. Since $\operatorname{span}(S - \{Y_{n+1}\}) \subset \operatorname{span}(S)$, we can apply part f of Theorem 2.7 to conclude that S is linearly dependent - the desired contradiction.

Thus, assuming $S = \{Y_1,\dots,Y_{n+1}\}$ is linearly independent, we are done if we prove that $\operatorname{span}(A) = \operatorname{span}(S - \{Y_{n+1}\})$. Consider the set $B_k = \{Y_1,\dots,Y_k,\ X_{k+1},\dots,X_n\}$. We know that $B_0 = A$, so that $\operatorname{span}(B_0) = \operatorname{span}(A) = W$. Then by induction we shall prove that $\operatorname{span}(B_k) = W \Longrightarrow \operatorname{span}(B_{k+1}) = W$. Since $\operatorname{span}(B_k)$ spans W, then Y_{k+1} is a linear combination of the elements of B_k. Because the Y's are assumed linearly independent, this linear combination must involve at least one of $X_{k+1},\dots,X_n$. Say it involves X_{k+1} (if not, relabel the X's to make it so). Then we can solve for X_{k+1} as a linear combination of $\operatorname{span}(B_{k+1})$. Therefore $W = \operatorname{span}(B_k) = \operatorname{span}(B_{k+1})$. Putting this part together, we find that $\operatorname{span}(A) = W = \operatorname{span}(B_0) = \operatorname{span}(B_1) = \dots = \operatorname{span}(B_n)$. But $B_n = S - \{Y_{n+1}\}$. Thus $\operatorname{span}(A) = \operatorname{span}(S - \{Y_{n+1}\})$, and the proof is completed.

Example. In $\mathbb{R}^2$, any three (or more) non-zero vectors are linearly dependent since the two vectors $X_1 = (1,0)$ and $X_2 = (0,1)$ span $\mathbb{R}^2$.

Exercises

1. Determine which of the following are linearly dependent.
   a) In $\mathcal{P}_2$ $\quad p_1(x) = 1$, $p_2(x) = 1 + x$, $\quad p_3(x) = x - x^2$
   b) In $\mathbb{R}^3$, $\quad X_1 = (0,1,1)$, $X_2 = (0,0,-1)$, $X_3 = (0,2,3)$.

   c) In $C[0,\pi]$, $f(x) = \sin x$, $g(x) = \cos x$.
   d) In $\mathbb{R}^n$, $e_1 = (1,0,0,\dots,0)$, $e_2 = (0,1,0,0),\dots,e_n = (0,0,\dots,0,1)$.

2. Use the result of (d) to show that any set of $n+1$ vectors in $\mathbb{R}^n$ must be linearly dependent.

3. a) Find a set that spans i) $\mathbb{R}^4$, ii) $\mathcal{P}_3$.
   b) Show that no finite set spans ℓ_1.

4. Let $X_1,\dots,X_k$ be any elements of a linear space $\mathcal{V}$.
   a) Prove that $\operatorname{span}(\{X_1,\dots,X_k\}) = \operatorname{span}(\{X_1 + aX_j, X_2,\dots,X_k\})$, where a is any scalar and X_j is any of the $X_2, X_3,\dots,X_k,$.
   b) Prove that $\operatorname{span}(\{X_1,\dots,X_k\}) = \operatorname{span}(\{aX_1, X_2,\dots,X_k\}), a \neq 0$.
   c) In $\mathbb{R}^n$, consider the *ordered* set of vectors $\{X_1, X_2,\dots,X_k\}$, where
   $$X_j = (x_{1j}, x_{2j},\dots,x_{nj}).$$
   They are said to be in *echelon form* if i) no X_j is zero, and ii) the *index* of the first non-zero entry in X_j is less than the index of the first non-zero entry in X_{j+1}, for each $j = 1,\dots,k-1$. Thus $X_1 = (0,1,0)$, $X_2 = (0,0,1)$ are in echelon form while $X_1 = (0,1,0)$, $X_2 = (1,0,1)$ are not in echelon form. Prove that any set of vectors in echelon form is always linearly independent. (I suggest a proof by induction).

5. For what real value(s) of the scalar α are the vectors $(\alpha,1,0)$, $(1,\alpha,1)$, and $(0,1,\alpha)$ in $\mathbb{R}^3$ linearly dependent?

6. a) In $\mathbb{R}^3$, let $X_1 = (3,-1,2)$. Express $(-6,2,-4)$ linearly in terms of X_1. Show that $(3,4,-7)$ cannot be expressed linearly in terms of X_1. Can $(1,2,1)$ be expressed linearly in terms of X_1?
   b) In $\mathbb{R}^3$, let $A = \{X_1, X_2\}$, where $X_1 = (1,3,-2)$ and $X_2 = (2,1,1)$. Express $(3,-1,4)$ linearly in terms of A. Show that $(0,0,2)$ cannot be expressed linearly in terms of A. Can $(0,5,-5)$ be expressed linearly in terms of A?

7. a) In $C[0,10]$, let $f_1,\dots,f_8$ be defined by
   $$\begin{array}{ll} f_1(x) = x^2 - x + 2, & f_5(x) = x^3 \\ f_2(x) = (x+1)^2 & f_6(x) = \sin x \\ f_3(x) = x + 3 & f_7(x) = \cos x \\ f_4(x) = 1 & f_8(x) = \sin(x + \pi/4). \end{array}$$

Let $S = \{f_1, f_2, f_3\}$. Express f_4 linearly in terms of S. Show that f_5 cannot be expressed linearly in terms of S. Is $f_6 \in \text{span}(S)$? Is $f_8 \in \text{span}(f_6, f_7)$? Is $f_6 \in \text{span}(f_5, f_7, f_8)$?

b) If we let $f_9(x) = (x-1)^3$, $f_{10}(x) = 2x - 1$, determine which of the following sets are linearly dependent:

$$A := \{f_1, f_3, f_{10}\}, \quad B := \{f_1, f_5, f_9\}, \quad C := \{f_3, f_4, f_{10}\}, \quad D := \{f_1, f_4, f_5, f_9\},$$

2.4 Bases and Dimension

If the set $\{X_1, \ldots, X_m\}$ spans the linear space W, is there any set with less than m vectors which also spans W? There certainly is if the $\{X_1, \ldots X_m\}$ are linearly dependent, for if say X_m depends linearly upon the $\{X_1, \ldots, X_{m-1}\}$, then by Theorem 2.9(f), $\text{span}(\{X_1, \ldots, X_m\}) = \text{span}(\{X_1, \ldots, X_m\}) = W$, so then $\{X_1, \ldots, X_{m-1}\}$ span W. We can continue and eliminate the extra linearly dependent elements until we obtain a set $\{X_1, \ldots, X_n\}$ of linearly independent vectors which still span W.

Definition. A set of vectors $X_j \in W, j = 1, \ldots, n$ which is i) linearly independent, and ii) spans W is called a *basis* for W.

Examples

1. In $\mathbb{R}^2$, the vectors $X_1 = (1, 0)$ and $X_2 = (0, 1)$ are linearly independent and span $\mathbb{R}^2$. Therefore X_1 and X_2 form a basis for $\mathbb{R}^2$. The vectors $X_3 = (3, -1)$ and $X_4 = (-2, 2)$ in $\mathbb{R}^2$ are also linearly independent and span $\mathbb{R}^2$. They thus constitute another basis for $\mathbb{R}^2$. Almost any two vectors in $\mathbb{R}^2$ span $\mathbb{R}^2$, as long as they do not point on the same or opposite direction.

2. In $\mathcal{P}_2$, the polynomials $p_1(x) = 1$, and $p_2(x) = x - x^2$ do *not* form a basis. They are linearly independent but do not span the space - since for example you can never obtain the polynomial $p(x) = x$ which is in $\mathcal{P}_2$. If we add the third polynomial, say $p_3(x) = x - 2x^2$, then p_1, p_2 and p_3 do form a basis for $\mathcal{P}_2$.

Bases have an important property.

Theorem 2.11 . *If $\{X_1, \ldots, X_n\}$ form a basis for the linear space W, then every $X \in W$ can be expressed uniquely as a linear combination of the X_j's.*

Remark: Every set which spans W has, by definition, the property that every $X \in W$ can be expressed as a linear combination of the X_j's. The point here is that for a basis, this linear combination is uniquely determined.

PROOF: Suppose that $X = \displaystyle\sum_1^n a_k X_k$ and also $X = \displaystyle\sum_1^n b_k X_k$. We must show that $a_k = b_k$ for all k. Subtracting the two equations we find that $0 = \displaystyle\sum_1^n c_k X_k$, where $c_k = a_k - b_k$. But

since the X_k's are linearly independent, by Corollary 2.6, the only way a linear combination can be zero is if $c_k = 0$, $k = 1, \ldots, n$, that is, $a_k = b_k$ for all k.

We have observed that a linear space may have several different bases. Is it possible that different bases contain a different number of elements? Our next theorem states that the answer is NO.

Theorem 2.12 . *If a linear space W has one basis with a finite number of elements, say n, then all other bases are finite and also have exactly n elements.*

PROOF: We invoke Theorem 2.10. Let A be a basis with n elements and B be a basis with m elements. Now A spans W and the elements of B are linearly independent, so the Theorem 2.10, $m \leq n$. Reversing the roles of A and B we find that $n \leq m$. Therefore $n = m$.

With this result behind us, we can now define the dimension of a linear space.

Definition. If a linear space W has a basis with n elements, then we say that the *dimension of W is n*. If a linear space W has the property that no finite set of elements spans it, we say it is *infinite dimensional.*

Remarks. Theorem 2.12 states that the dimension of W is independent of which basis we happened to pick. If we want to emphasize the dimension of a finite dimensional space, we will write W^n.

Announcement. The dimension of $\mathbb{R}^n$ is n, for the n elements $e_1 = (1, 0, 0, \ldots, 0)$, $e_2 = (0, 1, 0, \ldots, 0), \ldots,$ $e_n = (0, \ldots, 0, 1)$ are linearly independent and span $\mathbb{R}^n$.

A picture. We have seen that $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, $e_3 = (0, 0, 1)$ form a basis in $\mathbb{R}^3$. Thus every $X \in \mathbb{R}^3$ can be expressed uniquely as a linear combination of the e_j's, $X = a_1 e_1 + a_2 e_2 + a_3 e_3$. If we represent e_1 as a directed line segment from the origin to $(1, 0, 0)$, and similarly for e_2 and e_3, then X is the geometrical sum of $a_1 e_1 + a_2 e_2 + a_3 e_3$, and is represented as a directed line segment from the origin to (a_1, a_2, a_3). In $\mathbb{R}^3$, e_1 is usually written as $\mathbf{i}$, e_2 as $\mathbf{j}$ and e_3 as $\mathbf{k}$, so that a vector $X \in \mathbb{R}^3$ is written as $X = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$.

The points in the plane $x_3 = 0$, which is isomorphic to $\mathbb{R}^2$, are then represented as $X = a_1\hat{\imath} + a_2\hat{\jmath} + 0\hat{k} = a_1\hat{\imath} + a_2\hat{\jmath}$. We would retain this notation except that one runs out of letters when considering spaces of higher dimension. For that reason the subscript notation $e_1, e_2, \ldots$ is better suited to our purposes.

Example

We show that the linear space $C[0, 1]$ of functions continuous in the interval $[0, 1]$ is infinite dimensional. This will be done by proving that the functions $f_0(x) = 1$, $f_1(x) = e^x$, $f_2(x) = e^{2x}, \ldots, f_n(x) = e^{nx}, \ldots$ are linearly independent. Say $0 = \sum_{k=0}^N a_k e^{kx}$, where N is some non-negative integer. We must show that all the a_k's are zero.

The trick is to use induction. For $N = 0$, we know that $0 = a_0$ only if $a_0 = 0$. Suppose $1, e^x, e^{2x}, \ldots, e^{(N-1)x}$ are linearly independent. Then $\sum_{k=0}^{N-1} a_k e^{kx} = 0$ if and only if all of the a_k's are zero. Let us show that this implies that $\sum_{k=0}^{N} a_k e^{kx} = 0$ if and only if all the a_k's vanish. Take the derivative. The constant term drops out and we are left with

$$0 = a_1 e^x + 2a_2 e^{2x} + \cdots + N a_N e^{Nx}.$$

Factor out e^x
$$0 = e^x(a_1 + 2a_2e^x + \cdots + Na_Ne^{(N-1)x}).$$
Since e^x is never zero, we know that
$$0 = a_1 + 2a_2e^x + \cdots + Na_Ne^{(N-1)x}.$$
By our induction hypothesis, this linear combination of $!, e^x, \ldots, e^{(N-1)x}$ can be zero if and only if $a_1 = a_2 = a_3 = \cdots = a_N = 0$. It remains to show that $a_0 = 0$. This is an immediate consequence of $\sum_{k=0}^N a_ke^{kx} = 0$ and the vanishing of a_k, for $k \geq 1$.

Since the functions $1, e^x, e^{2x}, \ldots$, are in $C^k[a,b]$ for any k we have shown that these spaces are infinite dimensional too. Moreover, the exact same proof also shows that the set $\{e^{\alpha_1 x}, e^{\alpha_2 x}, \ldots, e^{\alpha_N x}\}$, where $\alpha_1, \ldots, \alpha_N$ are arbitrary distinct *complex* numbers, is linearly independent. This fact will be needed later. Perhaps we shall present a different proof - or several different ones - at that time. All of the other proofs still involve some calculus - but that should be no surprise since we used calculus to define the exponential function in the first place.

Not all spaces of functions are infinite dimensional. For example, the function space $A = \{f \in C[-1,1]\colon f(x) = a + be^x,\ a,\ b \in \mathbb{R}\}$ has dimension 2. The functions $f_1(x) = 1$ and $f_2(x) = e^x$ constitute a basis for A because every $f \in A$ can be written in the form $f = a_1f_1 + a_2f_2$, where a_1 and a_2 are real numbers. Another basis for A is $f_3(x) = 1 + e^x$ and $f_4 = 2 - e^x$. There are many ways to see this. One is to observe that $f_3 + f_4 = 3$ and $2f_3 - f_4 = 3e^x$. Thus if $f(x) = a + be^x \in A$, then
$$f = \frac{a}{3}(f_3 + f_4) + \frac{b}{3}(2f_3 - f_4) = (\frac{a}{3} + \frac{2b}{3})f_3 + (\frac{a}{3} - \frac{b}{3})f_4.$$

The function space $B = \{f \in C[-1,1]\colon f(x) = a\sin(x+\alpha),\ \alpha, a \in \mathbb{R}\}$, also has dimension two, since $f(x) = (a\cos\alpha)\sin x + (a\sin\alpha)\cos x = a_1\sin x + a_2\cos x$. Thus $f_1(x) = \sin x$ and $f_2(x) = \cos x$ form a basis. Actually, we have only shown that f_1 and f_2 span B, but not that they are linearly independent. You can settle that point yourselves.

A few more remarks should be added. If A is a subspace of an n dimensional space W^n, we would like to enlarge a basis $\{e_1, \ldots, e_k\}$ for A to a larger basis $\{e_1, \ldots, e_n\}$ for all of W. Since $A \subset W^n$, it is clear that $k = \dim A \leq n$. If $A = W^n$, we are done since $\{e_1, \ldots, e_k\}$ already span W^n. Otherwise there is some element e_{k+1} in W^n which is not in A. Let $A_1 = \operatorname{span}\{e_1, \ldots, e_{k+1}\} \subset W^n$. If $A_1 = W^n$, then $\{e_1, \ldots, e_{k+1}\}$ form a basis for W^n. Otherwise there is some element e_{k+2} in W^n which is not in A_1. Form $A_2 = sp\{e_1, \ldots, e_{m+2}\}$. Repeat this process until you finally get a basis for all W^n. Only a finite number of steps are needed since the dimension of W^n is finite. This proves

Theorem 2.13 . *If A is a subspace of (finite dimensional) space W, then any basis for A can be extended to a basis that spans all of W.*

Consider a subspace A of a linear space W. Somehow we would like to discuss - and give a name to - the part A' of V which is not in A. We would like A' to be a subspace of V such that the only element of V which A and A' share is 0, and such that every element in V can be written as the sum of an element in A and an element in A'.

Definition: . Let A be a subspace of the linear space V. A *complementary subspace* A' of A is a subset of V with the properties

1. A' is a subspace of V,

2. If $X \in V$, then $X = X_1 + X_2$, where $X_1 \in A$ and $X_2 \in A'$.

3. $A \cap A' = 0$. (The zero vector, *not* the empty set).

An immediate task is to prove

Theorem 2.14 . *Every subspace $A \subset V$ has at least one complement, A'.*

Proof: Let $\{e_1, \ldots, e_m\}$ be a basis for A, and $\{e_1, \ldots e_m, e_{m+1}, \ldots, e_n\}$ an extension to a basis for V. We shall verify that $A' = sp\{e_{m+1}, \ldots, e_n\}$ satisfies both criteria. Now if $X \in A$ and $X \in A'$, then we can write $X = a_1e_1 + \ldots + a_me_m \in A$, and $X = a_{m+1}e_{e+1} + \ldots + a_ne_n \in A'$. Subtracting these equations, we find
$$0 = a_1e_1 + \ldots + a_me_m - a_{m+1}e_{m+1} - \ldots - a_ne_n.$$
But since $\{e_1, \ldots, e_n\}$ is a basis for V, the elements are linearly independent. Thus $a_1 = a_2 = \ldots = a_m = a_{m+1} = \ldots = a_n = 0$, so $X = 0$. Therefore $A \cap A' = 0$.

Furthermore, if $X \in V$ since $\{e_1, \ldots, e_n\}$ is a basis for V, then
$$X = \sum_{j=1}^n c_je_j = \sum_{j=1}^m c_je_j + \sum_{j=m+1}^n c_je_j.$$
Thus we just let $X_1 = c_1e_1 + \ldots + c_me_m \in A$ and X_2 and $X_2 = c_{m+1}e_{m+1} + \ldots + c_ne_n$.

It is easy to see that the above construction of A' is *independent of the basis chosen for* A. This is because the construction of $e_{m+1}, \ldots, e_n$ (Theorem 2.13) did not depend on the particular basis for A. That construction only utilized the fact that we can pick elements *not* in A. However, the construction of A' *does depend* on which elements $e_{m+1}, \ldots e_n$ (not in A) we pick. For example, let $V = \mathbb{R}^2$, and A be some one dimensional subspace. Then we pick e, as any vector in A, and e_2 as any vector not in A. The resulting complement A' is then the span of e_2. But $\{e_1\}$ could have been extended to a basis for V by choosing another vector $\tilde{e}_2 \ni A$. This determines a different complement $\tilde{A}'$ of A. A subspace has many possible complements. This ambiguity will not bother us since we shall only use the properties of a particular complement which do not depend on which particular complement is chosen. The dimension of the complement is such a property. It only depends on the dimension of the subspace A and the larger space V, and has the reasonable formula $\dim A' = \dim V - \dim A$, which we now prove.

Theorem 2.15 . *If A is a subspace of a linear space V and if A' is any complement of A, then*

$$\dim A + \dim A' = \dim V.$$

Thus, the dimension of A' is determined by A and V alone.

PROOF: The $\dim A$ and $\dim V$ are given data. We shall compute $\dim A'$. Since the union of a basis for A with a basis for any A' spans V (property 2), it is clear that $\dim A + \dim A' \geq \dim V$. However A and any A' intersect only at the origin (property 3) and are subspaces of V. Thus the union of their bases can span at most V, that is, $\dim A + \dim A' \leq \dim V$. These two inequalities prove the theorem.

REMARK. Some people refer to $\dim A'$ as the *codimension* of A (complementary dimension). In this way they avoid mentioning A' at all. The last theorem can be written as $\dim A + \operatorname{codim} A = \dim V$.

A simple result closes the chapter.

Theorem 2.16 . *If A is a subspace of V and A' is a complement of A, then for $X \in V$ the decomposition $X = X_1 + X_2$, $X_1 \in A$, $X_2 \in A'$ is unique.*

PROOF: Assume there are two decompositions, $X = X_1 + X_2$ and $X = \tilde{X}_1 + \tilde{X}_2$. Then $\tilde{X}_1 + \tilde{X}_2 = X_1 + X_2$ or $\tilde{X}_1 - X_1 = X_2 - \tilde{X}_2$. However the left side of this equation is in A while the right is in A'. The only element in both A and A' is 0. Thus $\tilde{X}_1 = X_1$ and $\tilde{X}_2 = X_2$.

A FIGURE GOES HERE

Exercises

1. a) Let $A = \{ X \in \mathbb{R}^2 \colon x_1 = 0 \}$. Find a basis for A and extend it to a basis for all of $\mathbb{R}^2$. Use this to define a complement A' of A. Sketch A and A'. Extend the same basis for A in a different way to a basis for all of $\mathbb{R}^2$. Use this to define another complement $\tilde{A}'$ of A. Sketch $\tilde{A}'$.

   b) Find a basis for the subspace $A = \{ X \in \mathbb{R}^2 \colon x_1 + x_2 + x_3 = 0 \}$. Extend this basis to one for all of $\mathbb{R}^3$. Define a complement A' of A induced by this extension. Write $X = (-1, 0, 7)$ as $X = Y_1 + Y_2$ where $Y_1 \in A$ and $Y_2 \in A'$.

2. a) Let $A = \{ p \in \mathcal{P}_2 \colon p(0) = 0 \}$. Find a basis for A and extend it to a basis for all of $\mathcal{P}_2$. Define A' induced by this extension. Is the particular polynomial $p(x) = 1 + x^2$ in A? in A'? Write $p(x)$ as $p(x) = q_1(x) + q_2(x)$ where $q_1(x) \in A$, $q_2(x) \in A'$.

   b) Let $A = \{ p \in \mathcal{P}_2 \colon p(1) = 0 \}$. Find a basis for A and extend it to a basis for all of $\mathcal{P}_2$.

3. Let A be a subspace of a linear space V. Show by an example that a basis for V need not contain a basis for A.

4. If $\dim V = n$ and $V = sp\{ X_1, \ldots, X_n \}$, prove that $X_1, \ldots, X_n$ are linearly independent.

5. Let $V = \mathcal{P}_4$ and A the subspace spanned by $1, x^2$ and x^4. Find three different subspaces complementary to A (you may specify a subspace by giving a basis for it).

After all this about bases, it is probably best to notify you that *properties of linear spaces are best defined and proved without introducing a particular basis.* As soon as you define a property of a linear space in terms of a basis, you must then prove that the property is intrinsic to the space itself and does not depend upon the basis you choose. We met this problem in defining the dimension in terms of a basis - and were consequently forced to prove Theorem 2.12 which stated that the property really only depended on the space itself, not on the basis chosen.

This, in fact, corresponds to one of the major requisites for laws of physics: they should not depend upon the particular coordinate system you choose (picking a coordinate system is equivalent to picking a basis). Moreover, the laws should not depend on the units you choose for each axis of the coordinate system. But these are long, involved questions which must be investigated deeply to make our remarks precise.

One should, however, distinguish theoretical issues from computational ones. In *theoretical questions*, the rule is *never pick a specific basis unless there is no way out.* On the other hand, for *computational questions you must always pick a basis.* Just as in physics, on order to perform any measurements, you must pick some specific coordinate system and specific units. If the theoretical foundations are firm, then you can feel confident that no matter what choice of basis you make, the essential nature of the results will remain unchanged.

As an example, let us consider a point P and two different fixed coordinate systems in the plane of this paper. You should feel that any motion of the point P can be described adequately in either coordinate system - and that when the observers in the two coordinate systems get together and discuss the motion of P, they will agree as to what happened. A common example is the meeting of two people from countries using different units of money.

Exercises

1. Prove that any $n + 1$ elements in a linear space of dimension n must be linearly dependent.

2. Prove that $\mathcal{P}_n$ has dimension $n + 1$.

3. Since a basis for a linear space of dimension n must contain exactly n elements, all one must test is that the n elements which are candidates for a basis are linearly independent - or equivalently that they span the space. Show that the vectors $\{X_1, \ldots, X_n\}$ form a basis for $\mathbb{R}^n$ if and only if $e_1, e_2, \ldots, e_n$ can all be expressed as a linear combination of the $\{X_1, \ldots, X_n\}$.

4. Use Exercise 3 to determine which of the following sets from bases for $\mathbb{R}^3$.
   a) $X_1 = (1, 1, 0),\ X_2 = (1, 0, 1),\ X_3 = (0, 1, 1)$.
   b) $X_1 = (1, 0, 1),\ X_2 = (1, 1, 1)$.
   c) $X_1 = (1, 0, 1),\ X_2 = (1, 1, 0),\ X_3 = (0, -1, 1)$.
   d) $X_1 = (1, 1, 1),\ X_2 = (1, 2, 3),\ X_3 = (17, 3, 9),\ X_4 = (-2, 7, -1)$.
   e) $X_1 = (-1, 0, 2),\ X_2 = (1, 1, 1),\ X_3 = (\frac{1}{2}, \frac{1}{3}, -1)$.

5. Prove that the subspace of functions in $C[0, \pi]$ which vanish at $x = 0$ and at $x = \pi$ is infinite dimensional by showing that for any integer k the functions $f_1(x) = \sin x,\ f_2(x) = \sin 2x, \ldots, f_k(x) = \sin kx, \ldots$ are all linearly independent. [Hint: Assume that $0 = \sum_{k=1}^{N} a_k \sin kx$, and show that all the a_k's must be zero by multiplying both sides by $\sin nx$ and utilizing the important formula
$$\int_0^{\pi} \sin nx \, \sin kx \, dx = \left\{ \begin{array}{ccc} 0 & , & k \neq n \\ \frac{\pi}{2} & , & k = n. \end{array} \right\}.]$$

6. Let $C^*[a, b]$ denote the set of all complex-valued functions $f(x) = u(x) + iv(x)$ which are continuous for $x \in [a, b]$. The complex number field $\mathbb{C}$ is the field of scalars for C^*. What is the dimension of the subspace $A = \{ f \in C^*[-\pi, \pi] \colon f(x) = ae^{ix} + be^{-ix},\ a,\ b \in \mathbb{C} \}$? Show that $f_1(x) = \cos x$ and $f_2(x) = \sin x$ constitute a basis for A. [Hint: Use Euler's formula (1-20)].

7. Which of the following sets of vectors form a basis for $\mathbb{R}^4$?
   a) $X_1 = (1, 0, 0, 5),\quad X_2 = (0, 3, 2, 6),\quad X_3 = (0, 0, 1, 2),\quad X_4 = (0, 0, 0, 1)$.
   b) $X_1 = (1, 6, 7, 0),\quad X_2 = (-2, 2, 5, 0),\quad X_3 = (4, 5, 6, 0),\quad X_4 = (7, 8, 3, 0)$.
   c) $X_1 = (1, 2, 5, 7),\quad X_2 = (4, 9, 11, 8),\quad X_3 = (6, 3, 12, 2),\quad X_4 = (3, -4, 7, 6)$, $X_5 = (0, 0, 0, 1)$.
   d) $X_1 = (1, 2, 3, 4),\quad X_2 = (0, 2, 3, 4),\quad X_3 = (0, 0, 3, 4),\quad X_4 = (0, 0, 0, 4)$.

8. Find a basis for the following subspaces.
   a) $A = \{ X \in \mathbb{R}^2 \colon x_1 + x_2 = 0 \}$
   b) $B = \{ X \in \mathbb{R}^3 \colon x_1 + x_2 + x_3 = 0 \}$

   c) $C = \{ p \in \mathcal{P}_3 \colon p(0) = 0 \}$
   d) $D = \{ p \in \mathcal{P}_3 \colon p(1) = 0 \}$
   e) $E = \{ u \in C^1[-1, 1] \colon u' - u = 0 \}$
   f) $F = \{ u \in C^1[-1, 1] \colon u' + 2u = 0 \}$.

Chapter 3

Linear Spaces: Norms and Inner Products

3.1 Metric and Normed Spaces

Until now we have been contented with being able to add two elements X_1 and X_2 of a linear space, and to multiply them by scalars, aX. Since only these *algebraic* operations have been defined, only algebraic questions could have been raised and answered. Notably absent were any mention of convergence, because the idea of one element of a linear space being "close" to another was not defined. In this chapter we shall introduce a distance or *metric* structure into linear spaces. Instead of lingering in the realm of generalities, we shall define metric and norm in this first section and devote the balance of the chapter to a particular kind of metric which generalizes the "Pythagorean distance" of ordinary Euclidean space. Fourier series supply a wonderful and valuable application.

Our first notion of distance, that of a *metric*, makes sense for elements X, Y, Z of an *arbitrary set* S. The idea is to define the distance $d(X,Y)$ between any two elements of S. This distance is a function which assigns to every pair of points (X,Y) a *positive real number* $d(X,Y)$ called the "distance between X and Y".

Definition. Let S be a non-empty set. A *metric* on S is a real-valued function $d: \quad S \times S \to \mathbb{R}$, where X, $Y \in S$, which has the three properties:

i) $d(X,Y) \geq 0$. $\qquad d(X,Y) = 0 \Longleftrightarrow X = Y$

ii) (symmetry) $d(X,Y) = d(Y,X)$,

iii) (triangle inequality) $d(X,Z) \leq d(X,Y) + d(Y,Z)$.

Well, they certainly are reasonable requirements for any function we intend to think of as measuring distance.

Examples.

1. This first example is trivial but acts as an important check on intuition. With it, you see that every non-empty set can be regarded as a metric space with the following

metric

$$d(X,Y) = \begin{cases} 0 & , \text{ if } X = Y \\ 1 & , \text{ if } X \neq Y. \end{cases}$$

A moments reflection will show that this is a metric—but not too useful since it is so coarse.

2. For the real line, $\mathbb{R}$, with the usual definition of absolute value we define $d(X,Y) = |X - Y|$, which is clearly a metric.

3. Another less common metric may be given to $\mathbb{R}$. We define $d(X,Y) = \frac{|X-Y|}{1+|X-Y|}$. Only the triangle inequality is not evident—and that involves some algebra. This metric has the property that the distance between any two points is always less than one, $d(X,Y) < 1$ for all $X, Y \in \mathbb{R}$.

4. $\mathbb{R}^n$ can be endowed with many metrics. Let $X = (x_1, x_2, \ldots, x_n)$, $Y = (y_1, \ldots, y_n)$ and $Z = (z_1, \ldots z_n)$ be arbitrary points in $\mathbb{R}^n$. The metric you most expect is the Euclidean distance

$$d(X,Y) = [(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2]^{1/2} = [\sum_{k=1}^{n}(x_k - y_k)^2]^{1/2}$$

Again, only the triangle inequality is not obvious. It is a consequence of the *Cauchy-Schwarz inequality*

$$\left(\sum_{k=1}^{n} x_k y_k\right)^2 \leq \sum_{k=1}^{n} x_k^2 \sum_{k=1}^{n} y_k^2, \tag{3-1}$$

which in turn is an immediate consequence of the algebraic identity

$$\left(\sum_{k=1}^{n} x_k y_k\right)^2 = \sum_{k=1}^{n} x_k^2 \sum_{k=1}^{n} y_k^2 - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i y_j - x_j y_i)^2.$$

And now the triangle inequality. Let $a_k = x_k - y_k$, and $b_k = y_k - z_k$. Then $x_k - z_k = a_k + b_k$. Thus, using Cauchy- Schwarz in the second line below, we find that

$$\begin{aligned} [d(X,Z)]^2 &= \sum_{k=1}^{n}(a_k + b_k)^2 = \sum_{k=1}^{n} a_k^2 + 2\sum_{k=1}^{n} a_k b_k + \sum_{k=1}^{n} b_k^2 \\ &\leq \sum_{k=1}^{n} a_k^2 + 2\left[\sum_{k=1}^{n} a_k^2 \sum_{k=1}^{n} b_k^2\right]^{1/2} + \sum_{k=1}^{n} b_k^2 \\ &= \left[\left(\sum_{k=1}^{n} a_k^2\right)^{1/2} + \left(\sum_{k=1}^{n} b_k^2\right)^{1/2}\right]^2 = [d(X,Y) + d(Y,Z)]^2, \end{aligned} \tag{3-2}$$

so

$$d(X,Z) \leq d(X,Y) + d(Y,Z).$$

Another proof of the Schwarz and triangle inequalities for this metric will be given later in the chapter.

5. A second metric for $\mathbb{R}^n$ is

$$d(X,Y) = \sum_{k=1}^{n} |x_k - y_k|$$

The axioms for a metric are easily verified.

6. A third metric for $\mathbb{R}^n$ is

$$d(X,Y) = \left[\sum_{k=1}^{n} |x_k - y_k|^p\right]^{1/p}, \qquad 1 \leq p < \infty.$$

Example 4 is the special case $p = 2$, while example 5 is the special case $p = 1$. And again, all but the triangle inequality are obvious. However the triangle inequality, called *Minkowski's inequality* in this general case, is not simple. We shall not prove it here. Perhaps it will appear as an exercise later.

7. The usual metric for $C[a,b]$ is the *uniform metric*

$$d(f,g) = \max_{a \leq x \leq b} |f(x) - g(x)|.$$

Geometrically, this distance is the largest vertical distance between the graphs of f and g for all $x \in [a,b]$.

8. The space $L_1[a,b]$ of functions whose absolute value is integrable has the "natural" metric

$$d(f,g) = \int_a^b |f(x) - g(x)| \, dx,$$

which can be interpreted as the total area between the two curves. Since every function which is continuous for $x \in [a,b]$ is integrable there, i.e., $C[a,b] \subset L_1[a,b]$, this metric is another metric for $C[a,b]$.

9. For the function space $C^1[a,b]$, the standard metric is

$$d(f,g) = \max_{a \leq x \leq b} |f(x) - g(x)| + \max_{a \leq x \leq b} \left|f'(x) - g'(x)\right|$$

The metric for $C^k[a,b]$ is defined similarly.

There are many theorems one can prove about *metric spaces* (a metric space is a set S on which a metric is defined). Look in any book on general topology (or point set topology, as it is often called) and you will find more than enough to satisfy you. For most of our purposes metric spaces are too general. Normed linear spaces will suffice. The norm $\|X\|$ of an element X in a linear space $\mathcal{V}$ is the "distance" of X from the origin—the 0 element of $\mathcal{V}$.

Definition. Let $\mathcal{V}$ be a linear space over the real or complex field. If to every element $X \in \mathcal{V}$ there is associated a real number $\|X\|$, the *norm of* X, which has the three properties

i) $\|X\| \geq 0$. $\quad \|X\| = 0 \Longleftrightarrow X = 0$

ii) $\|aX\| = |a| \quad \|X\|$ (homogeneity), a is a scalar,

iii) $\|X+Y\| \leq \|X\| + \|Y\|$, (triangle inequality),

then we say that $\mathcal{V}$ is a *normed linear space.*

How does a norm differ from a metric?

First of all, a norm is only defined on a *linear space* (since aX and $X+Y$ appear in the definition) whereas a metric may be defined on any set (cf. example 1 above). But if we restrict our attention to linear spaces, how do the concepts of norm and metric differ? *Every normed linear space can be made into a metric space in such a way that* $\|X\|$ *is indeed the distance of* X *from the origin*, $d(X,0) = \|X\|$. The explicit formula for $d(X,Y)$ should surprise no one

$$d(X,Y) = \|X - Y\|.$$

It is easy to check that $d(X,Y)$ is a metric. Thus every normed linear space has a "natural" metric induced upon it. However, a *linear space which has a metric need not be a normed linear space.* For example in $\mathbb{R}$, the linear space of the real numbers, the metric of example 3

$$d(X,Y) = \frac{|X-Y|}{1+|X-Y|}$$

is not associated with a norm because axiom ii) for a norm is not satisfied.

Of the examples considered earlier, all but the first and third metrics arise from norms, in the sense that

$$d(X,Y) = d(X-Y,0) = \|X-Y\|.$$

By far the most common norm in $\mathbb{R}^n$ is that given by the Pythagorean theorem (example 4). Then

$$\|X\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \left[\sum_{k=1}^{n} x_k^2\right]^{1/2}$$

and the induced metric is

$$d(X,Y) = \|X-Y\| = \left[\sum_{k=1}^{n} (x_k - y_k)^2\right]^{1/2}$$

For obvious historical reasons, we shall refer to $\mathbb{R}^2$ with this Euclidean (or Pythagorean) norm as *Euclidean* n*-space*, and still denote it by $\mathbb{R}^n$. Note that here $\mathbb{R}^n$ is a linear space with a particular way of measuring length specified using the *Euclidean distance.* This metric removes the floppiness from $\mathbb{R}^n$, giving the additional structure needed to investigate those geometrical concepts which utilize the notion of distance.

There are other ways to measure distance that are occasionally useful. For instance

$$\|X\|_1 := |x_1| + |x_2| + \cdots + |x_n|.$$

For simplicity, *unless we specify otherwise,* $\mathbb{R}^n$ *will be assumed to have the Euclidean norm.*

Once we have a norm (or metric) it becomes possible to discuss convergence of a sequence of elements.

Definition: If $\mathcal{V}$ is a normed linear space, the *sequence* $X_n \in \mathcal{V}$ *converges* to $X \in \mathcal{V}$ if, given any $\epsilon > 0$, there in an N such that

$$\|X_n - X\| < \epsilon \quad \text{for all } n > N.$$

As an example, we shall prove the sample

Theorem 3.1 . *A sequence of points* $X_n = (x_1^{(n)}, x_2^{(n)}, \ldots x_k^{(n)})$ *in* $\mathbb{R}^k$ *converges to the point* $X = (x_1, \ldots, x_k)$ *in* $\mathbb{R}^k$ *if and only if each component* $x_j^{(n)}$ *converges to its respective limit,* $\lim_{n\to\infty} x_j^{(n)} = x_j, j = 1, \ldots, k$.

Proof: i) $X_n \to X \Rightarrow x_j^{(n)} \to x_j$. This is a consequence of the trivial inequality

$$\left|x_j^{(n)} - x_j\right| \leq \sqrt{(x_1^{(n)} - x_1)^2 + \cdots + (x_k^{(n)} - x_k)^2} = \|X_n - X\|;$$

for if $\|X_n - X\| < \epsilon$ for $n > N$, then $\left|x_j^{(n)} - x_j\right| < \epsilon$ for $n > N$ too. Thus $x_j^{(n)} \to x_j$. If the subscripts are cluttering up the proof, go through it again in a special case, say $x_2^{(n)} \to x_2$.

ii) $x_j^{(n)} \to x_j \Rightarrow X_n \to X$. By hypothesis, given any $\epsilon > 0$, there are numbers $N_1, N_2, \ldots N_k$ such that $\left|x_1^{(n)} - x_1\right| < \epsilon$, for all $n > N_1$, $\left|x_2^{(n)} - x_2\right| < \epsilon$ for all $n > N_2, \ldots, \left|x_k^{(n)} - x_k\right| < \epsilon$ for all $n > N_k$. Pick $N = \max(N_1, N_2, \ldots, N_k)$. This N will work for all the $x_j^{(n)}$'s, that is, for every j,

$$\left|x_j^{(n)} - x_j\right| < \epsilon \quad \text{for all } n > N.$$

Thus

$$\begin{aligned} \|X_n - X\| &= \sqrt{(x_1^{(n)} - x_1)^2 + \ldots + (x_k^{(n)} - x_k)^2} \\ &< \sqrt{\epsilon^2 + \ldots + \epsilon^2} = \epsilon\sqrt{k}, \quad \text{for all } \quad n > N. \end{aligned} \tag{3-3}$$

Since k is a fixed finite number, this shows that $\|X_n - X\|$ may be made arbitrarily small by picking n big enough, so X_n does converge to X.
Example. In $\mathbb{R}^4$, the sequence $X_n = (\frac{n}{n+1}, 2, -\frac{1}{n}, 0)$ converges to $X = (1,2,0,0)$ since $\frac{n}{n+1} \to 1$, $2 \to 2$, $-\frac{1}{n} \to 0$, and $0 \to 0$.

A useful elementary result is

Theorem 3.2 . *If $\mathcal{V}$ is a normed linear space, and if $X_n \to X$, $Y_n \to Y$ in $\mathcal{V}$, then for any scalars a and b, $aX_n + bY_n \to aX + bY$.*

Proof: There are essentially no changes from the case of $\mathbb{R}^1$. We must show that $\|aX_n + bY_n - aX - bY\|$ can be made arbitrarily small by picking n large enough. One application of the triangle inequality

$$\|aX_n + bY_n - aX - bY\| \le \|aX_n - aX\| + \|bY_n - bY\|,$$

and the homogeneity of a norm, yields

$$\le |a| \quad \|X_n - X\| + |b| \quad \|Y_n - Y\|.$$

Because $X_n \to X$ and $Y_n \to Y$, if $n > N_1$, then $\|X_n - X\| < \epsilon$. Also, if $n > N_2$, then $\|Y_n - Y\| < \epsilon$. Pick $N = \max(N_1, N_2)$. Thus

$$\|aX_n + bY_n - aX - bY\| < |a|\,\epsilon + |b|\,\epsilon = (|a| + |b|)\epsilon, \qquad n > N,$$

and the desired convergence is proved.

For a given linear space $\mathcal{V}$, there may be two (or even more) norms defined, say $\|\ \ \|$ and $\|\ \ \|_1$ to distinguish them. Why carry them both around? First of all, a sequence may converge in one norm and not in the other. Second, even if both norms yield the same convergent sequences, one norm may be more convenient in some particular computation.
Example. Consider the linear space $C[-1,1]$ of functions $f(x)$ continuous for $x \in [-1,1]$, with the two norms (Examples 7 and 8)

$$\|f\|_\infty = \max_{-1\le x\le 1} |f(x)| \quad ; \qquad \|f\|_1 = \int_{-1}^{1} |f(x)|\, dx,$$

that is, the uniform norm and the L_1 norm. We shall exhibit a sequence of functions which converge in the second norm but not in the first. Let $f_n(x)$ be

$$f_n(x) = \begin{cases} 0, & x \in [-1, -\frac{1}{n^2}] \\ n^3(x + \frac{1}{n^2}) & x \in [-\frac{1}{n^2}, 0] \\ -n^3(x - \frac{1}{n^2}) & x \in [0, \frac{1}{n^2}] \\ 0 & x \in [\frac{1}{n^2}, 1] \end{cases}$$

Then by inspection from the graph ($\|\ \ \|_1$ is the area), we see that $\|f_n\|_\infty = n$, and $\|f_n\|_1 = \frac{1}{n}$. As $n \to \infty$, $\|f_n - 0\|_1 \to 0$ so that $f_n \to 0$ in the L_1 norm. On the other hand, $\|f_n\|_\infty \to \infty$ so the limit does not exist in the uniform norm. If you look at the graph, f_n is zero except for a spike in the interval $[-\frac{1}{n^2}, \frac{1}{n^2}]$. As $n \to \infty$, the function is zero in essentially the whole interval, except for the bit around the origin where it blows up—but it blows up slowly enough that the area under the curve tends to zero.

However, we can prove the

Theorem 3.3 . *Let f_n and f be continuous functions, $n = 1, 2, \ldots$. If $f_n \to f$ in the uniform norm, then also $f_n \to f$ in the L_1 norm.*

Remark. We have just seen that the converse is false.
Proof: An immediate consequence of the

Lemma 3.4 $\|f_n - f\|_1 \le (b-a)\|f_n - f\|_\infty$

Proof:

$$\begin{aligned} \|f_n - f\|_1 &= \int_n^b |f_n(x) - f(x)|\, dx \le \int_a^b \|f_n - f\|_\infty\, dx \\ &= \|f_n - f\| \int_a^b dx = (b-a)\|f_n - f\|_\infty \end{aligned} \tag{3-4}$$

Exercises

1. In the set $\mathbb{Z}$, define $d(m,n) = |m - n|$ where $|x|$ is ordinary absolute value. Prove that $d(m,n)$ is a metric.

2. Suppose that $d_1(X,Y)$ and $d_2(X,Y)$ are both metrics for a set S, where $X, Y \in S$. a). Show that $[d_1(X,Y)]^2$ is *not*, in general, a metric. b). Prove that $d_1 + d_2$ and $\sqrt{d_1^2 + d_2^2}$ are also metrics for S.

3. Prove that the function $d(X,Y) = \frac{|X-Y|}{1+|X-Y|}$, $X, Y \in \mathbb{R}$, is a metric, but that it is not a norm on $\mathbb{R}$.

4. Let $X = (x_1, \ldots, x_k) \in \mathbb{R}^k$. Define $\|X\|_\infty = \max_{1\le l\le k} |x_l|$.

   a) Prove that $\|X\|_\infty$ is a norm for $\mathbb{R}^k$, and write down the induced metric.

   b) Let $\|X\|_1 = \sum_{l=1}^{k} |x_l|$, and $\|X\|_2 = \left(\sum_{l=1}^{k} |x_l|^2\right)^{1/2}$.

   Prove
   $$\|X\|_\infty \le \|X\|_2 \le \|X\|_1 \le k\|X\|_\infty.$$

   c) Consider the sequence of vectors in $\mathbb{R}^3$: $X_n := (1 - \frac{1}{n}, -7, \frac{1}{n^2})$. In which of the norms $\|\ \ \|_\infty, \|\ \ \|_2, \|\ \ \|_1$ does it converge, and to what?

5. Let X_n be a sequence of elements in a normed linear space V (not necessarily finite dimensional). Prove that if $X_n \to X$, then the sequence X_n is bounded in norm (a sequence X_n in a normed linear space is *bounded* if there is an $M \in \mathbb{R}$ such that $\|X_n\| \le M$ for all n). [Hint: Compare with Theorem 6, page ??].

6. Compute the $\| \quad \|_1, \| \quad \|_2, and \| \quad \|_\infty$ (cf. Ex. 4) norms of the following vectors in $\mathbb{R}^3$.

   a) $X = (1, 2, 2)$, b) $Y = (2, -2, 1)$, c) $Z = (0, 3, -4)$, d) $W = (0, -1, 0)$.

7. Compute the $\| \quad \|_1, \| \quad \|_2, and \| \quad \|_\infty$ norms of the following functions for the interval $[-1, 1]$.

   a) $f(x) = -2x + 3$ b) $g(x) = \sin \pi x$, c) $h_n(x) = x^n$

   d) As $n \to \infty$, does the sequence h_n converge in any of these three norms?

8. Which of the following define norms for the given linear spaces?

   a) For $\mathbb{R}^3$, $[X] = x_1^2 + x_x^2 + x_3^2$

   b) For $\mathcal{P}_3$, $[p] = \max\limits_{0 \le x \le 1} p(x)$

   c) For $\mathcal{P}_3$, $[p] = \max\limits_{0 \le x \le 1} |p(x)|$

   d) For $\mathbb{R}^3$, $[X] = |x_1| + |x_2|$

   e) For $\mathbb{R}^4$, $[X] = \sqrt{1 + x_1^2 + x_2^2 + x_3^2 + x_4^2}$.

9. Prove that $[X] = \sqrt{x_1^2 + x_2^2}$ defines a norm for $\mathbb{R}^2$ (some algebraic fortitude will be needed to prove the triangle inequality).

3.2 The Scalar Product in $\mathbb{R}^2$

In Euclidean space $\mathbb{R}^2$—which we remind you is $\mathbb{R}^2$ with the Euclidean norm $\|X\| = \sqrt{x_1^2 + x_2^2}$—one can introduce many geometric concepts and develop a corresponding geometric theory. Most important of these concepts is that of angle—especially *orthogonality* (perpendicularity). It turns out that these ideas generalize almost immediately to all $\mathbb{R}^n$, and even to some exceedingly important infinite dimensional spaces. This section is devoted to the most simple situation: $\mathbb{R}^2$. Please look at the pictures.

To begin, we introduce the *scalar product* (also called the *dot product*, or *inner product*) of two vectors X and Y.
Definition: If X and Y are two vectors in $\mathbb{R}^2$, their *scalar product* $\langle X, Y \rangle$ (sometimes written $X \cdot Y$) is defined by

$$\langle X, Y \rangle = \|X\| \quad \|Y\| \cos \theta,$$

where θ is the angle between X and Y.

Notice that the scalar product of two vectors is a real number, a scalar, *not* another vector. We need not specify the direction in which θ is measured, counterclockwise or clockwise, since $\cos \theta = \cos(-\theta)$. Further, we can use either the acute or obtuse angle between X and Y since $\cos(2\pi - \theta) = \cos \theta$. It is important to observe that the scalar product of two vectors is defined independent of any coordinate system.

We are immediately led to some simple consequences.

Lemma 3.5 . *Two vectors X and Y are orthogonal if and only if $\langle X, Y \rangle = 0$*

PROOF: If X and Y are orthogonal, the angle θ between them is $\frac{\pi}{2}$, so $\langle X, Y \rangle = \|X\| \quad \|Y\| \cos \frac{\pi}{2} = 0$. In the other direction, if $\langle X, Y \rangle = 0$, then $\|X\| \quad \|Y\| \cos \theta = 0$. If neither $\|X\|$ nor $\|Y\| = 0$, then $\cos \theta = 0$, that is $\theta = \frac{\pi}{2}$ or $\frac{3\pi}{2}$. Thus X is orthogonal to Y. If $\|X\|$ or $\|Y\| = 0$, then one of them is just the point at the origin, the zero vector. We agree to say that the zero vector is orthogonal to every other vector. With this agreement, $\langle X, Y \rangle = 0 \Rightarrow X \perp Y$, and the second half of the theorem is proved too.

There is a nice geometric interpretation of the scalar product. A hint of it appeared in our last lemma. Let e be a *unit vector*, $\|e\| = 1$. Consider $\langle X, e \rangle = \|X\| \cos \theta$ (see figure). This is the *length* of the *projection* of X in the direction of e, or in other words, the length of the projection of X into the subspace spanned by the single vector e. Strictly $\langle X, e \rangle$ is not really a “length”, since “length” carries the implication of being positive, whereas the real number $\langle X, e \rangle$ will be negative if the projection “points” in the direction opposite to e. We shall, however, allow ourselves this abuse of language. The vector U_1 which is the projection of X into the subspace spanned by e is $U_1 = \langle X, e \rangle e$.

If Y is a (non-zero) vector in $\mathbb{R}^2$ which is not a unit vector, the above geometric idea goes through by making the simple observation that given any vector $Y \neq 0$, the vector $e = Y / \|Y\|$ is a unit vector in the direction of Y.

Now you are certainly wondering how in the world we compute this scalar product. You could take out your ruler, protractor and table of cosines—but we will present a more convenient method. In order to compute this as is always the case, a particular basis must be chosen. Then the vectors X and Y can be given explicitly in terms of the basis. Since we want to show that the *concepts are independent of any particular basis*, you must relax and be patient. Only after the theory has been exposed will we reveal how to compute in terms of a given basis.

Theorem 3.6 (a) $\langle X, X \rangle = \|X\|^2$

(b) $\langle X, Y \rangle = \langle Y, X \rangle$

(c) $\langle aX, Y \rangle = a \langle X, Y \rangle$ *where* $a \in \mathbb{R}$.

(d) $\langle X, aY \rangle = a \langle X, Y \rangle$, *where* $a \in \mathbb{R}$.

(e) $\langle X + Y, Z \rangle = \langle X, Z \rangle + \langle Y, Z \rangle$

(f) $\langle X, Y + Z \rangle = \langle X, Y \rangle + \langle X, Z \rangle$

(g) $|\langle X, Y \rangle| \le \|X\| \quad \|Y\|$ *(Cauchy- Schwarz inequality)*

PROOF:

(a) Obvious since $\theta = 0$ and $\cos 0 = 1$.

(b) Obvious since $\cos(-\theta) = \cos \theta$.

(c) The vectors X and aX lie along the same line through the origin. There are two cases, $a > 0$ and $a < 0$ ($a = 0$ is trivial). *If* $a > 0$, the angle θ between X and Y is identical to that between aX and Y. Since $\|aX\| = a\|X\|$ for $a > 0$, this case is proved, for

$$\langle aX,\, Y\rangle = \|aX\| \quad \|Y\| \cos\theta = a\langle X,\, Y\rangle.$$

If $a < 0$, then aX points in the direction opposite to X. Thus the angle θ_1 between aX and Y equals $\pi - \theta$, where θ is the angle between X and Y. The following computation completes the proof:

$$\begin{aligned}\langle aX,\, Y\rangle &= \|aX\| \quad \|Y\| \cos\theta_1 = |a|\, \|X\| \|Y\| \cos(\pi - \theta)\\ &= -\,|a| \quad \|X\| \quad \|Y\| \cos\theta = a\|X\| \quad \|Y\| \cos\theta = a\langle X,\, Y\rangle\end{aligned} \tag{3-5}$$

(d) By (b) and (c) and (b) again we are done

$$\langle X,\, aY\rangle = \langle aY,\, X\rangle = a\langle Y,\, X\rangle = a\langle X,\, Y\rangle.$$

(e) This is the most subtle part. We shall rely on the interpretation of the scalar product $\langle U,\, e\rangle$ as the length of the projection of U in the subspace spanned by e. First, let $e = Z/\|Z\|$ be the unit vector in the direction of Z. We shall show that $\langle X + Y,\, e\rangle = \langle X,\, e\rangle + \langle Y,\, e\rangle$. A picture is all that is needed now. Two situations are illustrated, where both X and Y are on the same side of the line perpendicular to e and

A FIGURE GOES HERE

where X and Y are on opposite sides of that line. The vector $X + Y$ is found from X and Y by the parallelogram rule for addition. Interpreting the scalar product of a vector with e as the length of the projection into the subspace (line) spanned by e, we see (look) that we must prove

$$\vec{OP} = \vec{OQ} + \vec{OM}\ .$$

But since $\vec{OA}$ and $\vec{BC}$ are on opposite sides of the same parallelogram, know that $\vec{OM} = \vec{QP}$ both in magnitude and direction. The natural substitution yields

$$\vec{OP} = \vec{OQ} + \vec{QP},$$

which is indeed all we desired. Thus

$$\langle X + Y,\, e\rangle = \langle X,\, e\rangle + \langle Y,\, e\rangle$$

To prove the general result for $Z = \|Z\|e$, multiply the last equation by $\|Z\|$, which is a scalar. Then by part a we find

$$\langle X + Y,\, \|Z\|e\rangle = \langle X,\, \|Z\|e\rangle + \langle Y,\, \|Z\|e\rangle,$$

or

$$\langle X + Y,\, Z\rangle = \langle X,\, Z\rangle + \langle Y,\, Z\rangle,$$

We are done.

(f) By parts (b), (e) and (b) again we obtain the result.

$$\langle X,\, Y + Z\rangle = \langle Y + Z,\, X\rangle = \langle Y,\, X\rangle + \langle Z,\, X\rangle = \langle X,\, Y\rangle + \langle X,\, Z\rangle$$

(g) Obvious since $|\cos\theta| \le 1$. It is evident that equality occurs when and only when $\cos\theta \pm 1$, that is, when X and Y lie along the same line (possibly pointing in opposite directions).

If e is a unit vector, we know how to find the projection U_1 of a given vector X into the subspace spanned by e, it is $U_1 = \langle X,\, e\rangle e$. Similarly, if Y is any vector—not necessarily of length one, then since $Y/\|Y\|$ is a unit vector in the direction of Y, the projection of X into the subspace spanned by Y is $\langle X,\, Y/\|Y\|\rangle Y/\|Y\| = \langle X,\, Y\rangle Y/\|Y\|^2$. We can also find the projection U_2 of X into the subspace orthogonal to the unit vector e. Since the sum of U_1 and U_2 must add up to X, $X = U_1 + U_2$, we find that $U_2 = X - U_1 = X - \langle X,\, e\rangle e$. Thus, we have proved

Theorem 3.7 . *If X and Y are any two vectors, $\|Y\| \ne 0$, then X can be decomposed into two vectors U_1 and U_2, $X = U_1 + U_2$ such that U_1 is in the subspace spanned by Y and U_2 is in the orthogonal subspace. The decomposition is given by $U_1 = \langle X,\, Y\rangle Y/\|Y\|^2$ and $U_2 = X - \langle X,\, Y\rangle Y/\|Y\|^2$, so that*

$$X = \langle X,\, Y\rangle \frac{Y}{\|Y\|^2} + (X - \langle X,\, Y\rangle \frac{Y}{\|Y\|^2}).$$

Without further delay, we shall show how to compute the scalar product of two vectors. In order to carry this out we must introduce a basis. Let X_1 and X_2 be any two vectors in $\mathbb{R}^2$ which span $\mathbb{R}^2$. Then every vector $X \in \mathbb{R}^2$ can be written in the form $X = a_1X_1 + a_2X_2$, where the scalars a_1 and a_2 are determined uniquely by the vector X. Now it is most convenient to have a basis whose vectors are i) orthogonal to each other and ii) have unit length. Such a basis is called an *orthonormal* basis (orthogonal and normalized to have unit length). In other words e_1 and e_2 are an orthonormal basis for $\mathbb{R}^2$ if $\|e_j\| = 1$ and $\langle e_1,\, e_2\rangle = 0$. This requirement is most conveniently stated by introducing the *Kronecker symbol* δ_{jk}

$$\delta_{jk} = \begin{cases} 0 & j \ne k\\ 1 & j = k.\end{cases}$$

Then the orthonormality property reads $\langle e_j,\, e_k\rangle = \delta_{jk}$, $j, k = 1, 2$. The notation is perhaps excessive for this simple case, but will really be useful in our generalizations.

Therefore, let e_1 and e_2 be an orthonormal basis for $\mathbb{R}^2$, so that if $X \in \mathbb{R}^2$, $X = x_1e_1 + x_2e_2$. *Fix* the basis throughout the ensuing discussion. Observe that x_1 and x_2 can be computed in terms of X, and the basis vectors e_1 and e_2, viz $\langle X,\, e_1\rangle = \langle x_1e_1 + x_2e_2,\, e_1\rangle = x_1\langle e_1,\, e_1\rangle + x_2\langle e_2,\, e_1\rangle = x_1$, since $\langle e_1,\, e_1\rangle = 1$ and $\langle e_1,\, e_2\rangle = 0$. Similarly, $\langle X,\, e_x\rangle = x_2$. Thus we have proved

Theorem 3.8 . *If* $\{\, e_j \,\}$, $\quad j = 1, 2$, *form an orthonormal basis for* $\mathbb{R}^2$, *then every vector* $X \in \mathbb{R}^2$ *can be written as* $X = \sum_{j=1}^{2} x_j e_j$, *where* x_j *is the length of the projection of* X *into the subspace spanned by* e_j, $\quad x_j = \langle X,\, e_j \rangle$.

If $X = x_1e_1 + x_2e_2$ and $Y = y_1e_1 + y_2e_2$ are any two vectors in $\mathbb{R}^2$, then

$$\begin{aligned}
\langle X,\, Y\rangle &= \langle x_1e_1 + x_2e_2,\, y_1e_1 + y_2e_2\rangle \\
&= \langle x_1e_1 + x_2e_2,\, y_1e_1\rangle + \langle x_1e_1 + x_2e_2,\, y_2e_2\rangle \\
&= \langle x_1e_1,\, y_1e_1\rangle + \langle x_2e_2,\, y_1e_1\rangle + \langle x_1e_1,\, y_2e_2\rangle + \langle x_2e_2,\,,\, y_2\rangle \\
&= x_1y_1\langle e_1,\, e_1\rangle + x_2y_x\langle e_2,\, e_1\rangle + x_1y_2\langle e_1,\, e_2\rangle + x_2y_2\langle e_2,\, e_2\rangle \\
&= x_1y_1 + 0 + 0 + x_2y_2 = x_1y_1 + x_2y_2.
\end{aligned}$$

Now you see how easy it is to compute the scalar product of X and Y in terms of the representation from an orthonormal basis. Let us rewrite our result formally.

Theorem 3.9 . *Let* $\{\, e_j \,\}$, $j = 1, 2$, *form an orthonormal basis for* $\mathbb{R}^2$. *If* $X = \sum_{j=1}^{2} x_j e_j$ *and* $Y = \sum_{j=1}^{2} v_j e_j$, *then*

$$\langle X,\, Y\rangle = \sum_{j=1}^{2} x_j y_j = x_1y_1 + x_2y_2.$$

Some numerical examples should reassure you of the basic simplicity of the computation. As our orthonormal basis in $\mathbb{R}^2$, we choose the vectors $e_1 = (1, 0)$ and $e_2 = (0, 1)$. These both have unit length, and are perpendicular (one is on the horizontal axis, the other on the vertical axis). Let $X = (-2, 3)$. Then $X = -2e_1 + 3e_2$. Notice that $-2e_1$ and $3e_2$ are exactly the projections of X into the subspaces spanned by e_1 and e_2 respectively. If $Y = (1, -2)$, then our theorem shows that

$$\langle X,\, Y\rangle = (-2)(-1) + (3)(-2) = -2 - 6 = -8.$$

From this computation we can reverse the geometric procedure and find the angle θ between X and Y, for we know the formula

$$\cos\theta = \frac{\langle X,\, Y\rangle}{\|X\| \quad \|Y\|}.$$

In this example, $\langle X,\, Y\rangle = -8$, $\|X\| = \sqrt{4+9} = \sqrt{13}$ and $\|Y\| = \sqrt{1+4} = \sqrt{5}$. Thus $\theta = \cos^{-1}(\frac{-8}{\sqrt{65}})$ which can be evaluated by consulting your favorite numerical tables.

It is equally simple to check if two vectors are orthogonal. Let $X = (2, -3)$ and $Y = (6, 4)$. Then $\langle X,\, Y\rangle = (2)(6) + (-3)(4) = 0$; consequently X and Y are orthogonal.

Another consequence is the law of cosines. Let $X = (x_1, x_2)$ and $Y = (y_1, y_2)$. Then from the parallelogram construction, the length of the segment joining the tip of X to the tip of Y has length $\|Y - X\|$. But

$$\begin{aligned}
\|Y - X\|^2 &= \langle Y - X,\, Y - X\rangle \\
&= \|X\|^2 + \|Y\|^2 - 2\langle X,\, Y\rangle \\
&= \|X\|^2 + \|Y\|^2 - 2\|X\| \quad \|Y\|??\theta.
\end{aligned}$$

One more example. We shall find the distance of the point $P = (-3, 2)$ from the coset $A = \{\, X = (x_1, x_2) \in \mathbb{R}^2 \colon x_1 - 2x_2 = 2 \,\}$. Pick some point in X_0 in A, say $X_0 = (3, \frac{1}{2})$. The distance d from P to A is then the length of the projection of the segment X_0P onto a line l orthogonal to A. First of all, we can replace the segment X_0P by a vector from the origin 0 to the point $Q = P - X_0 = (-6, \frac{3}{2})$, for the length of the projection of $\bar{0Q}$ onto a line l orthogonal to A is equal to the length of the projection of $\bar{X_0P}$ onto l (see figure). Now we have the vector $Q = (-6, \frac{3}{2})$; all we need to compute the desired projection is another vector N orthogonal to A, for then $d = |\langle Y,\, N/\|N\|\rangle|$.

To find a vector N orthogonal to A, we realize that N will also be orthogonal to the subspace $\mathcal{S}$ parallel to the coset A so $A = \mathcal{S} + X_0$, where $\mathcal{S} = \{\, X = (x_1, x_2) \in \mathbb{R}^2 \colon x_1 - 2x_2 = 0 \,\}$. If $N = (n_1, n_2)$ and X is any element of $\mathcal{S}$, since $N \perp \mathcal{S}$, we must have $0 = \langle X,\, N\rangle = x_1n_1 + x_2n_2$. However $X \in \mathcal{S}$ so $x_1 - 2x_2 = 0$. We want the equation $x_1n_1 + x_2n_2 = 0$ to hold for *all* points on $x_1 - 2x_2 = 0$, that is for all $X \in \mathcal{S}$. This is only possible if $n_1 = 1 \cdot c$ and $n_2 = -2 \cdot c$, where c is any constant. Thus $N = c(1, -2)$ and $\|N\| = |c|\,\sqrt{5}$. The distance d between the point P and the coset A is then

$$\begin{aligned}
d = |\langle Y,\, N/\|N\|\rangle| &= \left|\langle(-6, \frac{3}{2}),\, \frac{c}{|c|\,\sqrt{5}}(1, -2)\rangle\right| \\
&= \left|(-6)(\frac{c}{|c|\,\sqrt{5}}) + (\frac{3}{2})(\frac{-2c}{|c|\,\sqrt{5}})\right| = \frac{9}{\sqrt{5}}.
\end{aligned} \tag{3-6}$$

This example contained a plethora of ideas. It would be wise to go through it again and list the constructions and concepts used. The exercises will develop many of them in greater generality.

Now you should try some problems on your own.

Exercises

1. If $X = (3, 4)$ and $Y = (5, -12)$ are two points in $\mathbb{R}^2$, find the angle between $\vec{OX}$ and $\vec{OY}$, where 0 is the origin.

2. If $X = (3, -4)$ and $Y = (5, 12)$ are two vectors in $\mathbb{R}^2$, find vectors $U_1 \in \text{span}(Y)$ and U_2 orthogonal to $\text{span}(Y)$ such that $X = U_1 + U_2$.

3. Show that the vector $N = (a_1, a_2)$ is perpendicular to the straight line whose equation is $a_1x_1 + a_2x_2 = c$ (you will have to supply the natural definition of what it means for a vector to be perpendicular to a straight line).

4. a) Find the distance of the point $P = (2, -1)$ from the coset $A = \{\, X \in \mathbb{R}^2 \colon x_1 + x_2 = -2 \,\}$.

b) Find the distance between the two "parallel" cosets A defined above and $B = \{X \in \mathbb{R}^2 \colon x_1 + x_2 = 1\}$. (Hint: Draw a figure and observe that $P \in B$).

5. a) Prove that the distance d of the point $P = (y_1, y_2)$ from the coset $A = \{X \in \mathbb{R}^2 \colon a_1x_1 + a_2x_2 = c\}$ is given by

$$d = \frac{|a_1y_1 + a_2y_2 - c|}{\sqrt{a_1^2 + a_2^2}}.$$

b) Prove that the distance d between the two "parallel" cosets $A = \{X \in \mathbb{R}^2 \colon a_1x_1 + a_2x_2 = c_1\}$, and $B = \{X \in \mathbb{R}^2 \colon a_1x_1 + a_2x_2 = c_2\}$ is given by

$$d = \frac{|c_1 - c_2|}{\sqrt{a_1^2 + a_2^2}}.$$

(Hint: If you use part (a) and are cunning, the derivation takes but one line).

6. a) If it is known that $\langle X, Y_1\rangle = \langle X, Y_2\rangle$, and that $\|X\| \neq 0$ for *a fixed* X, can you "cancel" X from both sides and conclude that $Y_1 = Y_2$? Reason?

b) If it is known that $\langle X, Y\rangle = 0$ for *every* X, can you conclude that $Y = 0$? Reason?

c) If it is known that $\langle X, Y_1\rangle = \langle X, Y_2\rangle$ for *every* X, can you conclude that $Y_1 = Y_2$? Reason?

7. a) Show that the vector

$$Z = \frac{\|X\|Y + \|Y\|X}{\|X\| + \|Y\|}.$$

bisects the angle between the vectors X and Y.

b) Show that the vector $\|X\|Y + \|Y\|X$ is perpendicular to the vector $\|Y\|X - \|X\|Y$.

8. Express the angle between an edge and a diagonal of a rectangle in terms of the scalar product.

9. Let two of the sides of a parallelogram be given by the vectors X and Y. The *parallelogram theorem* states that the sum of the squares of the sides is equal to the sum of the squares of the diagonals, that is,

$$\|X + Y\|^2 + \|X - Y\|^2 = 2\|X\|^2 + 2\|Y\|^2.$$

Prove this in two ways: i) using elementary geometry, and ii) using only the fact that X and Y are elements of a linear space, and the properties of the scalar product contained in Theorem 4 (using 4a to define $\|\quad\|$).

10. Let X be any vector in $\mathbb{R}^2$, and let e be a unit vector. Define the vector $U = ae$, where $a = \langle X, e\rangle$ is the length of the projection of X into the subspace spanned by e, and $V = \alpha e$, where α is any scalar. Prove that

$$\|X - V\|^2 \geq \|X - U\|^2 = \|X\|^2 - \|U\|^2 = \|X\|^2 - a^2.$$

This shows that in the subspace spanned by e, the vector closest to X is the projection U of X into that subspace.

11. If X is orthogonal to Y, prove the Pythagorean theorem $\|X + Y\|^2 = \|X\|^2 + \|Y\|^2$ using only $\|V\|^2 = \langle V, V\rangle$ and the properties of a scalar product in Theorem 4.

12. Let X and Y be orthogonal elements of $\mathbb{R}^2$, with neither $\|X\|$ nor $\|Y\|$ zero. Prove that X and Y are linearly independent. Do *not* introduce a basis.

3.3 Abstract Scalar Product Spaces

We shall turn the tables around. Whereas in the last section we defined the scalar product geometrically and deduced its properties, in this section we define a scalar product space as a linear space upon which a scalar product is defined, and the scalar product is stipulated to have the properties deduced earlier. After presenting our abstract definition, we shall give examples—other than $\mathbb{R}^2$ —of scalar product spaces.

Definition. A linear space H is called a *real scalar product space* if to every pair of elements $X, Y \in H$ is associated a real number $\langle X, Y\rangle$, *the scalar product of X and Y*, which has the properties

1. $\langle X, X\rangle \geq 0$ with equality if and only if $X = 0$.
2. $\langle X, Y\rangle = \langle Y, X\rangle$
3. $\langle aX, Y\rangle = a\langle X, Y\rangle$, $a \in \mathbb{R}$
4. $\langle X + Y, Z\rangle = \langle X, Z\rangle + \langle Y, Z\rangle$

You should observe that the scalar product in $\mathbb{R}^2$ does have these properties (Theorem 4). Using $\mathbb{R}^2$ as our model, it is natural to *define* $\|X\| = \sqrt{\langle X, X\rangle}$ and suspect that $\|\quad\|$ is indeed a norm on the linear space H. This is true, but proving the triangle inequality for this norm using *only* properties 1-4 will take some work. We shall do just that after presenting

Examples

1. Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ be points in the linear space $\mathbb{R}^2$. We define

$$\langle X, Y\rangle = x_1y_1 + x_2y_2 + \ldots + x_ny_n.$$

Only easy algebra is needed to verify that the real number $\langle X, Y\rangle$ satisfies all of the properties of a scalar product. It turns out (after we prove the triangle inequality) that the natural norm $\|X\| = \sqrt{\langle X, X\rangle}$ is the Euclidean norm, so this is $\mathbb{R}^2$.

2. This example is the first hint that our abstractions are fruitful. Let the functions $f(x)$ and $g(x)$ be points in the linear function space $C[a,b]$ of real-valued functions continuous for $a \le x \le b$. We define

$$\langle f,\, g\rangle = \int_a^b f(x)g(x)\,dx.$$

You might be surprised; in any event let us verify that the real number $\langle f,\, g\rangle$ associated with the pair of functions f and g does satisfy the four properties of a scalar product.

a) $\langle f,\, f\rangle = \int_a^b f^2(x)\,dx$. This is clearly non-negative and $f = 0$ implies that $\langle f,\, f\rangle = 0$. All we must show is that if $\langle f,\, f\rangle = \int_a^b f^2(x)\,dx = 0$, then $f = 0$. By contradiction, assume $f(x) \ne 0$. Then there is some point $x_0 \in [a,b]$ such that $f(x_0) = c \ne 0$. Thus $f^2(x_0) = c^2 > 0$. Since f—and hence f^2—is continuous, this means that f^2 is positive in some interval about x_0 (p. 29b, Theorem I), so that $\int_a^b f^2(x)\,dx > 0$, the desired contradiction.

b) $\langle f,\, g\rangle = \int_a^b f(x)g(x)\,dx = \int_a^b g(x)f(x)\,dx = \langle g,\, f\rangle$.

c) $\langle \alpha f,\, g\rangle = \int_a^b \alpha f(x)g(x)\,dx = \alpha \int_a^b f(x)g(x)\,dx = \alpha\langle f,\, g\rangle$, where $\alpha \in \mathbb{R}$.

d) $$\begin{aligned}\langle f+g,\, h\rangle &= \int_a^b (f(x)+g(x))h(x)\,dx\\ &= \int_a^b f(x)h(x)\,dx + \int_a^b g(x)h(x)\,dx\\ &= \langle f,\, h\rangle + \langle g,\, h\rangle.\end{aligned}$$

There. We did it. After we prove the triangle inequality for an abstract scalar product space, the natural candidate for a norm $\|f\|$ is a norm:

$$\|f\| = \sqrt{\int_a^b f^2(x)\,dx}.$$

I like this space very much. You will be meeting it often, becoming much more intimate with its finer features. We shall—somewhat improperly—refer to this linear space with the given scalar product as $L_2[a,b]$. The name is improper since $L_2[a,b]$ is customarily used for our space but with more general functions and an extended notion of integration.

3. Let $f(x)$ and $g(x)$ be in $C[0,\infty]$. This time define

$$\langle f,\, g\rangle = \int_0^\infty f(x)g(x)e^{-x}\,dx.$$

Since e^{-x} is continuous and positive for all x, we are assured that $\langle f,\, f\rangle \ge 0$, with equality if and only if $f = 0$. The other properties of an inner product follow from simple manipulations. Do them.

Remark: Complex scalar product spaces are defined similarly. For them, $\langle X,\, Y\rangle$ may be a *complex* number, and *complex* scalars are admitted. The only change in the axioms is that property 2 is dropped in favor of

$$\bar{2}.\quad \langle Y,\, X\rangle = \overline{\langle X,\, Y\rangle},$$

where the bar means take the complex conjugate of the complex number $\langle X,\, Y\rangle$. Since we shall not develop the theory far enough, our attention henceforth will be restricted to real scalar product spaces.

The first order of business is to prove that the natural candidate for a *norm* $\|X\| = \sqrt{\langle X,\, X\rangle}$ is in fact a norm for the linear space $\mathcal{V}$. Only properties 1-4 may be used.

1. $\|X\| \ge 0$, with equality if and only if $X = 0$. This follows immediately from the corresponding property of $\langle X,\, X\rangle$.

2. $\|aX\| = |a| \quad \|X\|$. For $\|aX\| = \sqrt{\langle aX,\, aX\rangle} = \sqrt{a^2\langle X,\, X\rangle} = |a|\sqrt{\langle X,\, X\rangle} = |a| \quad \|X\|$.

   The proof of the triangle inequality

3. $\|X+Y\| \le \|X\| + \|Y\|$ involves more labor. We shall first need to prove the Cauchy-Schwarz inequality (cf. Theorem 4,g).

Theorem 3.10 *(Cauchy-Schwarz inequality).*

$$|\langle X,\, Y\rangle| \le \|X\|\|Y\|.$$

PROOF: If either $\|X\|$ or $\|Y\|$ is zero, this is immediate. Thus, assume that neither $\|X\|$ nor $\|Y\|$ is zero and define

$$U = \frac{X}{\|X\|}, \qquad V = \frac{Y}{\|Y\|},$$

so that both U and V are unit vectors, $\|U\| = \|V\| = 1$. Then

$$\begin{aligned}0 \le \|U \pm V\|^2 &= \langle U \pm V,\, U \pm V\rangle\\ &= \langle U,\, U\rangle \pm \langle U,\, V\rangle \pm \langle V,\, U\rangle + \langle V,\, V\rangle\ .\\ &= \|U\|^2 \pm 2\langle U,\, V\rangle + \|V\|^2,\end{aligned}$$

Since $\|U\| = 1$ and $\|V\| = 1$, this shows $\pm\langle U,\, V\rangle \le 1$. Substituting for U and V, we obtain the inequality sought:

$$|\langle X,\, Y\rangle| \le \|X\|\|Y\|.$$

Theorem 3.11 *(Triangle inequality)* $\|X+Y\| \le \|X\| + \|Y\|$.

PROOF: This is identical to that given in section 1. $\|X+Y\|^2 = \langle X+Y,\, X+Y\rangle = \|X\|^2 + 2\langle X,\, Y\rangle + \|Y\|^2$. By Cauchy-Schwarz, $\langle X,\, Y\rangle \le \|X\| \quad \|Y\|$, so

$$\|X+Y\|^2 \le \|X\|^2 + 2\|X\| \quad \|Y\| + \|Y\|^2 = (\|X\| + \|Y\|)^2.$$

Now take square root of both sides to find

$$\|X+Y\| \le \|X\| + \|Y\|.$$

Nice, eh? See how clean everything is. We have proved

Theorem 3.12 . *If H is a scalar product space and we define $\|X\| = \sqrt{\langle X,\, X\rangle}$ in terms of the scalar product, then $\| \quad \|$ is a norm and H is a normed linear space with that norm. This special case where the norm is induced by a scalar product is called a pre-Hilbert space (an honest Hilbert space has the additional property of being "complete").*

Let us state two easy algebraic consequences of our axioms for a scalar product. The proofs are identical to those of Theorem 4 in the previous section.

Theorem 3.13

$$\langle X,\, aY\rangle = a\langle X,\, Y\rangle, \quad a \in \mathbb{R} \tag{3-7}$$

$$\langle X,\, Y+Z\rangle = \langle X,\, Y\rangle + \langle X,\, Z\rangle, \tag{3-8}$$

Needless to say, we hope you are still thinking in the geometric terms presented earlier. In particular, the next definition should be reasonable.

Definition Two vectors X, Y are said to be *orthogonal* if $\langle X,\, Y\rangle = 0$.

The Pythagorean theorem suggests

Theorem 3.14 . *If X and Y are orthogonal, then*

$$\|X \pm Y\|^2 = \|X\|^2 + \|Y\|^2,$$

and conversely.

PROOF: Both parts are an immediate consequence of the identity

$$\|X \pm Y\|^2 = \langle X+Y,\, X+Y\rangle = \|X\|^2 \pm 2\langle X,\, Y\rangle + \|Y\|^2.$$

Examples.

1. Let $X = (2,3,-1)$ and $Y = (1,-1,-1)$ be points in $\mathbb{R}^3$, where we use the scalar product of example 1 in this section. Then $\langle X,\, Y\rangle = 2\cdot 1 + 3(-1) + (-1)(-1) = 0$ so X and Y are orthogonal. Similarly $X = (2,3,1,-1)$ and $Y = (3,-3,3,0)$ in $\mathbb{R}^4$ are orthogonal. A useful example is supplied by the vectors $e_1 = (1,0,0,\dots,0)$, $e_2 = (0,1,0,0,\dots,0),\dots,e_n = (0,0,\dots,0,1)$ in $\mathbb{R}^n$. These are *orthonormal* since $\langle e_k,\, e_k\rangle = 1$, but $\langle e_k,\, e_l\rangle = 0$, $k \ne l$, that is, $\langle e_k,\, e_l\rangle = \delta_{kl}$.

2. Consider the functions $\Phi_k(x) = \sin kx$ in $L_2[-\pi,\pi]$, where $k = 1,2,3,\dots$. Then, since $\sin\theta\sin\Psi = \frac{1}{2}[\cos(\theta-\Psi) - \cos(\theta+\Psi)]$, we find that

$$\langle \Phi_k,\, \Phi_k\rangle = \int_{-\pi}^{\pi} \sin^2 kx\, dx = \pi$$

and for $k \ne \ell$

$$\langle \Phi_k,\, \Phi_\ell\rangle = \int_{-\pi}^{\pi} \sin kx\, \sin \ell x\, dx = 0.$$

as a computation reveals. Thus in $L_2[-\pi,\pi]$ the function $\sin kx$ is orthogonal to the function $\sin lx$ when $k \ne l$. The whole computation may be summarized by

$$\langle \Phi_k,\, \Phi_l\rangle = \langle \sin kx,\, \sin \ell x\rangle = \pi\delta_{k\ell}.$$

It is only the factor π which does not allow us to say that the Φ_k are *orthonormal—but that is easily patched up. Let $e_k(x) = \frac{\sin kx}{\sqrt{\pi}}$. Then*

$$\langle e_k,\, e_\ell\rangle = \langle \frac{\sin kx}{\sqrt{\pi}},\, \frac{\sin \ell x}{\sqrt{\pi}}\rangle$$
$$= \frac{1}{\pi}\langle \sin kx,\, \sin \ell x\rangle,$$

or

$$\langle e_k,\, e_\ell\rangle = \delta_{k\ell}.$$

Therefore the functions $e_k(x) = \frac{\sin kx}{\sqrt{\pi}}$ are orthonormal. Don't attempt to imagine it. Just keep on thinking of a big $\mathbb{R}^2$ and all will be well.

So far we have discussed the notion of two vectors X and Y being orthogonal. This can be restated as one vector X being orthogonal to the subspace A spanned by Y, for all vectors in A are of the form aY where a is a scalar, and $\langle X,\, aY\rangle = 0 \iff \langle X,\, Y\rangle = 0$ since $\langle X,\, aY\rangle = a\langle X,\, Y\rangle$. One can also introduce the concept of a vector X being orthogonal to an arbitrary subspace A. Think of A as being a plane (through the origin of course).

Definition The vector X is *orthogonal to the subspace* A if X is orthogonal to every vector in the subspace A.

In practice, the usual way to check if X is orthogonal to the subspace A is as follows. Pick some basis $\{Y_1, Y_2, \dots\}$ for A. Then every $Y \in A$ is of the form

$$Y = \sum a_k Y_k$$

(if the basis has an infinite number of elements—that is, if A is infinite dimensional—one should worry about convergence; however we shall ignore that issue for now). By the algebraic rules for the scalar product, we find that

$$\langle X,\, Y\rangle = \langle X,\, \sum a_k Y_k\rangle = \sum a_k \langle X,\, Y_k\rangle.$$

Thus, X is orthogonal to the subspace A if X is orthogonal to every element in some basis for A $\langle X, Y_k\rangle = 0$.

For example, if A is the $x_1 \quad x_2$ plane in $\mathbb{R}^3$, and X is the vector $(0,0,1)$, then we can show that $X=(0,0,1)$ is orthogonal to A by showing it is orthogonal to both the vector $e_1=(1,0,0)$ and to $e_2=(0,1,0)$, since e_1 and e_2 form a basis for A. The computation $\langle X, e_1\rangle = 0$ and $\langle X, e_2\rangle = 0$ is immediate. Because $Y_1=(1,2,0)$ and $Y_2=(1,-1,0)$ also form a basis for A, we could prove that X is orthogonal to A by showing that $\langle X, Y_1\rangle = 0$ and $\langle X, Y_2\rangle = 0$—which is equally simple.

A less obvious example is supplied by the function $\Psi(x) = \cos x$ which is orthogonal to the subspace A spanned by $\Phi_1(x) = \sin x$, $\Phi_2(x) = \sin 2x, \ldots, \Phi_n(x) = \sin nx$ in $L_x(-\pi,\pi)$. The proof is a consequence of the integration formula

$$\langle \Psi, \Phi_k\rangle = \int_{-\pi}^{\pi} \cos x \sin kx\, dx = 0 \quad \text{for all } k.$$

Even more general than a vector being orthogonal to a subspace is the idea that *two subspaces A and B are orthogonal*, by which we mean that every vector in A is orthogonal to every vector in B. If A is a subspace of a scalar product space H, then it is natural to define the *orthogonal complement* $A^\perp$ of A as the set

$$A^\perp = \{\, X \in H \colon \langle X, Y\rangle = 0 \ \text{ for all } \ Y \in A \,\}$$

of vectors X orthogonal to A, that is, orthogonal to every vector $Y \in A$. *The set $A^\perp$ is a subspace* since it is closed under vector addition and multiplication by scalars (Theorem 2, p. 142).

Without fear of evoking surprise, we define the angle θ between two vectors X and Y by the formula

$$\cos\theta = \frac{\langle X, Y\rangle}{\|X\|\,\|Y\|}.$$

No matter what X and Y are, this defines a real angle since the right side of the equation is a real number between -1 and $+1$ (by the Cauchy-Schwarz inequality). To be honest, there is little use for the concept of angles other than right angles. In $\mathbb{R}^3$ the formula has some use, but is totally unused for more general scalar product spaces.

If we are given a set of linearly independent vectors $\{\, X_1, X_2, \ldots \}$ which span a linear scalar product space H, how can we construct an orthonormal set $\{\, e_1, e_2, \ldots \}$ which also spans the space? The process is carried out inductively. Let $e_1 = \frac{X_1}{\|X_1\|}$. Now we want a unit vector e_2 orthogonal to e_1. A reasonable candidate is

$$\tilde{e}_2 = X_2 - \langle X_2, e_1\rangle e_1,$$

which is X_2 with the projection of X_2 onto e_1 subtracted off (see fig.) This vector $\tilde{e}_2$ is orthogonal to e_1 since $\langle \tilde{e}_2, e_1\rangle = 0$. We divide by its length to obtain the unit vector e_2,

$$e_2 = \frac{X_2 - \langle X_2, e_1\rangle e_1}{\|X_2 - \langle X_2, e_1\rangle e_1\|}.$$

Next we take X_2 and subtract off both its projection into the subspace spanned by e_1 and e_2

$$\tilde{e}_3 = X_3 - [\langle X_3, e_1\rangle e_1 + \langle X_3, e_2\rangle e_2].$$

This vector $\tilde{e}_3$ is orthogonal to both e_1 and e_2. Normalize it to get $e_3 = \tilde{e}_3/\|\tilde{e}_3\|$.

More generally, say we have used the vectors $X_1, X_2, \ldots, X_k$ to obtain the orthonormal set $e_1,\ e_2, \ldots, e_k$. Then e_{k+1} is given by

$$e_{k+1} = \frac{X_{k+1} - \sum_{l=1}^{k} \langle X_{k+1}, e_l\rangle e_l}{\|X_{k+1} - \sum_{l=1}^{k} \langle X_{k+1}, e_l\|}$$

This procedure is called the *Gram-Schmidt orthogonalization process.* With it we can assert that if some set of linearly independent vectors spans a linear space A, we might as well suppose that those vectors constitute an orthonormal set, for if they don't just use Gram-Schmidt to construct a set that is orthonormal.

The next result is a useful observation.

Theorem 3.15 . *A set $\{\, X_1, X_2, \ldots, X_n \,\}$ of orthogonal vectors, none of which is the zero vector, is necessarily linearly independent.*

PROOF: The hypothesis states that $\langle X_j, X_k\rangle = 0$, $j \neq k$ and that $\langle X_j, X_j\rangle \neq 0$. Assume there are scalars $a_1, a_2, \ldots a_n$ such that

$$0 = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n.$$

We shall show that $a_1 = a_2 = \ldots = a_n = 0$. Take the scalar product of both sides with the vector X_1. Then

$$\langle 0, X_1\rangle = a_1\langle X_1, X_1\rangle + a_2\langle X_2, X_1\rangle + \cdots + a_n\langle X_n, X_1\rangle.$$

so that

$$0 = a_1\langle X_1, X_1\rangle.$$

Since $\langle X_1, X_1\rangle \neq 0$, we conclude that $a_1 = 0$. Similarly, by taking the scalar product with X_2 we find that $a_2 = 0$, and so on.

An easy consequence of this theorem is the fact that the functions $f_n(x) = \sin nx$, $n = 1, 2, \ldots, N$ where $x \in [-\pi, \pi]$ are linearly independent, for they are orthogonal (cf. Exercise 5, p. ???).

Say we are given an orthonormal set of n vectors, $\{\, e_j \,\}$, $\quad j = 1, \ldots, n, \quad \langle e_j, e_k\rangle = \delta_{jk}$, and X an element of the linear space A spanned by the $\{\, e_j \,\}$. Then

$$X = \sum_{j=1}^{n} x_j e_j,$$

where the x_j are uniquely determined just from the general theory of linear spaces (p. 160, Theorem 10). In the special case of a scalar product space we can conclude even more.

Theorem 3.16 . *Let* $\{e_j,\ j=1,\ldots,n\}$ *be an orthonormal set of vectors which span* A*. Then every vector* $X\in A$ *can be uniquely written as* $X=\sum_{j=1}^{n}x_je_j$*, where* x_j *is the length of the projection of* X *into the subspace spanned by* e_j*, that is,* $x_j=\langle X,\, e_j\rangle$*. The* x_j *are the Fourier coefficients of* X *with respect to the orthonormal basis* $\{e_j\}$*.*

PROOF: This is identical to Theorem 6 of the last section. Take the inner product of both sides of $X=\sum_{n=1}^{n}x_je_j$ with e_k. Then

$$\begin{aligned}\langle X,\, e_k\rangle &= \langle\sum_{j=1}^{n}x_je_j,\, e_k\rangle\\ &= \sum_{j=1}^{n}x_j\langle e_j,\, e_k\rangle=\sum_{j=1}^{n}x_j\delta_{jk},\end{aligned}\tag{3-9}$$

so that

$$\langle X,\, e_k\rangle = x_k.$$

Furthermore,

Theorem 3.17 . *Let* $\{e_j\}$*,* $j=1,\ldots,n$ *be an orthonormal set of vectors which span* A*. If* $X=\sum_{j=1}^{n}x_je_j$ *and* $Y=\sum_{j=1}^{n}y_je_j$ *are vectors in* A*, then*

$$\langle X,\, Y\rangle=\sum_{j=1}^{n}x_jy_j=x_1y_1+x_2y_2+\cdots+x_ny_n.$$

PROOF: Identical to Theorem 7 of the last section.

$$\begin{aligned}\langle X,\, Y\rangle &= \langle\sum_{j=1}^{n}x_je_j,\quad \sum_{k=1}^{n}y_ke_k\rangle\\ &= \sum_{j=1}^{n}x_j\langle e_j,\, \sum_{k=1}^{n}y_ke_k\rangle\\ &= \sum_{j=1}^{n}x_j\left(\sum_{k=1}^{n}y_k\langle e_j,\, e_k\rangle\right)\\ &= \sum_{j=1}^{n}x_j\left(\sum_{k=1}^{n}y_k\delta_{jk}\right),\end{aligned}\tag{3-10}$$

so that

$$\langle X,\, Y\rangle=\sum_{j=1}^{n}x_jy_j.$$

Remark: We shall see that these two theorems extend to the case $n=\infty$.
Examples.

1. The vectors $e_1=(1,0,0)$, $e_2=(0,1,0)$, and $e_3(0,0,1)$ clearly form an orthonormal basis for $\mathbb{R}^3$. Let $X=(2,-1,4)$. We shall compute the x_j in

   $$X=\sum_{j=1}^{3}x_je_j.$$

   Since $x_j=\langle X,\, e_j\rangle$, we find $z_1=\langle X,\, e_1\rangle=\langle(2,-1,4),(1,0,\ 0)\rangle=2\cdot1+(-1)\cdot0+4\cdot0=2$, and similarly, $x_2=-1$, $x_3=4$ as expected. Thus

   $$(2,-1,4)=2e_1-e_2+4e_3.$$

   In the same way, if $Y=(7,1,-3)$, then

   $$Y=7e_1+e_2-3e_3.$$

   Also,

   $$\langle X,\, Y\rangle=(2)(7)+(-1)(1)+(4)(-3)=1.$$

   The projection of X into the subspace spanned by Y is

   $$\begin{aligned}\langle X,\, Y/\|Y\|\rangle\frac{Y}{\|Y\|} &= \frac{1}{59}(7,1,-3)\\ &= \frac{7}{59}e_1+\frac{1}{59}e_2-\frac{3}{59}e_3.\end{aligned}\tag{3-11}$$

   Another orthonormal basis for $\mathbb{R}^3$ is $\tilde{e}_1=(\frac{1}{\sqrt2},\frac{1}{\sqrt2},0)$, $\tilde{e}_2=(-\frac{1}{\sqrt2},\frac{1}{\sqrt2},0)$, and $\tilde{e}_3=(0,0,1)$, since $\langle\tilde{e}_j,\,\tilde{e}_k\rangle=\delta_{jk}$. The expansion for X in this basis is

   $$X=\sum_{j=1}^{3}\tilde{x}_j\tilde{e}_j,$$

   where

   $$\tilde{x}_1=\langle X,\,\tilde{e}_1\rangle=\langle(2,-1,4),\,(\frac{1}{\sqrt2},\frac{1}{\sqrt2},0)\rangle=\frac{1}{\sqrt2},\tag{3-12}$$

   $$\tilde{x}_2=-\frac{3}{\sqrt2},\ \text{ and }\tilde{x}_3=4.\tag{3-13}$$

   Thus

   $$X=\frac{1}{\sqrt2}\tilde{e}_1-\frac{3}{\sqrt2}\tilde{e}_2+4\tilde{e}_3.$$

   Similarly,

   $$Y=\frac{8}{\sqrt2}\tilde{e}_1-\frac{6}{\sqrt2}\tilde{e}_2-3\tilde{e}_3.$$

Therefore
$$\langle X, Y\rangle = (\frac{1}{\sqrt{2}})(\frac{8}{\sqrt{2}}) + (-\frac{3}{\sqrt{2}})(-\frac{6}{\sqrt{2}}) + (4)(-3) = 1.$$

Notice that the number $\langle X, Y\rangle$ is the same no matter which basis is used. This is *not* a coincidence. Recall that the scalar product $\langle X, Y\rangle$ was defined independently of any basis. Hence its value should not be dependent upon which basis we happen to choose. If you think of $\langle X, Y\rangle$ geometrically in terms of the projection, it should be clear that the number should not depend upon which particular basis is used to describe the vectors.

2. For our second example, we consider the set of orthonormal functions $e_1(x) = \frac{\sin x}{\sqrt{\pi}}$, $e_2(x) = \frac{\sin x}{\sqrt{\pi}}$ and let A be the set in $L_2(-\pi, \pi)$ which they span. We would like to expand some function
$$f(x) + \sum_{j=1}^{2} f_j e_j(x).$$

The only trouble is that Theorems 14 and 15 only allow us to expand functions f which are in the subspace A, that is, are a linear combination of the basis elements e_1 and e_2. Since we secretly know that $f(x) = \sin x \cos x (= \frac{1}{2}\sin 2x)$ is such a function, let us find its expansion. By elementary integration,
$$f_1 = \langle f, e_1\rangle = \int_{-\pi}^{\pi} (\sin x \cos x)\frac{\sin x}{\sqrt{x}}\, dx = 0,$$
and
$$f_2 = \langle f, e_2\rangle = \int_{-\pi}^{\pi} (\sin x \cos x)\frac{\sin 2x}{\sqrt{\pi}}\, dx = \frac{\sqrt{\pi}}{2}.$$
Therefore
$$f = 0 \cdot e_1 + \frac{\sqrt{\pi}}{2}e_2 = \frac{\sqrt{\pi}}{2}e_2$$
or
$$\sin x \cos x = \frac{\sqrt{\pi}}{2}(\frac{\sin 2x}{\sqrt{\pi}}) = \frac{\sin 2x}{2},$$
which we knew was the case from trigonometry.

If the orthonormal set $\{e_j\}$, $j = 1, \ldots, m$ spans a *subspace* A of a linear scalar product space H, and if $X \in H$, can any sense be made of the expansion
$$X \stackrel{?}{=} \sum_{j=1}^{m} x_j e_j?$$

One way to seek an answer is to examine a special case. Again geometry will supply the key. Let $H = \mathbb{R}^3$ and let A be the subspace spanned by the orthonormal vectors $e_1 = (1, 0, 0)$

and $e_2 = (0, 1, 0)$. Then if $X \in \mathbb{R}^3$, how can we interpret
$$X \stackrel{?}{=} \sum_{j=1}^{2} x_j e_j = x_1 e_1 + x_2 e_2?$$

Plowing blindly ahead, we take the scalar product of both sides with e_1 and then with e_2. This gives us $x_j = \langle X, e_j\rangle$. Thus the right side, $x_1 e_1 + x_2 e_2$, is the *projection* of X into the subspace A spanned by $\{ej\}$. It is now clear how our original quandary is resolved.

Definition If the orthonormal set $\{e_j\}$, $j = 1, \ldots, m$ spans a subspace A of a linear scalar product space H, and if $X \in H$, then the vector $\sum_{j=1}^{m} x_j e_j$, where $x_j = \langle X, e_j\rangle$, *is the projection of X into the subspace A.*

Remark. It is customary to denote the projection of X into A by $P_A X$. Think of P_A as an operator (function) which maps the vector X into its projection in A. With this notation the above definition reads
$$P_A X = \sum_{j=1}^{m} x_j e_j,$$
where $x_j = \langle X, e_j\rangle$ and the orthonormal set $\{e_j\}$ spans A.

Since the projection $P_A X$ is defined in terms of a particular basis for A, we should show that this geometrical object is independent of the basis you choose for A. But we shall not take the time right now. In reality, Theorem 17 below leads us to make a better definition of projection.

Theorem 3.18 . *If the orthonormal set $\{e_j\}$, $j = 1, \ldots, m$ spans a subspace $A \subset H$, and if X and Y are in H, then*
$$a) \quad \langle P_A X, P_A Y\rangle = \sum_{j=1}^{m} x_j y_j,$$
where $x_j = \langle X, e_j\rangle$ and $y_j = \langle Y, e_j\rangle$. In particular
$$b) \quad \|P_A X\| = \sqrt{\sum_{j=1}^{m} x_j^2}.$$

Furthermore, $X - P_A X \in A^{\perp}$, that is, for every $Y \in A$
$$c) \quad \langle X - P_A X, Y\rangle = 0$$
Every $X \in H$ can be written as
$$d) \quad X = P_A X + P_{A^{\perp}} X, \ \text{where } P_{A^{\perp}} X \equiv X - P_A X \text{ is in } A^{\perp}.$$

PROOF: Since both vectors $P_AX = \sum_{j=1}^{m} x_je_j$ and $P_AY = \sum_{j=1}^{m} y_je_j$ are in A itself, a) and b) are immediate consequences of Theorem 15. Although the equation c) is geometrically clear, we shall compute it too.

Since the e_j span A, this is equivalent to showing it is orthogonal to all the e_j. Now $\langle X-P_AX,\, e_j\rangle = \langle X,\, e_j\rangle - \langle P_AX,\, e_j\rangle = x_j - x_j = 0$. Since trivially $X = P_AX + (X-P_AX)$, the only content of part d) is that $(X - P_AX) \in A^\perp$, which is just what part c) proved.

Corollary 3.19

$$\textit{a)}\quad \left.\begin{array}{l} \|X\|^2 = \|P_AX\|^2 + \|X-P_AX\|^2 \\ \|X\|^2 = \|P_AX\|^2 + \|P_{A^\perp}X\|^2 \end{array}\right\} \qquad \textit{(Pythagorean Theorem)}$$

$$\textit{b)}\quad \|X\|^2 \ge \|P_AX\|^2 = \sum_{j=1}^{m} x_j^2 \qquad \textit{(Bessel's Inequality)}$$

PROOF: a) is a result of the fact that $P_AX \in A$ is orthogonal to $X - P_AX \in A^\perp$ and Theorem 12. The inequality b), Bessel's inequality, is simply a weaker form of a)—since $\|X-P_AX\| \ge 0$. There is equality if and only if $X \in A$, for only then does $\|X-P_AX\| = 0$.

Examples:

1. Let A be the subspace of $\mathbb{R}^3$ spanned by $e_1 = (1,0,0)$ and $e_2 = (0,1,0)$. The projection of $X = (3,-1,7)$ into A is represented by

   $$P_AX = \langle X,\, e_1\rangle e_1 + \langle X,\, e_2\rangle e_2 = 3e_1 - e_2 \in A$$

   Also
   $$P_{A^\perp}X = X - P_AX = 3e_1 - e_2 + 7e_3 - (3e_1 - e_2) = 7e_3 \in A^\perp.$$

   Since $\tilde{e}_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)$ and $\tilde{e}_2 = (-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)$ also form an orthonormal basis for A, we can equally well write

   $$P_AX = \langle X,\, \tilde{e}_1\rangle \tilde{e}_1 + \langle X,\, \tilde{e}_2\rangle \tilde{e}_2 = \frac{2}{\sqrt{2}}\tilde{e}_1 - \frac{4}{\sqrt{2}}\tilde{e}_2.$$

2. Let A be the subspace of $L_2[-\pi,\pi]$ spanned by the orthonormal functions $e_1(x) = \frac{\sin x}{\sqrt{\pi}}$, $e_2(x) = \frac{\sin 2x}{\sqrt{\pi}}$. The projection of the function $f(x) \equiv x$ into A is represented by

   $$P_Af = \langle f,\, e_1\rangle e_1 + \langle f,\, e_2\rangle e_2.$$

   Since an integration by parts shows that

   $$\int_{-\pi}^{\pi} x\sin kx\, dx = \frac{-x\cos kx}{k}\Big|_{-\pi}^{\pi} \cos kx\, dx$$

   $$= -\frac{x\cos kx}{k}\Big|_{-\pi}^{\pi} = -\frac{2\pi}{k}\cos k\pi = \left\{ \begin{array}{ll} \frac{2\pi}{k}, & k \text{ odd} \\ -\frac{2\pi}{k}, & k \text{ even} \end{array} \right\} = (-1)^{k+1}\frac{2\pi}{k},$$

   we find
   $$\langle f,\, e_1\rangle = \langle x,\, e_1\rangle = \int_{-\pi}^{\pi} x\frac{\sin x}{\sqrt{\pi}}\, dx = 2\sqrt{\pi}$$

   and
   $$\langle f,\, e_2\rangle = \langle x,\, e_2\rangle = \int_{-\pi}^{\pi} x\frac{\sin 2x}{\sqrt{\pi}}\, dx = -\sqrt{\pi}.$$

   Thus
   $$P_Ax = 2\sqrt{\pi}\frac{\sin x}{\sqrt{\pi}} - \sqrt{\pi}\frac{\sin 2x}{\sqrt{\pi}},$$

   or
   $$P_Ax = 2\sin x = \sin 2x.$$

   Also,
   $$\|P_AX\|^2 = \langle f,\, e_1\rangle^2 + \langle f,\, e_2\rangle^2 = 5\pi.$$

   More generally, we can let $\tilde{A}$ be the subspace of $L_2[-\pi,\pi]$ spanned by $\{e_k\}$, $k = 1, 2, \ldots, N$, where $e_k(x) = \frac{\sin kx}{\sqrt{\pi}}$. Then the projection of x onto $\tilde{A}$ is given by

   $$P_{\tilde{A}}x = \sum_{k=1}^{N} \langle x,\, e_k\rangle e_k(x).$$

Since
$$\langle x,\, e_k\rangle = \int_{-\pi}^{\pi} x\frac{\sin kx}{\sqrt{\pi}}\, dx = (-1)^{k+1}\frac{2\sqrt{\pi}}{k},$$

we have
$$\begin{aligned} P_{\tilde{A}}x &= \sum_{k=1}^{N} (-1)^{k+1}\frac{2\sqrt{\pi}}{k}\frac{\sin kx}{\sqrt{\pi}} \\ &= 2\sum_{k=1}^{N} \frac{(-1)^{k+1}}{k}\sin kx \\ &= 2(\sin x - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \ldots + (-1)^{N+1}\frac{\sin Nx}{N}). \end{aligned} \tag{3-14}$$

Furthermore,
$$\|P_{\tilde{A}}x\|^2 = \sum_{k=1}^{N} \langle x,\, e_k\rangle^2 = \sum_{k=1}^{N} \frac{4\pi}{k^2} = 4\pi\sum_{k=1}^{N}\frac{1}{k^2}.$$

It is from this formula that we eventually intend to obtain the famous formula

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots = \frac{\pi^2}{6}.$$

We will observe that
$$\|f\|^2 = \|X\|^2 = \int_{-\pi}^{\pi} x \cdot x\, dx = \frac{2\pi^3}{3}$$
and prove
$$\lim_{N\to\infty} \|P_{\tilde{A}^\perp} x\| = \lim_{N\to\infty} \|x - P_{\tilde{A}} x\| = 0$$
Then from the Corollary to Theorem 16,
$$\|X\|^2 = \lim_{N\to\infty} \|P_{\tilde{A}} x\|^2,$$
or
$$\frac{2\pi^3}{3} \overset{?}{=} 4\pi \sum_{k=1} \frac{1}{k^2} \Rightarrow \frac{\pi^2}{6} = \sum_{k=1}^{\infty} \frac{1}{k^2}.$$

Geometry leads us to the next theorem—and the proof too. Let X be a given vector and $P_A X$ its projection into the subspace A. Since distance is measured by dropping a perpendicular, we expect that $P_A X$ is the vector in A which is closest to X, that is, most closely approximates X.

Theorem 3.20 . *Let X be a vector in a scalar product space H and A a subspace of H. Then if V is any vector in A,*
$$\|X - P_A X\| \le \|X - V\|.$$

PROOF: We shall prove the stronger statement (cf. fig. above)
$$\|X - P_A X\|^2 + \|V - P_A X\|^2 = \|X - V\|^2.$$
Observe that $(V - P_A X) \in A$, since both terms are in A and A is a subspace. Moreover $X - P_A X \in A^\perp$ (Theorem 16c). Therefore $X - P_A X$ is orthogonal to $P_A X - V$, so the identity is a consequence of Theorem 12.

Remark. With this theorem in mind, we could *define* the projection $P_A X$ into a subspace A as the element in A which is closest to X. This definition is independent of any basis, whereas our original definition was not. One must, however, be somewhat careful when defining the projection into an infinite dimensional subspace. Although it is clear that the number $\|X - V\|$ has a g.l.b. as V wanders throughout A, it is *not* clear that it has an actual min, that is, there really is a vector $U \in A$ such that $\|X - U\|$ takes on its g.l.b. as a min. If there is such a U, we call it $P_A X$. Otherwise there *is no* projection. When projecting into a finite dimensional space this difficulty does not arise (but we will stop without further explanation of this detail).

Some discussion of these results is needed to place the material in its proper perspective. If you are given an orthonormal set of vectors $\{ e_j \}$ which span some subspace A of a scalar product space H, then for any X in H you can find a representation for $P_A X$ in terms of that basis, $P_A X = \sum x_j e_j$. If the vector X happened to already lie in A, then $P_A X = X$ so $X = \sum x_j e_j$ and $\|X\| = \sqrt{\sum x_j^2}$. This last equation for the length of X is the Pythagorean

Theorem. If X did not lie entirely in A, but "stuck out" of it into the rest of H, then $P_A X = \sum x_j e_j$ only represents a piece of X, its projection into A. Since part of X has been omitted, we expect that $\|X\| > \|P_A X\| = \sqrt{\sum x_j^2}$. This inequality was the content of the Corollary to Theorem 16. Informally, if no vector $X \in H$ sticks out of the linear space spanned by the $\{ e_j \}$, then the set $\{ e_j \}$ is said to be complete (do not confuse this with the complete of Chapter 0; they are entirely different concepts, an unfortunate coincidence). More precisely,

Definition An *orthonormal set is complete* for the scalar product space H if that orthonormal set is not properly contained in a larger orthonormal set.

There are many ways to check if a given orthonormal set is complete for H. Geometry suggests them all.

Theorem 3.21 . *Let $\{ e_j \}$ be an orthonormal set which spans the subspace A of the scalar product space H. The following statements are equivalent*

(a) *The set $\{ e_j \}$ is complete for H.*

(b) *If $\langle X, e_j \rangle = 0$ for all j, then $X = 0$.*

(c) *$A = H$.*

(d) *If $X \in H$, then $X = \sum x_j e_j$, where $x_j = \langle X, e_j \rangle$.*

(e) *If X and $Y \in H$, then $\langle X, Y \rangle = \sum x_j y_j$, where $x_j = \langle X, e_j \rangle$ and $y_j = \langle Y, e_j \rangle$*

(f) *If $X \in H$, then (Pythagorean Theorem) $\|X\|^2 = \sum x_j^2$, where $x_j = \langle X, e_j \rangle$*

PROOF: We shall use the chain of reasoning $a \Rightarrow b \Rightarrow c \ldots \Rightarrow f \Rightarrow a$.

$a \Rightarrow b$. If $\langle X, e_j \rangle = 0$ but $X \neq 0$, then $X/\|X\|$ is a unit vector orthogonal to all the e_j. This means that $\{ \frac{X}{\|X\|}, e_1, e_2, \ldots \}$ is an orthonormal set which contains $\{ e_1, e_2, \ldots \}$ as a proper subset.

$b \Rightarrow c$. If there is an $X \in H$ but $X \ni A$, then $P_{A^\perp} X = X - P_A X \in A^\perp$ and is not zero. Since all the $e_j \in A$, we have $\langle P_{A^\perp} X, e_j \rangle = 0$ for all j but $P_{A^\perp} X \neq 0$, contradicting b). Thus $H \subset A$. Since $A \subset H$ by hypothesis, this proves that $H = A$.

$c \Rightarrow d$. Since every $X \in A$ has the form $X = \sum x_j e_j$ (by Theorem 14) and since $H = A$, the conclusion is immediate.

$d \Rightarrow e \Rightarrow f$. A restatement of Theorem 16 since for every $X \in H$, we know that $P_A X = P_H X = X$.

$f \Rightarrow a$. If $\{ e_j \}$ is not complete, it is contained in a larger orthonormal set. Let e be a vector in that larger set which is not one of the e_j. Then by f), and the fact that $\langle e, e_j \rangle = 0$,
$$\|e\|^2 = \sum \langle e, e_j \rangle^2 = 0.$$
Therefore $e = 0$.

Remarks. 1. Because each of the six conditions a-f are equivalent, any one of them could have been used as the definition of a complete orthonormal set.

2. If the orthonormal set $\{e_j\}$ has a (countably) infinite number of elements, the theorem is still valid but some convergence questions for d-f arise because of the then *infinite* series $X = \sum_1^\infty x_j e_j$. The appropriate sense of convergence is that the remainder after N terms, $\sum_{N+1}^\infty x_j e_j = X - \sum_1^N x_j e_j$ tends to zero *in the norm of the scalar product space*, that is, if

$$\lim_{N\to\infty} \|X - \sum_1^N x_j e_j\| = 0.$$

We shall meet this in the next section for the space $L_2[-\pi,\pi]$. Condition f) gives us no convergence problems since the series is an infinite series of positive terms which is always bounded by $\|X\|^2$ (Bessel's Inequality—Corollary b to Theorem 16), and so always converges. This criterion just asks if the sum of the series actually equals $\|X\|^2$ (we know it is no larger).

Examples

1. The set of orthonormal vectors $e_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)$ and $e_2 = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0)$ are not complete for $\mathbb{R}^3$ since any basis for $\mathbb{R}^3$ must have three elements because its dimension is 3. This could also be seen geometrically from the fact that, for example $X = (1,2,3)$ sticks out of the space spanned by e_1 and e_2, or from the fact that $e_3 = (0,0,2)$ is a non-zero vector orthogonal to both e_1 and e_2, or in many other ways. The dimension argument is the easiest to apply *if H is finite dimension*, for *then the number of elements in a complete orthonormal set* $\{e_k\}$ *must equal the dimension of H.*

2. The set $\{\tilde{e}_n\}$ where $\tilde{e}_n(x) = \frac{\sin nx}{\sqrt{\pi}}$ is an orthonormal set of functions in the scalar product space $L_2[-\pi,\pi]$, but it is *not* a complete orthonormal set for that space since the function $\cos x$ is a non-zero function in $L_2[-\pi,\pi]$ which is orthogonal to all the $\tilde{e}_n$,

$$\langle \cos x, \tilde{e}_n\rangle = \int_{-\pi}^{\pi} \cos \frac{\sin nx}{\sqrt{\pi}}\, dx = 0.$$

   Thus, although the set $\{\tilde{e}_n\}$ has an infinite number of elements, it is still not big enough to span all of $L_2[-\pi,\pi]$. The next section will be devoted to proving that the larger orthonormal set, $e_0, e_1, \tilde{e}_1, e_2, \tilde{e}_2, \ldots$, where

$$e_0 = \frac{1}{\sqrt{2\pi}},\ e_n(x) = \frac{\cos nx}{\sqrt{\pi}},\ \tilde{e}_n(x) = \frac{\sin nx}{\sqrt{\pi}}$$

   is a complete orthonormal set for the scalar product space $L_2[-\pi,\pi]$. This is a difficult theorem.

Specific applications of the ideas in this section are contained in the exercises. For many of them you would be wise if you referred to their corresponding special cases which appeared in Section 2.

Exercises

1. Let X and Y be points in $\mathbb{R}^n$. Determine which of the following make $\mathbb{R}^n$ into a scalar product space, and why—or why not.

   a) $\langle X, Y\rangle = \sum_{k=1}^n \frac{1}{k} x_k y_k$.

   b) $\langle X, Y\rangle = \sum_{k=1}^n (-1)^k x_k y_k$.

   c) $\langle X, Y\rangle = \sqrt{\sum_{k=1}^n x_k^2 y_k^2}$.

   d) $\langle X, Y\rangle = \sum_{k=1}^n a_k x_k y_k$, where $a_k > 0$ for all k.

2. Let f and g be continuous real-valued functions in the interval $[0,1]$, so $f, g \in C[0,1]$. Determine which of the following make $C[0,1]$ into a scalar product space, and why—or why not.

   a) $\langle f, g\rangle = \int_0^1 f(x)g(x)\frac{1}{1+x^2}\, dx$.

   b) $\langle f, g\rangle = \int_0^1 f(x)g(x)\sin 2\pi x\, dx$.

   c) $\langle f, g\rangle = \int_0^1 f(x)g^2(x)\, dx$

   d) $\langle f, g\rangle = \int_0^1 f(x)g(x)\rho(x)\, dx$, where $\rho(x)$ is a fixed continuous function with the property $\rho(x) > 0$.

   e) $\langle f, g\rangle = f(0)g(0)$.

3. This is the analogue of L_2 for sequences. Let l_2 be the set of all sequences $X = (x_1, x_2, x_e, \ldots)$ with the property that $\|X\| = \sqrt{\sum_{j=1}^\infty x_j^2} < \infty$. Prove that l_2 is a normed linear space (cf. the example for l_1 in Section 1).

4. Use the Cauchy-Schwarz inequality to prove that if $\sum_{n=1}^{\infty} n^2 a_n^2 < \infty$, then $\sum_{n=1}^{\infty} |a_n| < \infty$. (Hint: $|a_n| = \frac{1}{n}|na_n|$).

5. Consider the following linearly independent vectors in $\mathbb{R}^3$:

$$X_1 = (1, 0, -1), \quad X_2 = (0, 3, 1), \quad X_3 = (2, -1, 0).$$

a) Use the Gram-Schmidt orthogonalization process to find an orthonormal set of vectors, e_1, e_2 and e_3 such that e_1 is in the subspace spanned by X_1.

b) Write $X = (1, 2, 3)$ as $X = \sum_{j=1}^{3} x_j e_j$, where the e_j are those of part a). Also, compute $\|X\|$ and $\|PX\|$.

6. Consider the following linearly independent set of functions in $L_2[-1, 1]$

$$f_1(x) = 1, \quad f_2(x) = x, \quad f_3(x) = x^2.$$

a) Use the Gram-Schmidt orthogonalization process to find an orthonormal set of functions $e_1(x), e_2(x)$ and $e_3(x)$ such that e_1 is in the subspace spanned by f_1.

b) Find the projection of the function $f(x) = (1 + x)^3$ into the subspace of $L_2[-1, 1]$ spanned by $e_1(x), e_2(x)$, and $e_3(x)$. Also, compute $\|f\|$ and $\|Pf\|$.

7. Let $P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(1 - x^2)^n, n = 0, 1, 2, \ldots$. These are the *Legendre Polynomials.*

a) Prove that $\langle P_n, P_m\rangle = \int_{-1}^{1} P_n(x)P_m(x)\,dx = 0, \quad n \neq m$, that is, the P_n are orthogonal in $L_2[-1, 1]$ by first proving that

$$\int_{-1}^{1} P_n(x)x^m\,dx = 0, \; m < n.$$

b) Show that $\|P_n\|^2 = \frac{2}{2n+1}$. Thus the functions

$$e_n(x) = \sqrt{\frac{2n+1}{2}}P_n(x)$$

are an orthonormal set of functions for $L_2[-1, 1]$. Compute $e_0(x), e_1(x)$, and $e_2(x)$ and compare with Exercise 6a.

8. a) Show that the vector $N = (a_1, a_2, a_3)$ is orthogonal to the coset (a plane in $\mathbb{R}^3$) $A = \{ X \in \mathbb{R}^3 \colon a_1x_1 + a_2x_2 + a_3x_3 = c \}$.

b) Show that the vector $N = (a_1, \ldots, a_n)$ is orthogonal to the coset (a hyperplane in $\mathbb{R}^n$) $A = \{ X \in \mathbb{R}^n \colon a_1x_1 + \ldots a_nx_n = c \}$.

c) Find the coset $A \subset \mathbb{R}^3$ which passes through the point $X_0 = (1, -1, 2)$ and is orthogonal to $N = (1, 3, 2)$. In ordinary language, A is the plane containing the point X_0 which is orthogonal to N.

d) Show that the coset $A \subset \mathbb{R}^n$ which passes through the point $X_0 = (\tilde{x}_1, \ldots, \tilde{x}_n)$ and is orthogonal to $N = (a_1, \ldots, a_n)$ is

$$A = \{ X \in \mathbb{R}^n \colon \langle X, N\rangle = \langle X_0, N\rangle \}.$$

9. a) Use Problem 8a to show that the distance d from the point $P = (y_1, y_2, y_3) \in \mathbb{R}^3$ to the coset $a_1x_1 + a_2x_2 + a_3x_3 = c$ in $\mathbb{R}^3$ is

$$d = \frac{|a_1y_1 + a_2y_2 + a_3y_3 - c|}{\sqrt{a_a^2 + a_2^2 + a_3^2}} = \frac{|\langle N, P\rangle - c|}{\|N\|}$$

b) Show that the distance d from the point $P = (y_1, \ldots, y_n) \in \mathbb{R}^n$ to the coset $a_1x_1 + \ldots + a_nx_n = c$ in $\mathbb{R}^n$ is

$$d = \frac{|a_1y_1 + a_2y_2 + \cdots + a_ny_n - c|}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}} = \frac{|\langle N, P\rangle - c|}{\|N\|}.$$

c) Show that the distance d between the "parallel" cosets $a_1x_1 + \cdots + a_nx_n = c_1$ and $a_1x_1 + \cdots + a_nx_n = c_2$ in $\mathbb{R}^n$ is

$$d = \frac{|c_1 - c_2|}{\sqrt{a_1^2 + \cdots + a_n^2}} = \frac{|c_1 - c_2|}{\|N\|}.$$

(Hint: Pick a point P in one of the cosets and apply part b).

10. Find the angle between the diagonal of a cube and one of its edges.

11. Let Y_1 and Y_2 be fixed vectors in a scalar product space H.

a). If $\langle X, Y_1\rangle = 0$ for all $X \in H$, prove that $Y_1 = 0$.

b). If $\langle X, Y_1\rangle = \langle X, Y_2\rangle$ for all $X \in H$, prove that $Y_1 = Y_2$.

12. Let Y_0 be a fixed vector in a scalar product space H. Let $A = \{ X \in H \colon \langle Y, Y_0\rangle = 0 Rightarrow \langle X, Y\rangle = 0 \}$. Prove that A is the span of $Y_0 : \{ X \in A Rightarrow X = cY_0 \}$ for some scalar c. Make sure to see the geometrical situation for the case $H = \mathbb{R}^3$. [*Hint*: Let B be the set of all vectors orthogonal to Y_0, so $Y \in B$. Since H is composed of two parts, Y_0 and B, every $X \in H$ can be written as $X = cY_0 + Z$, where cY_0 is the projection of X into the subspace spanned by Y_0 (so $c = \langle X, Y_0\rangle/\|Y_0\|^2$) and $Z = (X - cY_0) \in B$. Now show that $X \in A \Rightarrow Z = 0$).

13. a) Let $X = (1, 3, -1)$ and $Y = (2, 1, 1)$. Find a vector N which is orthogonal to the subspace spanned by X and Y.
    b) Let $X = (x_1, x_2, x_3)$ and $Y = (y_1, y_2, y_3)$. Find a vector N which is orthogonal to the subspace spanned by X and Y. [Answer. $N = c(x_2y_3 - y_2x_3, y_1x_3 - x_1y_3, x_1y_2 - y_1x_2)$, where c is any non-zero scalar].

14. Let A be the subspace of $L_2[-\pi, \pi]$ spanned by the orthonormal set $\{e_n(x)\}$, $n = 1, 2, \ldots, N$, where $e_n(x) = \frac{\sin nx}{\sqrt{\pi}}$.
    a) Find the projection of $f(x) = x^2$, into A. (The answer should surprise you). Compute $\|f\|$ and $\|P_A f\|$ too.
    b) Find the projection of $f(x) = 1 + \sin^3 x$ into A. Compute $\|f\|$ and $\|P_A f\|$.
    c) If $f(x)$ is an even function, $f(x) = f(-x)$, show that its projection into A is zero. Now look at part (a) again.

15. a) If $f \in C[a, b]$, show that

$$\left(\int_a^b f(x)\, dx\right)^2 \le (b-a) \int_a^b f^2(x)\, dx.$$

    [Hint: Write $f(x) = 1 \cdot f(x)$ and use the Cauchy-Schwarz inequality for $L_2[a, b]$].
    b) If $f \in C^1[a, b]$, prove that

$$|f(x) - f(a)|^2 \le (x - a) \int_a^b f'(x)^2\, dx,\ x \in (a, b).$$

    [Hint: Write $f(x) - f(a) = \int_a^x f'(t)\, dt$ and apply part a)]
    c) If $f \in C^1[a, b]$ and $f(a) = 0$, use part b to prove that

$$\int_a^b f^2(x)\, dx \le \frac{(b-a)^2}{2} \int_a^b f'(x)^2\, dx.$$

16. a) Let $A = \{h \in C^1[a, b] \colon h(a) = h(b)\}$ and let $B = \{h \in C^1[a, b] \colon \langle 1, h' \rangle = 0\}$, where $h' = \frac{dh}{dx}$. Show that the subspaces A and B are identical $h \in A \Longleftrightarrow h \in B$.
    b) Let $f(x)$ be any continuous function such that $\int_a^b f(x)h'(x)\, dx = 0$ for all $h(x) \in C^1[a, b]$ with $h(a) = h(b)$. Show that $f \equiv$ constant. [Hint: Use part (a) and the result of Exercise 12].

17. If $f(x) \in C[a, b]$ and satisfies the condition $\int_a^b f(x)h(x)\, dx = 0$ for all $h(x) \in C[a, b]$ which satisfy the conditions

$$\int_a^b h(x)\, dx = 0,\ \int_a^b x h(x)\, dx = 0, \ldots,\ \int_a^b x^n h(x)\, dx = 0,$$

prove that $f \in \mathcal{P}_n$, that is, f is of the form

$$f(x) = a_0 + a_1 x + \cdots + a_n x^n,$$

where the a_j are constants. [Hint: Use Exercise 12].

18. Determine which of the following orthonormal sets are complete for their respective spaces.
    a) In $\mathbb{R}^3$, $e_1 = (0, 1, 0)$, $e_2 = (\frac{3}{5}, 0, \frac{4}{5})$, $e_3 = (-\frac{4}{5}, 0, \frac{3}{5})$
    b) In $\mathbb{R}^4$, $e_1 = (1, 0, 0, 0)$, $e_x = (0, 1, 0, 0)$, $e_3 = (0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$.
    c) In $\mathbb{R}^4$, e_1, e_2, e_3 as in (b), and $e_4 = (0, 0, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$.

19. Let e_1, e_2, and e_3 be an orthonormal basis for $\mathbb{R}^3$, and let A be the subspace spanned by $X_1 = 3e_1 - 4e_3$. Find an orthonormal basis for $A^\perp$.

20. Let A be a subspace of a scalar product space H. If $X \in H$, prove that $P_A(P_A X) = P_A X$ and interpret this geometrically. This result can be written as $P_A^2 = P_A$.

21. Let A be any operator (not necessarily linear) on a scalar product space. Prove the *polarization identity*

$$2\langle AX,\, AY \rangle = \|AX + AY\|^2 - \|AX\|^2 - \|AY\|^2.$$

3.4 Fourier Series.

Throughout this section we shall only use the scalar product of $L_2[a, b]$,

$$\langle f,\, g \rangle = \int_a^b f(x) g(x)\, dx.$$

We begin with the observation that in the interval $[-\pi, \pi]$

$$\langle \sin nx,\, \sin mx \rangle = \int_{-\pi}^{\pi} \sin nx \sin mx\, dx = \pi \delta_{nm}, \tag{3-15}$$

$$\langle \sin nx,\, \cos mx \rangle = \int_{-\pi}^{\pi} \sin nx \cos mx\, dx = 0, \tag{3-16}$$

and

$$\langle \cos nx,\, \cos mx \rangle = \int_{-\pi}^{\pi} \cos nx \cos mx\, dx = \pi \delta_{nm}, \tag{3-17}$$

where $n,\, m = 0, 1, 2, 3, \ldots$. Thus the functions

$$e_0(x) = \frac{1}{\sqrt{2\pi}},\; e_n(x) = \frac{\cos nx}{\sqrt{\pi}},\; \tilde{c}_n(x) = \frac{\sin nx}{\sqrt{\pi}}$$

form an orthonormal set:

$$\langle e_n,\, \tilde{e}_m\rangle = \delta_{nm}\langle e_n,\, \tilde{e}_m\rangle = 0,\; \langle \tilde{e}_n,\, \tilde{e}_m\rangle = \delta_{nm}.$$

Thus, if $f \in L_x[-\pi,\pi]$, we can find the projection $P_N f$ of f into the subspace spanned by $e_0, e_1, \tilde{e}, \ldots, e_N, \tilde{e}_N$.

$$(P_N f) = a_0 e_0 + \sum_{n=1}^{N} a_n e_n + b_n \tilde{e}_n, \tag{3-18}$$

where

$$a_k = \langle f,\, e_k\rangle \quad \text{and} \quad b_k = \langle f,\, \tilde{e}_k\rangle \tag{3-19}$$

More explicitly,

$$(P_N f)(x) = a_0 \frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{N} a_n \frac{\cos nx}{\sqrt{\pi}} + b_n \frac{\sin nx}{\sqrt{\pi}}, \tag{3-20}$$

where

$$a_0 = \int_{-\pi}^{\pi} f(x) \cdot \frac{1}{\sqrt{2\pi}}\, dx,$$

and

$$a_n = \int_{-\pi}^{\pi} f(x) \frac{\cos nx}{\sqrt{\pi}}\, dx, \quad b_n = \int_{-\pi}^{\pi} f(x) \frac{\sin nx}{\sqrt{\pi}}\, dx \tag{3-21}$$

A natural question arises: as $N \to \infty$, does the series converge: $P_N f \to f$, in the sense that $\|f - P_N f\| \to 0$? In other words, is the set $\{\, e_j(x), \tilde{e}_j(x)\,\}$, $j = 0, 1, 2, \ldots$ a *complete* orthonormal set of functions for $L_2[-\pi,\pi]$? The answer is yes, as we shall prove. Thus for any $f \in L_2[-\pi,\pi]$,

$$f(x) = a_0 \frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{\infty} a_n \frac{\cos nx}{\sqrt{\pi}} + b_n \frac{\sin nx}{\sqrt{\pi}}, \tag{3-22}$$

where the *Fourier coefficients*, a_n, b_n are determined by the formulas (2). The expansion (3) is called the *Fourier series for* f.

Historically, Fourier series did not arise from the geometrical considerations we have developed. Mathematical physics—in particular the vibrations of strings and the flow of heat in a bar—take the credit for these ideas. Only in recent years has the geometrical viewpoint been investigated. Later on we shall discuss some of the fascinating problems in mathematical physics to which Fourier series can be applied.

Beware. The equality which appears in (3) is equality in the $L_2[-\pi,\pi]$ norm, viz.

$$\|f - P_N f\| = \sqrt{\int_a^b [f(x) - (P_N f)(x)]^2\, dx} \to 0$$

This is quite different than the convergence of infinite series to which you're accustomed, which is the uniform norm

$$\|f - P_N f\|_\infty = \max_{-\pi \le x \le \pi} |f(x) - (P_N f)(x)|\,.$$

In Section 1 (p. 176) you saw one instance of where a sequence of functions converged in some norm (the L_1 norm there) but did not converge in the uniform norm. Such is also the case here. In fact, contrasting the situation in the L_2 norm, there do exist continuous functions f whose Fourier series (3) does not converge to f in the uniform norm. However if the function f has one derivative, then its Fourier series does converge to f in the uniform norm.

In addition, there are some *discontinuous* functions whose Fourier series converge. These ideas will become clearer later on.

You should be warned that our definition (1), (3) of a Fourier series is not the standard one. Most books do *not* work with the *orthonormal* set $e_0 = \frac{1}{\sqrt{2\pi}}$, $e_n = \frac{\cos nx}{\sqrt{\pi}}$, $\tilde{e}_n = \frac{\sin nx}{\sqrt{\pi}}$, but rather use just an *orthogonal* set which is *not* normalized $\theta_0 = \frac{1}{2}$, $\theta_n = \cos nx$, $\tilde{\theta}_n = \sin nx$. For these people,

$$f(x) = \frac{A_0}{2} + \sum_{n=1}^{\infty} A_n \cos nx + B_n \sin nx,$$

where

$$A_n = \int_{-\pi}^{\pi} f(x) \frac{\cos nx}{\pi}\, dx, \quad B_n = \int_{-\pi}^{\pi} f(x) \frac{\sin nx}{\pi}\, dx,$$

$n = 0, 1, 2 \ldots$. As you can see, these differ from our formulas only by factors of $\sqrt{\pi}$. Needless to say, the resulting Fourier series for a given function f does not depend which intermediate formulas you use. We prefer the less standard ones because they are more intimately tied to geometry (so there is less to remember).

Before discussing the difficult issues of convergence in detail, we will find the Fourier series associated with some specific functions.

Examples.

1. Find the Fourier series associated with the functions $f(x) = x$, $-\pi \le x \le \pi$. We actually found this in the previous section. A computation (involving integration by parts) shows that

$$a_0 = \langle f,\, e_0\rangle = \int_{-\pi}^{\pi} x \cdot \frac{1}{\sqrt{2\pi}}\, dx = 0$$

$$a_n = \langle f,\, e_n\rangle = \int_{-\pi}^{\pi} x \frac{\cos nx}{\sqrt{\pi}}\, dx = 0, \qquad n = 1, 2, \ldots$$

$$b_n = \langle f,\, \tilde{e}_n\rangle = \int_{-\pi}^{\pi} x \frac{\sin nx}{\sqrt{\pi}}\, dx = \frac{2(-1)^{n+1}\sqrt{\pi}}{n}, \qquad n = 1, 2, \ldots$$

Thus, upon substituting into (3) we find that

$$x = \sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n}\sqrt{\pi}\,\frac{\sin nx}{\sqrt{\pi}} = 2\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}\sin nx$$

or

$$x = 2[\sin x - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \frac{\sin 4x}{4} + \cdots].$$

Again we remind you that the equality here is in the sense of convergence in L_2. For this particular function, there is also equality in the usual sense of convergence for infinite series for all $x \in (-\pi, \pi)$. Direct substitution reveals that it does *not* converge in the usual sense at $x = \pm\pi$. These remarks are based upon convergence theorems we have yet to prove. At $x = \frac{\pi}{2}$, this yields

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$$

2. Since the formulas (2)' make sense even if the function $f(x)$ has a finite number of discontinuities, we are tempted to find the Fourier series for discontinuous functions (in contrast, recall that the coefficients of an infinite *power series* are only defined if the function had an infinite number of derivatives). We shall find the Fourier series associated with the discontinuous function

$$f(x) = \begin{cases} 0, & -\pi \le x \le 0 \\ \pi, & 0 < x < \pi \end{cases}$$

The computations are particularly simple.

$$a_0 = \langle f, e_0\rangle = \int_{-\pi}^{0} 0\cdot\frac{1}{\sqrt{2\pi}}\,dx + \int_0^{\pi}\pi\cdot\frac{1}{\sqrt{2\pi}}\,dx = \frac{\pi^2}{\sqrt{2\pi}}, \qquad (3\text{-}23)$$

$$a_n = \langle f, e_n\rangle = \int_{-\pi}^{0} 0\cdot\frac{\cos nx}{\sqrt{\pi}}\,dx + \int_0^{\pi}\pi\cdot\frac{\cos nx}{\sqrt{\pi}}\,dx = 0, \quad n > 0, \qquad (3\text{-}24)$$

$$b_n = \langle f, \tilde{e}_n\rangle = \int_{-\pi}^{0} 0\cdot\frac{\sin nx}{\sqrt{\pi}}\,dx + \int_0^{\pi}\pi\cdot\frac{\sin nx}{\sqrt{\pi}}\,dx \qquad (3\text{-}25)$$

$$= \frac{\sqrt{\pi}}{n}(1 - \cos n\pi) = \begin{cases} \frac{2\sqrt{\pi}}{n} & , \ n \text{ odd} \\ 0 & , \ n \text{ even} \end{cases} \qquad (3\text{-}26)$$

Therefore the Fourier series associated with this function is

$$f(x) = \frac{\pi^2}{\sqrt{2\pi}}\cdot\frac{1}{\sqrt{2\pi}} + 2\sqrt{\pi}\left(\frac{\sin x}{\sqrt{\pi}} + \frac{\sin 3x}{3\sqrt{\pi}} + \frac{\sin 5x}{5\sqrt{\pi}} + \cdots\right),$$

or

$$f(x) = \frac{\pi}{2} + 2(\sin x + \frac{\sin 3x}{3} + \frac{\sin 5x}{5} + \cdots).$$

As usual, the equality is meant in the sense of convergence in the L_2 norm. The series also converges to the function f in the uniform norm in the whole interval except for a neighborhood of $x = 0$. At 0 it hasn't got a chance because of the discontinuity of f there. A glance at the series reveals that at $x = 0$, the right side is $\pi/2$—the arithmetic mean between the values of f just to the left and right of 0. This is the usual case at a discontinuity: *a Fourier series converges to the average of the function values to the right and left of the point where f is discontinuous.* We still offer no proof for these statements.

Observe that the Fourier series (3) for any function $f(x)$ depends only upon the values of x in the interval $-\pi \le x \le \pi$. However the series itself is periodic with period 2π. If the function $f(x)$, which we considered only for $x \in [-\pi, \pi]$ is defined for all other x by the formula $f(x+2\pi) = f(x)$ (making f periodic too), then both sides of the Fourier series (3) are periodic with period 2π. Therefore whatever they do in the interval $[-\pi, \pi]$ is repeated every 2π.

For example, the function $f(x) = x$, $x \in [-\pi, \pi]$ when continued outside the interval $[-\pi, \pi]$ as a function periodic with period 2π becomes

A FIGURE GOES HERE

Since the Fourier series for this particular function converges uniformly for all $x \in (-\pi, \pi)$, it also converges uniformly to the periodically continued function for all $x \in (k\pi, k\pi+2\pi)$, $k = 1, \pm1, \pm2, \ldots$. This also makes it clear why the Fourier series for $f(x) = x$ converges to zero at $x = \pm\pi$, for the series is just converging to the arithmetic mean of its neighboring values at the discontinuity.

It is pleasant to look at a picture. Let us see how the first four terms of its Fourier series approximates the function x

$$x = 2(\sin x - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \frac{\sin 4x}{4} + \cdots)$$

A FIGURE GOES HERE

Notice that as more terms are used, the projection $P_N x$ $P_N x = 2(\sin x - \frac{\sin 2x}{2} + \cdots + (-1)^{N+1}\frac{\sin Nx}{N})$ more and more closely approximate x. This reflects the convergence of the Fourier series, $P_N f \to f$.

One popular interpretation of a Fourier series is as a sum of "waves" which approximate a given function. Thus the function x is the sum of 2 times the wave $\sin x$ plus (-1) times the wave $\sin 2x$ and so on. In other words, the Fourier series for the function $f(x) = x$ represents that function as the *superposition* of sine waves. The term $2\sin x$ is spoken of as the first *harmonic*, the term $-\sin 2x$ as the second harmonic, the term $\frac{2}{3}\sin 3x$ as the third harmonic, etc.

Although it is difficult to believe, the ear hears by taking the sound wave $f(x)$ which impinges on the ear drum and splitting it up into its Fourier components (3). It then analyzes each component $a_n e_n + b_n\tilde{e}_n$—only considering the coefficients a_n and b_n. These Fourier coefficients measure the intensity of the nth harmonic. Particular sounds are then heard in terms of the intensity of their various harmonics. We recognize familiar sounds by recognizing that the sound waves have similar Fourier coefficients. Amazing.

It is time to consider the convergence of Fourier series. The question is: does the partial Fourier series

$$P_N f = a_0 e_0 + \sum_{n=0}^{N} a_n e_n + b_n \tilde{e}_n$$

converge to the function f as $N \to \infty$. Since there are several norms, in particular the L_2 norm $\| \quad \|$ and the uniform norm $\| \quad \|_\infty$, we must investigate convergence in each norm. Even though our proofs are reasonably slick, they are neither short nor particularly simple. A great deal of analytical technique will be needed. The proofs to be presented have been chosen because each of the devices invoked are important devices in their own right.

We begin with some useful facts which have nothing especially to do with Fourier series.

Theorem 3.22 *(Weierstrass Approximation Theorem). If $f(x)$ is continuous in the interval $[-\pi, \pi]$ and $f(-\pi) = f(\pi)$, then given any $\epsilon > 0$ there is a trigonometric polynomial*

$$\begin{aligned} T_N(x) &= \alpha_0 + \sum_{n=1}^{N} \alpha_n \cos nx + \beta_n \sin nx \\ &= \hat{\alpha}_0 e_0 + \sum_{n=1}^{N} \hat{\alpha}_n e_n + \hat{\beta}_n \tilde{e}_n, \end{aligned} \tag{3-27}$$

(where $\alpha_0 = \hat{\alpha}_0 \sqrt{2\pi}$, $\alpha_n = \hat{\alpha}_n \sqrt{\pi}$, $\beta_n = \hat{\beta}_n \sqrt{\pi}$), such that

$$\|f - T_N\|_\infty = \max_{-\pi \le x \le \pi} |f(x) - T_N(x)| < \epsilon.$$

Note that the numbers α_n and β_n are *not* necessarily the Fourier coefficients of f. The proof, which is placed as an appendix at the end of this section, will indicate how they can be found.

The following theorem states that convergence in the uniform norm implies convergence in the L_2 norm.

Theorem 3.23 . *If $\theta(x)$ is any bounded integrable function, then (if $b > a$)*

$$\|\theta\| \le \sqrt{b-a}\, \|\theta\|_\infty.$$

Proof: Since $\|\theta\|_\infty = \max_{x \in [a,b]} |\theta(x)|$ we find immediately that

$$\int_a^b \theta(x)^2\, dx \le \int_a^b \|\theta\|_\infty^2\, dx = \|\theta\|_\infty^2 \int_a^b dx = (b-a)\|\theta\|_\infty^2$$

from which the conclusion is obvious. On geometrical grounds the theorem is even easier, since $\|\theta\|_\infty$ is the greatest height of the curve $\theta(x)$.

Although convergence in the L_2 norm does *not* imply convergence in the uniform norm (the example in Section 1 comparing L_1 convergence and uniform convergence also works for L_2), a useful weaker statement is true.

Theorem 3.24 . *(cf. Ex. 15 Section 3). If $\theta \in C^1[a, b]$ and $\theta(x_0) = 0$, where $x_0 \in [a, b]$, then for every $x \in [a, b]$*

$$|\theta(x)| \le \sqrt{b-a} \sqrt{\int_a^b \theta'\, dt} = \sqrt{b-a}\, \|\theta'\|.$$

Since the right side is independent of x, this implies that

$$\|\theta\|_\infty = max_{x \in [a,b]} |\theta(x)| \le \sqrt{b-a}\, \|\theta'\| = \sqrt{b-a} \|D\theta\|$$

Proof: By the fundamental theorem of calculus,

$$\theta(x) = \theta(x) - \theta(x_0) = \int_{x_0}^{x} \theta'(t)\, dt.$$

Thus the Cauchy-Schwarz inequality yields

$$\begin{aligned} |\theta(x)|^2 &= \left(\int_{x_0}^{x} 1 \cdot \theta'(t)\, dt \right)^2 \le \int_{x_0}^{x} 1^2\, dt \quad \int_{x_0}^{x} \theta'(t)^2\, dt \\ &= (x - x_0) \int_{x_0}^{x} \theta'(t)^2\, dt \\ &\le (b-a) \int_a^b \theta'(t)^2\, dt. \end{aligned} \tag{3-28}$$

Therefore

$$|\theta(x)|^2 = (b-a)\|\theta'\|^2.$$

With these preliminaries behind us we turn to the convergence of Fourier series. First up is convergence in the L_2 norm.

Theorem 3.25 . *Assume f is continuous in the interval $[-\pi, \pi]$ and $f(-\pi) = f(\pi)$. Denote the sum of the first N terms of its Fourier series by $P_N f$. Then*

$$\lim_{N \to \infty} \|f - P_N f\| = \lim_{N \to \infty} \sqrt{\int_{-\pi}^{\pi} [f(x) - (P_N f)(x)]^2\, dx} = 0$$

Proof: Given any $\epsilon > 0$, let $T_N(x)$ be the trigonometric polynomial given by Weierstrass Approximation Theorem. The trick is to apply Theorem 17. Using the N of T_N, we know that

$$P_N f = a_0 e_0 + \sum_{n=1}^{N} a_n e_n + b_n \tilde{e}_n$$

and

$$T_N = \hat{a}_0 e_0 + \sum_{n=1}^{N} \hat{\alpha}_n e_n + \hat{\beta}_n \tilde{e}_n.$$

Let A be the subspace of $H = L_2[-\pi,\pi]$ spanned by $e_0, e_1, \tilde{e}_1, \dots e_N, \tilde{e}_N$ Then both $P_N f$ and T_N are in A. Thus by Theorem 17 of the last section (where slightly different notation was used),

$$\|f - P_N f\| \le \|f - T_N\|,$$

and by Theorem 20

$$\le \sqrt{b-a}\,\|f - T_N\|_\infty < \sqrt{b-a}\,\epsilon.$$

Thus

$$\lim_{N\to\infty} \|f - P_N f\| = 0,$$

proving the theorem.

Corollary 3.26 *(Parseval's Theorem). If $f(x)$ is continuous in the interval $[-\pi,\pi]$ and $f(-\pi) = f(\pi)$, then*

$$\|f\|^2 = \lim_{N\to\infty} \|P_N f\|^2,$$

that is,

$$\int_{-\pi}^{\pi} f^2(x)\,dx = a_0^2 + \sum_{n=1}^{\infty}(a_n^2 + b_n^2),$$

where the Fourier coefficients a_j and b_j are determined by equations (2) or (2)'.

PROOF: The Corollary to Theorem 16 states that

$$\|f\|^2 = \|P_N f\|^2 + \|f - P_N f\|^2.$$

If we now let $N \to \infty$, the second term on the right vanishes by the theorem just proved. *Remark*: The theorem and corollary state that the orthonormal set of functions $e_0 = \frac{1}{\sqrt{2\pi}}$, $e_n(x) = \frac{\cos nx}{\sqrt{\pi}}$, and $\tilde{e}_n(x) = \frac{\sin nx}{\sqrt{\pi}}$ is a complete orthonormal set for the scalar product space $L_2[-\pi,\pi]$. The formula contained in the corollary is a generalization of the Pythagorean Theorem to $L_2[-\pi,\pi]$.

The proof of convergence in the uniform norm if the function has one continuous derivative is only slightly more difficult. We shall need a preliminary

Lemma 3.27 . *Assume $f \in C^1[-\pi,\pi]$. Extend it as a periodic function with period 2π by $f(x+2\pi) = f(x)$. Let $(P_N f)$ be the sum of the first N terms of its Fourier series. Then the sum of the first N terms in the Fourier series for $Df = \frac{df}{dx}$ is $P_N(Df)$, that is*

$$P_N(Df) = D(P_N f).$$

This in not necessarily true for other bases in $L_2[-\pi,\pi]$.

PROOF: We know that

$$(P_N f)(x) = a_0\frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{N} a_n\frac{\cos nx}{\sqrt{\pi}} + b_n\frac{\sin nx}{\sqrt{\pi}}.$$

Since we can differentiate a *finite* sum term by term, we find that

$$D(P_N f)(x) = \sum_{n=1}^{N} -na_n\frac{\sin nx}{\sqrt{\pi}} + nb_n\frac{\cos nx}{\sqrt{\pi}},$$

where the a_n and b_n are found by using formulas (2)'. If

$$P_N(Df) = A_0\frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{N} A_n\frac{\cos nx}{\sqrt{\pi}} + B_n\frac{\sin nx}{\sqrt{\pi}},$$

where the A_n and B_n are also found by using (2)', we must show that

$$A_0 = 0, \quad A_n = nb_n, \text{ and } B_n = -na_n$$

But

$$A_0 = \int_{-\pi}^{\pi}(Df(x))\frac{1}{\sqrt{2\pi}}\,dx = \frac{1}{\sqrt{2\pi}}[f(\pi) - f(-\pi)] = 0 \quad \text{since } f \text{ is periodic.}$$

Integrating by parts, we further find

$$A_n = \int_{-\pi}^{\pi}(Df(x))\frac{\cos nx}{\sqrt{\pi}}\,dx = n\int_{-\pi}^{\pi} f(x)\frac{\sin nx}{\sqrt{\pi}}\,dx = nb_n$$

and

$$B_n = \int_{-\pi}^{\pi}(Df(x))\frac{\sin nx}{\sqrt{\pi}}\,dx = -n\int_{-\pi}^{\pi} f(x)\frac{\cos nx}{\sqrt{\pi}}\,dx = -na_n.$$

Our result is now only a few steps away.

Theorem 3.28 . *If $f \in C^1[-\pi,\pi]$ and if both f and f' are periodic with period 2π, then the Fourier series $P_N f$ converges to f in the uniform norm*

$$\lim_{N\to\infty} \|f - P_N f\|_\infty = 0.$$

PROOF: The key observation is that f' is a continuous function, so that Theorem 22 can be applied to its Fourier series. This shows that

$$\lim_{N\to\infty} \|Df - P_N(Df)\| = 0.$$

By the above lemma,

$$D(f - P_n f) = Df - D(P_N f) = Df - P_N(Df).$$

Thus

$$\lim_{N\to\infty} \|D(f - P_N f)\| = 0. \tag{3-29}$$

We would like to apply Theorem 21 to the function $\theta_N = f - P_N f$. In order to do so, we must only verify that θ_N vanishes somewhere in $[-\pi, \pi]$. But the area under $\theta_N = f - P_N f$ is

$$\int_{-\pi}^{\pi} \theta_N(x)\, dx = \sqrt{2\pi}\langle \theta_N,\, e_0\rangle = 0$$

since $f - P_N f$ is orthogonal to the space spanned by $e_0, e_1, \tilde{e_1}, \ldots, e_N, \tilde{e}_N$ (Theorem 16c). Because $\theta_N(x)$ is a continuous function (the difference of the C^1) function f and the infinitely differentiable trigonometric polynomial $P_N f$), the area under it can be zero only if θ_N vanishes somewhere. Thus Theorem 21 is applicable and yields the inequality

$$\|f - P_N f\|_\infty \le \sqrt{b-a}\|D(f - P_N f)\|.$$

We now pass to the limit $N \to \infty$ and use equation (4) to complete the proof of the theorem:

$$\lim_{N\to\infty} \|f - P_N f\|_\infty \le \lim_{N\to\infty} \sqrt{b-a}\|D(f - P_N f)\| = 0.$$

Remarks. The hypothesis that $f \in C^1[-\pi, \pi]$ and is periodic with period 2π has been proved a sufficient condition for the Fourier series to converge to the function in the uniform norm. Much weaker hypotheses also suffice to prove the same result—but mere continuity is not enough. Convergence of Fourier series or generalizations thereof is a vast and deep subject, one still the object of intense study.

On the basis of the theorems we have proved, many other problems are reasonably accessible—like the convergence of the Fourier series for a function which is nice except for a finite number of jump discontinuities. But there is not time for this pleasant excursion.

A FIGURE GOES HERE

3.5 Appendix. The Weierstrass Approximation Theorem

.

The proof—which is difficult—will be given as a series of lemmas.

Lemma 3.29 . *If $f(x)$ is continuous and periodic with period 2π, then for any $a \in \mathbb{R}$, the following equality holds*

$$\int_a^{a+2\pi} f(x)\, dx = \int_0^{2\pi} f(x)\, dx.$$

Proof: This is clear from a graph of f, since the area under one period of f does not depend upon where you begin measuring. We also offer a computational proof. Write

$$\int_a^{a+2\pi} f(x)\, dx = \int_a^0 f(x)\, dx + \int_0^{2\pi} f(x)\, dx + \int_{2\pi}^{a+2\pi} f(x)\, dx.$$

Let $x = t + 2\pi$ in the last integral and use the fact that $f(t + 2\pi) = f(t)$. The last integral is then

$$-\int_a^0 f(t)\, dt,$$

which cancels the unwanted term in the last equation and proves the lemma.

Lemma 3.30 . $\displaystyle\int_0^{\pi/2} \cos^{2n} t\ dt = \frac{1}{2c_n}$, *where* $c_n = v\dfrac{1}{\pi}\dfrac{2\cdot 4\cdot 6\cdots(2n)}{1\cdot 3\cdot 5\cdots(2n-1)}$.

Proof: A computation. Integrate by parts to show that

$$I_{2n} = \int_0^{\pi/2} \cos^{2n} t\, dt = (2n-1)(I_{2n-2} - I_{2n}).$$

Thus $I_{2n} = \frac{2n-1}{2n} I_{2n-2}$. Now induction can be used to do the rest, since by observation $I_0 = \pi/2$.

Lemma 3.31 . *Assume $f(x)$ is continuous and periodic with period 2π. Let*

$$T_N(x) = \frac{c_N}{2}\int_{-\pi}^{\pi} f(t)\cos^{2N}(\frac{t-x}{2})\, dt \tag{3-30}$$

Then given any $\epsilon > 0$, there is an N such that

$$\|f - T_N\|_\infty = \max_{-\pi\le x\le\pi} |f(x) - T_N(x)| < \epsilon.$$

Proof: How did we guess the formula (4)? We observed that $\cos^{2N} x$ is one at $x = 0$, and strictly less than one for all other $x \in [-\pi, \pi]$. Thus, for large N, $\cos^{2N} x$ is one at $x = 0$, and decreases sharply thereafter so $\cos^{2N}(\frac{t-x}{2})$ has the same property at $x - t = 0$, where $x = t$. Then essentially the only values of $f(t)$ which will count are those about $t = x$, so what comes out will be $f(x)$. Let us proceed with the details.

Take $s = \frac{t-x}{2}$. Then

$$T_N(x) = c_N \int_{-\pi/2}^{\pi/2} f(x + 2s)\cos^{2N} s\, ds.$$

Split the integral into two pieces, from $-\frac{\pi}{2}$ to 0 and from 0 to $\frac{\pi}{2}$, and then replace s by $-s$ in the first one. This gives

$$T_N(x) = c_N \int_0^{\pi/2} [f(x) + 2s) + f(x - 2s)]\cos^{2N} s\, ds.$$

From Lemma 2 we know that

$$f(x) = c_N \int_0^{\pi/2} 2f(x)\cos^{2N} s\, ds,$$

since $f(x)$ is a constant in the integration with respect to s. Therefore

$$T_N(x) - f(x) = c_N \int_0^{\pi/2} [f(x+2s) - 2f(x) + f(x-2s)] \cos^{2N} s\, ds.$$

Now given any $\epsilon > 0$, from the continuity of f we can pick a $\delta > 0$ independent of x such that

$$|f(x_1) - f(x_2)| < \frac{\epsilon}{2} \quad \text{when} \quad |x_1 - x_2| < \delta.$$

This ϵ will be the ϵ of our conclusion. Break the integral into two parts, one from 0 to δ and the other from δ to $\pi/2$, where δ is the δ we just found. Then in the $[0, \delta]$ interval,

$$|f(x+2s) - 2f(x) + f(x-2s)| \le |f(x+2s) - f(x)| + |f(x) - f(x-2s)| < \epsilon,$$

while in the $[\delta, \frac{\pi}{2}]$ interval,

$$|f(x+2s) - 2f(x) + f(x-2s)| \le |f(x+2s)| + 2\,|f(x)| + |f(x-2s)| \le 4M,$$

where $M = \max\limits_{x \in [-\pi,\pi]} |f(x)|$. Hence

$$|f(x) - T_N(x)| < c_N[\epsilon \int_0^{\delta} \cos^{2N} s\, ds + 4M \int_0^{\pi/2} \cos^{2N} s\, ds].$$

Now we observe that

$$\int_0^{\delta} \cos^{2N} s\, ds < \int_0^{\pi/2} \cos^{2N} s\, ds = \frac{1}{2c_N},$$

and that, since $\cos s$ decreases as s goes to $\pi/2$,

$$\int_{\delta}^{\pi} \cos^{2N} s\, ds < \int_{\delta}^{\pi/2} \cos^{2N} \delta\, ds < \frac{\pi}{2}\gamma^N,$$

where $\gamma = \cos^2 \delta < 1$. Thus

$$|f(x) - T_N(x)| < \frac{\epsilon}{2} + 2\pi M c_N \gamma^N.$$

Now $\pi c_N = \left(\frac{2}{3} \cdot \frac{4}{5} \cdots \frac{2N-2}{2N-1}\right) \cdot 2N < 2N$, so that $2\pi M c_N \gamma^N < 4MN\gamma^N$. Because $\gamma < 1$, we know that $\lim\limits_{N\to\infty} N\gamma^N = 0$. Thus, pick N so large that $N\gamma^N < \frac{\epsilon}{8M}$, where this is the same ϵ as before. Consequently, for this N,

$$|f(x) - T_N(x)| < \epsilon.$$

Since ϵ is independent of x,

$$\|f(x) - T_N(x)\| = \max_{x \in [-\pi,\pi]} |f(x) - T_N(x)| < \epsilon$$

too. A difficult lemma is thereby proved.

The whole proof is completed in the following simple

Lemma 3.32 . *The function $T_N(x)$ defined by (3)*

$$T_N(x) = \frac{c_N}{2} \int_{-\pi}^{\pi} f(t) \cos^{2N}(\frac{t-x}{2})\, dt$$

is a trigonometric polynomial.

PROOF: This can be horribly messy unless one is shrewd. We shall use the formula $e^{i\theta} = \cos\theta + i\sin\theta$ and the binomial theorem (top p. 108). First notice that

$$\cos^{2N}\theta = \left(\frac{e^{i\theta} + e^{-i\theta}}{2}\right)^{2N} = \frac{1}{2^{2N}} \sum_{k=0}^{2N} \frac{(2N)!}{(2N-k)k!} e^{ik\theta} e^{-i(2N-k)\theta}.$$

Let $d_k = (2N)!/2^{2N}(2N-k)!k!$ Then

$$\begin{aligned} \cos^{2N}\theta &= \sum_{k=0}^{2N} d_k e^{-i(2N-2k)\theta} \\ &= \sum_{k=0}^{2N} d_k[\cos(2N-2k)\theta - i\sin(2N-2k)\theta]. \end{aligned} \tag{3-31}$$

Since $\cos^{2N}\theta$ is real, the sum of the imaginary terms on the right must be zero. Thus, replacing 2θ by $t - x$, we find that

$$\begin{aligned} \cos^{2N}(\frac{t-x}{)}2 &= \sum_{k=0}^{2N} d_k \cos(N-k)(t-x) \\ &= \sum_{k=0}^{2N} d_k[\cos(N-k)t\cos(N-k)x + \sin(N-k)t\sin(N-k)x]. \end{aligned} \tag{3-32}$$

Split the sum into two parts, one from 0 to N, the other from $N+1$ to $2N$, and let $n = N - k$ in the first, $n = k - N$ in the second. This gives

$$\begin{aligned} \cos^{2N}(\frac{t-x}{2}) &= \sum_{n=0}^{N} d_{N-n}[\cos nt\, \cos nx + \sin nt\, \sin nx] \\ &\quad + \sum_{n=1}^{N} d_{N+n}[\cos\, nt\, \cos\, nx + \sin\, nt\, \sin\, nx], \end{aligned} \tag{3-33}$$

so

$$\cos^{2N}(\frac{t-x}{2}) = d_N + \sum_{n=1}^{N} (d_{N+n} + d_{N-n})[\cos nt\, \cos nx + \sin nt\, \sin nx],$$

which is much more simple than one might have anticipated. Substituting this into (4) and realizing that the t integrations just yield constants, we find that $T_N(x)$ is indeed a trigonometric polynomial. Coupled with Lemma 3, the proof of Weierstrass' Approximation Theorem is completely proved.

Exercises

1. Find the Fourier series with period 2π for the given functions.
   a) $f(x) = \begin{cases} 0, & -\pi \le x \le 0 \\ 2, & 0 < x < \pi \end{cases}$
   b) $f(x) = \begin{cases} -2, & -\pi \le x < 0 \\ 2, & 0 \le x < \pi \end{cases}$
   c) $f(x) = \sin 17x + \cos 2s, \quad -\pi \le x < \pi$
   d) $f(x) = \sin^2 x,\ -\pi \le x \le \pi;$
   e) $f(x) = x^2,\ -\pi \le x \le \pi$
   f) $f(x) = \begin{cases} x+\pi, & -\pi \le x \le 0 \\ -x+\pi, & 0 \le x \le \pi \end{cases}$ (Also, compute $\|f\|^2$ and $a_0^2 + \sum_{n=1}^{\infty}(a_n^2 + b_n^2)$ for (a)-(f)).

2. a) Apply Parseval's Theorem (Corollary to Theorem 22) to the function $f(x) = x$ and its Fourier series to deduce that
   $$\frac{\pi^2}{6} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots$$
   (cf. the example before Theorem 17 of Section 3).
   b) Do the same for the function $f(x) = x^2$ (Ex. 1, e above) to evaluate
   $$1 + \frac{1}{2^4} + \frac{1}{3^4} + \frac{1}{4^4} + \cdots = ?$$

3. A function $f(x)$ is *even* if $f(-x) = f(x)$, *odd* if $f(-x) = -f(x)$. Thus $2 + x^2$ is an even function, $x^3 - \sin x$ is an odd function, while $1 + x$ is neither even nor odd. Let a_n and b_n be the Fourier coefficients of the piecewise continuous function $f(x)$. Prove the following statements.
   a) If f is an *odd* function,
   $$a_n = 0, \quad b_n = 2\int_0^{\pi} f(x)\frac{\sin nx}{\sqrt{\pi}}\,dx$$
   b) If f is an *even* function
   $$a_n = 2\int_0^{\pi} f(x)\frac{\cos nx}{\sqrt{\pi}}\,dx, \quad b_n = 0$$

   c) A function f defined in $[0,\pi]$ may be extended to $[-\pi,\pi]$ as either an even or odd function by the formulas
   $$\textit{even extension}: \quad f(-x) = f(x), \quad x \ge 0,$$
   or
   $$\textit{odd extension}: \quad f(-x) = -f(x), \quad x \ge 0.$$
   The even extension of $f(x) = x$, $x \in [0,\pi]$ is $f(x) = |x|$, $x \in [-\pi,\pi]$, while its odd extension is $f(x) = x$, $x \in [-\pi,\pi]$. The odd extension of $f(x) = x^2$, $x \in [0,\pi]$ is $f(x) = \begin{cases} x^2, & x \in [0,\pi] \\ -x^2, & x \in [-\pi,0] \end{cases}$. Extend the function $f(x) = 1$, $x \in [0,\pi]$ to the interval $[-\pi,\pi]$ as an odd function and sketch its graph. Find its Fourier series using part (a).

4. a) Let $f(x)$ be a given function. Find a solution of the O.D. E. $u'' + \lambda^2 u = f$, where λ is a real number and $u(x)$ satisfies the boundary condition $u(-\pi) = u(\pi) = 0$, by the following procedure: Expand f in its Fourier series and assume u has a Fourier series whose coefficients are to be found. Find a formula for the Fourier coefficients of u in terms of those for f in the case where λ is not an integer.
   b) If $\lambda = n$ is an integer, show that there is a solution if and only if $0 = \langle f, \tilde{e}_n\rangle = \int_{-\pi}^{\pi} f(x)\frac{\sin nx}{\sqrt{\pi}}\,dx$.

5. a) State Parseval's Theorem for the special cases i) f is a continuous even function in $[-\pi,\pi]$, and ii) f is a continuous odd function in $[-\pi,\pi]$.
   b) If f is a continuous even function in $[-\pi,\pi]$ and
   $$\int_0^{\pi} f(x)\cos nx\,dx = 0, \quad n = 0, 1, 2, 3, \ldots,$$
   show that $f = 0$ in $[-\pi,\pi]$.
   c) State and prove a theorem similar to (b) in the case of a continuous odd function.

6. In this exercise you show how a function $f \in L_2[-A,A]$ can be expanded in a modified Fourier series (so far we know only $L_2[-\pi,\pi]$). Let $y = \frac{\pi x}{A}$—this maps the interval $[-A,A]$ onto $[-\pi,\pi]$—and define $g(y)$ by
   $$f(x) = f(\frac{Ay}{\pi}) = g(y) = g(\frac{\pi x}{A}).$$
   Since $g(y) \in L_2[-\pi,\pi]$, it can be expanded in a Fourier series
   $$g(y) = a_0\frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{\infty} a_n\frac{\cos ny}{\sqrt{\pi}} + b_n\frac{\sin ny}{\sqrt{\pi}},$$
   where the a_n and b_n are given by the usual formulas (2)'.

a) Prove that $f(x) \in L_2[-A, A]$ has the modified Fourier series

$$f(x) = a_0 \frac{1}{\sqrt{2A}} + \sum_{n=1}^{\infty} \cos\frac{nx}{A}x + \frac{b_n}{\sqrt{A}} \sin\frac{n\pi}{A}x,$$

where

$$a_0 = \frac{1}{\sqrt{2A}} \int_{-A}^{A} f(x)\,dx$$

$$a_n = \frac{1}{\sqrt{A}} \int_{-A}^{A} f(x) \cos\frac{n\pi x}{A}\,dx, \quad b_n = \frac{1}{\sqrt{A}} \int_{-A}^{A} f(x) \sin\frac{n\pi x}{A}\,dx.$$

b) Find the modified Fourier series for $f(x) = |x|$, in the interval $[-1, 1]$.

The following exercises all concern the Weierstrass Approximation Theorem.

7. Prove the following version of the Weierstrass Approximation Theorem. Let $f \in C[a, b]$. Then given any $\epsilon > 0$, there is a polynomial $Q(x)$ such that

$$\|f - Q\|_\infty = \max_{x \in [a,b]} |f(x) - Q(x)| < \epsilon.$$

(Hint: Let $y = -\pi + 2\frac{(x-a)}{b-a}\pi$. This maps $[a, b]$ into $[-\pi, \pi]$. Define $g(y)$, $y \in [-\pi, \pi]$ by

$$f(x) = f(a + \frac{(b-a)}{2\pi}(y + \pi) = g(y) = g(-\pi + 2\frac{(x-a)}{b-a}\pi).$$

Use the version of the theorem proved to approximate $g(y)$, $y \in [-\pi, \pi]$ by a trigonometric polynomial $T_N(y)$ to within $\epsilon/2$. Then approximate $\sin ny$ and $\cos ny$ to within $c\epsilon$ (you pick c) by a finite piece of their Taylor series—which are polynomials. Put both parts together to obtain the complete proof for $g(y)$. The transition back to $f(x)$ is trivial.]

8. (Riemann-Lebesgue Lemma). Let $f \in C[a, b]$. Prove that

$$\lim_{\lambda \to \infty} \int_a^b f(x) \sin \lambda x\,dx = 0.$$

[*Hint*: Integrate by parts to prove it first for all $f \in C^1[a, b]$. For arbitrary f, approximate f by a polynomial—Ex. 7 above—to within $\epsilon/2$ and realize that every polynomial is in $C^1[a, b]$].

9. If $f \in C[0, 1]$, prove that

$$\lim_{n \to \infty} n \int_0^1 f(x)x^n\,dx = f(1).$$

[*Hint*: Use the hint in Ex. 8].

10. If $f \in C[a, b]$, and if

$$\int_a^b f(x)x^n\,dx = 0, \quad n = 0, 1, 2, 3, \ldots,$$

show that $f = 0$. [*Hint*: This implies that $\int_a^b f(x)Q(x)\,dx = 0$, where Q is *any* polynomial. f can be approximated by some polynomial $\tilde{Q}$. Now show that $\int_a^b f^2(x)\,dx = 0$.]

3.6 The Vector Product in $\mathbb{R}^3$

.

As you grasped many years ago, the world we live in has three space dimensions. For this reason the material in this section is important in many applications. What we intend to do is define a way to multiply two vectors X and Y in $\mathbb{R}^3$. Whereas the scalar product $\langle X, Y\rangle$ is a *scalar*, this product $X \times Y$, the *vector product*, or *cross product* as it is often called, is a *vector*.

For several reasons [i) we shall not cover this in class, and ii) I can probably not do as good a job as appears in many books] we shall let you read about this topic elsewhere. But make sure to read about it even though you'll never be examined on it.

Chapter 4

Linear Operators: Generalities. $V^1 \to V_n, V_n \to V^1$

4.1 Introduction. Algebra of Operators

.

Let **V** by a linear space. So far we have considered the algebraic structure of such a space; however most significant reason for studying linear spaces is so that one can study operators defined on them. *Operator* is another, more organic, name for function. Thus an operator

$$T \colon A \to B$$

T maps elements in its domain A into elements of B, where B contains the range of T. If $X \in A$, then $T(X) = Y \in B$. Think of feeding X into the operator T, and Y being

A FIGURE GOES HERE

what T sends out in return. It is useful to think of T as some type of machine or factory, the input (raw material) is X, and the output is Y. Some examples should illustrate the situation and its potential power.

EXAMPLES:

(1) Let $\mathbf{V} = \mathbb{R}^2$. If $X = (x_1, x_2) \in \mathbb{R}^2$, and $Y = (y_1, y_2, y_3)$, we define $T(X) = Y$ by

$$T(X) = \left\{ \begin{array}{c} x_1 + 2x_2 = y_1 \\ x_1 + x_2 = y_2 \\ 3x_1 + x_2 = y_3 \end{array} \right\},$$

or

$$T(X) = T(x_1, x_2) = (x_1 + 2x_2,\ x_1 + x_2,\ 3x_1 + x_2) = (y_1, y_2, y_3) = Y.$$

This operator T has the property that to every $X \in \mathbb{R}^2$ it assigns a $Y \in \mathbb{R}^3$. In other words T maps the two dimensional space $\mathbb{R}^2$ into the three dimensional space $\mathbb{R}^3$

$$T\colon \mathbb{R}^2 \to \mathbb{R}^3.$$

$\mathbb{R}^2$ is the *domain* of T, denoted by $\mathcal{D}(T)$, while the *range* of T, $\mathcal{R}(T)$ is contained in $\mathbb{R}^3$,

$$\mathcal{D}(T) = \mathbb{R}^2, \quad \mathcal{R}(T) \subset \mathbb{R}^3.$$

Since $y_1 = y_2 = 0$ implies that $x_1 = x_2 = 0$, which in turn implies that $y_3 = 0$, we see that the point $(0,0,1) \in \mathbb{R}^3$ is *not* in the range of T. Thus, T is *not surjective* onto $\mathbb{R}^3$. It is *injective* (one-to-one) since every point $Y \in \mathcal{R}(T)$ is the image of exactly one $X \in \mathcal{D}(T)$. This an be seen by observing that y_1 and y_2 suffice to determine $X = (x_1, x_2)$ uniquely by solving the first two equations

$$\begin{aligned} -y_1 + 2y_2 &= x_1 \\ y_1 - y_2 &= x_2. \end{aligned}$$

Hence if $Y = T(X_1)$ and also $Y = T(X_2)$, then $X_1 = X_2$. Since the operator T is completely determined by the coefficients in the equations, it is reasonable to represent this T by the *matrix*

$$T = \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 3 & 1 \end{pmatrix}$$

If you care to think of X as the input into a paint-making machine, then x_1 might represent the quantity of yellow and x_2 the quantity of blue used. In this case y_1, y_2 and y_3 represent the quantities of three different shades of green the machine yields. For this machine, as soon as you specify the desired quantities of any two of the greens, say y_1 and y_2, the quantities x_1 and x_2 of the input colors are completely determined, as is the quantity y_3 of the remaining shade of green.

(2) Let $\mathbf{V}$ be $\mathbb{R}^2$ again. With $X = (x_1, x_2) \in \mathbb{R}^2$, and $Y = (y_1) \in \mathbb{R}^1$, define T by

$$x_1^2 + x_2^2 = y_1,$$

or

$$T(X) = x_1^2 + x_2^2.$$

This operator T maps $\mathbb{R}^2$ into $\mathbb{R}^1$

$$T\colon \mathbb{R}^2 \to \mathbb{R}^1.$$

It is not surjective onto $\mathbb{R}^1$ since the negative half of $\mathbb{R}^1$ is completely omitted from $\mathcal{R}(T)$. Furthermore, it is not injective either since each point $y_1 \in \mathcal{R}(T)$ other than zero is the image of infinitely many points—all of those on the circle $x_1^2 + x_2^2 = y_1$.

(3) Let $\mathbf{V}$ be $C[-1,1]$. If $f \in C[-1,1]$, we define T by

$$T(f) = f(0).$$

Thus, if $f(x) = 2 + \cos x$, then $Tf = 3$. This operator T is usually denoted by δ and called the *Dirac delta functional*. It was first used by Dirac in his work on quantum mechanics and is extremely valuable in modern mathematics and physics. T assigns to each continuous function f its value at $x = 0$, a real number. Therefore

$$T\colon C[-1,1] \to \mathbb{R}^1.$$

The operator T is not injective, since for example the element $2 \in \mathbb{R}^1$ is the image of both $f(x) = 1 + e^x$ and $f(x) = 2$. It is surjective since every element $a \in \mathbb{R}^1$ is the image of at least one element in $C[-1,1]$ (if $f(x) \equiv a$, then clearly $T(f) = a$).

(4) Let $\mathbf{V}$ be $C[-1,1]$. If $f \in C^1[-1,1]$ then the differentiation operator D is defined by

$$(Df)(x) = \frac{df}{dx}(x).$$

It maps each function into its derivative. If $f(x) = x^2$, then $(Df)(x) = 2x$. Since the derivative of a continuously differentiable function (a function in C^1) is necessarily continuous, we see that

$$D\colon C^1[-1,1] \to C[-1,1].$$

D is not injective since, for example, the function $g(x) = 1$ is the image of both $f_1(x) = x$ and $f_2(x) = 2 + x$. D is surjective onto $C[-1,1]$.

$$\mathcal{R}(D) = C[-1,1],$$

since if $g(x)$ is any element of $C[-1,1]$, then g is the image of the particular function $f \in C^1[-1,1]$ defined by

$$f(x) = \int_0^x g(s)\, ds,$$

because $Df = g$ by the fundamental theorem of calculus.

Throughout this and the next chapter we will study some of the elementary aspects of linear operators. It is reasonable to denote a linear operator by L.

Definition Let $\mathbf{V}_1$ and $\mathbf{V}_2$ both be linear spaces over the same field of scalars. An operator L mapping $\mathbf{V}_1$ into $\mathbf{V}_2$ is called a *linear operator* if for every X and $\tilde{X}$ in $\mathbf{V}_1$ and any scalar a, L satisfies the two conditions

1. $L(X + \tilde{X}) = L(X) + L(\tilde{X})$
2. $L(aX) = aL(X)$.

Whenever ambiguity does not arise, we will omit the parentheses and write LX instead of $L(X)$.

An equivalent form of the definition is

Theorem 4.1 . *L is a linear operator* $\Longleftrightarrow$

$$L(aX + b\tilde{X}) = aL(X) + bL(\tilde{X}),$$

where $X, \tilde{X} \in \mathbf{V}_1$ and a and b are any scalars.

PROOF: $\Rightarrow$

$$\begin{aligned} L(aX + b\tilde{X}) &= L(aX) + L(b\tilde{X}) \quad \text{(property 1)} \\ &= aLX + bL\tilde{X} \quad \text{(property 2).} \end{aligned} \tag{4-1}$$

$\Leftarrow$ Property 1 is the special case $a = b = 1$. Property 2 is the special case $b = 0$.
REMARK:

It is useful to observe that always $L(0) = L(0 \cdot X) = 0L(X) = 0$. This identity is often the easiest way to test if an operator is *not* linear.

EXAMPLES:

(1) The operator L defined by example 1 where $L\colon \mathbb{R}^2 \to \mathbb{R}^3$ is

$$LX = (x_1 + 2x_2, x_1 + x_2, 3x_1 + x_2),$$

is linear. Let $X = (x_1, x_2)$ and $\tilde{X} = (\tilde{x}_1, \tilde{x}_2)$. Then

$$\begin{aligned} L(X + \tilde{X}) &= (x_1 + \tilde{x}_1 + 2x_2 + 2\tilde{x}_2, x_1 + \tilde{x}_1 + x_2 + \tilde{x}_2, 3x_1 + 3\tilde{x}_1 + x_2 + \tilde{x}_2) \\ &= (x_1 + 2x_2, x_1 + x_2, 3x_1 + x_2) + (\tilde{x}_1 + 2\tilde{x}_2, \tilde{x}_1 + \tilde{x}_2, 3\tilde{x}_1 + \tilde{x}_2) \\ &= LX + L\tilde{X} \end{aligned}$$

and

$$\begin{aligned} L(aX) &= (ax_1 + 2ax_2, ax_1 + ax_2, 3ax_1 + ax_2) \\ &= a(x_1 + 2x_2, x_1 + x_2, 3x_z + x_2) \\ &= aLX. \end{aligned} \tag{4-2}$$

(2) The operator $TX = x_1^2 + x_2^2$ with domain $\mathbb{R}^2$ and range $\mathbb{R}^1$ is *not* linear, since

$$T(aX) = (ax_1)^2 + (ax_2)^2 = a^2[x_1^2 + x_2^2] \neq aTX$$

except for the particular scalars $a = 0, 1$.

(3) The operator $Df = \frac{df}{dx}$ with domain $C^1[-1,1]$ and range $C[-1,1]$ is linear since if f_1 and f_2 are in $C^1[-1,1]$ and a and b are any real numbers, then by elementary calculus

$$\begin{aligned} D(af_1 + bf_2) &= \frac{d}{dx}(af_1 + bf_2) = a\frac{df_1}{dx} + b\frac{df_2}{dx} \\ &= aDf_1 + bDf_2. \end{aligned} \tag{4-3}$$

(4) The operator L defined as

$$Lu = a_2(x)u'' + a_1(x)u' + a_0(x)u, \quad ('= \frac{d}{dx}),$$

where $u(x) \in \mathcal{D}(L) = C^2$, and where $a_0(x)$, $a_1(x)$, and $a_2(x)$ are continuous functions, is a linear operator,

$$L\colon C^2 \to C.$$

If A and B are any constants (scalars for C^2), then for any u_1 and $u_2 \in C^2$,

$$\begin{aligned} L(Au_1 + Bu_2) &= a_x[Au_1 + Bu_2]'' + a_1[Au_1 + Bu_2]' + a_0[Au_1 + Bu_2] \\ &= a_2Au_1'' + a_2B_2'' + a_1Au_1' + a_1Bu_2' + a_0Au_1 + a_0Bu_2 \\ &= A[a_2u_1'' + a_1u_1' + a_0u_1] + B[a_2u_2'' + a_1u_2' + a_0u_2] \\ &= ALu_1 + BLu_2. \end{aligned} \tag{4-4}$$

(5) The *identity operator* I is the operator which leaves everything unchanged. Because it is so simple, it can be defined on an arbitrary set S and maps S into *itself* $S \to S$ in a trivial way. If $X \in S$, then we define

$$IX = X.$$

What could be more simple? If S is a linear space $\mathbf{V}$ (so aX and $X_1 + X_2$ are defined), then I is trivially a linear operator, since

$$I(aX_1 + bX_2) = aX_1 + bX_2 = aIX_1 + bIX_2$$

Why are linear operators important? There are several reasons. First, they are much easier to work with than nonlinear operators. Second, most of the operators which arise in applications are linear. The feature possessed by linear operators which is central to applications is that of *superposition*. If $Lu_1 = f$ and $Lu_2 = g$, then $L(u_1 + u_2) = f + g$. In other words, if u_1 is the response to some external influence f and u_2 the response to g, then the response to $f + g$ is found by adding the separate responses.

The special case of a linear operator whose range is the real number line $\mathbb{R}^1$ arises often enough to receive a name of its own.
DEFINITION: A linear operator whose range is $\mathbb{R}^1$ is called a *linear functional*, $\ell\mathbf{V} \to \mathbb{R}^1$.

The Dirac delta functional is such an operator. So is the operator

$$l(f) = \int_0^1 f(x)\,dx,$$

which assigns to every continuous function $f \in C[0,1]$ the real number equal to the area between the graph of f and the x-axis. Check that ℓ is linear.

If the linear operator $L\colon V_1 \to V_2$ the range of L—a subset of the linear space V_2—has a particularly nice structure. In fact, $\mathcal{R}(L)$ is not just any clump of points in V_2 but

Theorem 4.2 . *The range of a linear operator $L\colon V_1 \to V_2$ is a linear subspace of V_2.*

REMARK: Even more is true. We shall prove (p. 312-3) that $\dim \mathcal{R}(L) \le \dim \mathcal{D}(L)$ so that no matter how large V_2 is, the range has at most the same dimension as the domain.

PROOF: The range of L consists of all elements $Y \in V_2$ of the form $Y = LX$ where $X \in V_1$. We know that $\mathcal{R}(L)$ is a subset of the linear space V_2. The only task is to prove that it is actually a subspace. Since V_2 is a linear space, it is sufficient to show that the set $\mathcal{R}(L)$ is closed under multiplication by scalars, and under addition of vectors. i) $\mathcal{R}(L)$ is closed under multiplication by scalars. If $Y \in \mathcal{R}(L)$, there is an $X \in V_1 = \mathcal{D}(L)$ such that $Y + LX$. We must find some $\tilde{X}$ in V_1 such that $aY = L\tilde{X}$, where a is any scalar. Since $aY = aLX = L(aX)$, we take $\tilde{X} = aX$.

ii) $\mathcal{R}(L)$ is closed under addition of vectors. If Y_1 and Y_2 are in $\mathcal{R}(L)$, there are elements X_1 and X_2 in $V_1 = \mathcal{D}(L)$ such that $Y_1 = LX_1$ and $Y_2 = LX_2$. We must show that $Y_1 + Y_2 \in \mathcal{D}(L)$, that is, find some $\tilde{X} \in V_1$ such that $Y_1 + Y_2 = L\tilde{X}$. But $Y_1 + Y_2 = LX_1 + LX_2 = L(X_1 + X_2)$. Thus we can take $\tilde{X} = X_1 + X_2$.

Before moving further on into the realm of special linear operators, we shall take this opportunity to define algebraic operations (addition and multiplication) for linear operators. But first we define equality, $L_1 = L_2$, in a straightforward way.

DEFINITION: (EQUALITY) If L_1 and L_2 both map V_1 into V_2, where V_1 and V_2 are linear spaces, and if $L_1X = L_2X$ for all X in V_1, then L_1 *equals* L_2. Thus, two operators are equal if they have the same effect on any vector.

Addition is equally simple.

DEFINITION: (ADDITION). If $L_1\colon V_1 \to V_2$ and $L_2\colon V_1 \to V_2$ then their sum, $L_1 + L_2$, is defined by the rule

$$(L_1 + L_2)X = L_1X + L_2X, \quad X \in V_1$$

EXAMPLES:

(1) Let $L_1\colon \mathbb{R}^2 \to \mathbb{R}^3$ be defined by

$$L_1(X) = (x_1 + x_2, x_1 + 2x_2, -x_2), \quad X = (x_1, x_2) \in \mathbb{R}^2$$

and $L_2\colon \mathbb{R}^2 \to \mathbb{R}^3$ be defined by

$$L_2X = (-3x_1 + x_2, x_1 - x_2, x_1), \quad X = (x_1, x_2) \in \mathbb{R}^2.$$

Then $L_1 + L_2$ is defined, and is

$$\begin{aligned}(L_1 + L_2)X + L_1X + L_2X &= (x_1 + x_2, x_1 + 2x_2, -x_2) + (-3x_1 + x_2, x_1 - x_2, x_1)\\ &= (-2x_1 + 2x_2, 2x_1 + x_2, x_1 - x_2)\end{aligned} \tag{4-5}$$

(2) Let $D\colon C^1 \to C$ be defined by

$$Du = \frac{du}{dx} \qquad u \in C^1,$$

and $L\colon C^1 \to C$ be defined by

$$\begin{aligned}Lu &= \int_0^1 e^{x-t}u(t)\,dt \qquad u \in C^1\\ &= e^x \int_0^1 e^{-t}u(t)\,dt.\end{aligned} \tag{4-6}$$

(In reality, L may be defined on a much larger class of functions—$u \in C$ is plenty, while its image is the smaller space, constant $e^x \subset C$. We have decided on the smaller domain and larger image space so that the sum $D + L$ is defined). Then for any $u \in C^1$.

$$(D + L)u = Du + Lu = \frac{du}{dx} + \int_0^1 e^{x-t}u(t)\,dt.$$

The following theorem is a statement of some simple facts about the sum of two linear operators.

Theorem 4.3 . *Let $L_1, L_2, L_3, \ldots$ be any linear operators which map $V_1 \to V_2$, so that their sums are defined. Then*

0. $L = L_1 + L_2$ is a linear operator

(1) $L_1 + (L_2 + L_3) + (L_1 + L_2) + L_3$,

(2) $L_1 + L_2 = L_2 + L_1$

(3) *Let 0 be the operator which maps every element of V_1 into $0 \in V_2$, so $0X = 0$. Then*

$$L_1 + 0 = L_1.$$

(4) $L_1 + (-L_1) = 0$. *Here $-L_1$ is the operator which maps every element $X \in V_1$ into $-(L_1X)$.*

Proof: These are just computations. Let $X_1, X_2 \in V_1$.

0.

$$\begin{aligned}
L(aX_1 + bX_2) &= (L_1 + L_2)(aX_1 + bX_2) \\
&= L_1(aX_1 + bX_2) + L_2(aX_1 + bX_2) \\
&= aL_1X_1 + bL_1X_2 + aL_2X_1 + bL_2X_2 \\
&= a(L_1X_1 + L_2X_1) + b(L_1X_2 + L_2X_2) \\
&= a(L_1 + L_2)X_1 + b(L_1 + L_2)X_2 \\
&= aLX_1 + bLX_2.
\end{aligned} \tag{4-7}$$

(1) $(L_1+(L_2+L_3))X = L_1X+(L_2+L_3)X = L_1X+L_2X+L_3X = (L_1+L_2)X+L_3X = ((L_1 + L_2) + L_3)X$.

(2) $(L_1 + L_2)X = L_1X + L_2X = L_2X + L_1X = (L_2 + L_1)X$. The step $L_1X + L_2X = L_2X + L_1X$ is justified on the grounds that the vectors $Y_1 := L_1X$ and $Y_2 := L_2X$ are elements of V_2—which is a linear space—so that $Y_1 + Y_2 = Y_2 + Y_1$.

(3) $(L_1 + 0)X = L_1X + 0X = L_1X + 0 = L_1X$

Note that the 0 in $0X$ is an operator, while the 0 in the next step is an element of V_2. This ambiguity causes no trouble once you understand it.

(4) $(L_1 + (-L_1))X = L_1X + (-L_1)X = L_1X - L_1X = 0$

The crucial step $(-L_1)X = -L_1X$ is the definition of the operator $(-L_1)$.

Remark: This theorem states that the set of all linear operators mapping one linear space V_1 into another V_2 form an abelian group under addition.

Multiplication of operators is not much more difficult. If L_1 and L_2 are linear operators, then their product L_2L_1 *in that order* is defined by the rule $L_2L_1X = L_2(L_1X)$. In other words, *first* operate on X with L_1 giving a vector $Y = L_1X$. *Then* operate on this new vector Y with L_2, giving $L_2Y = L_2(L_1X)$. It is clear that in order for this to make sense, for every $X \in \mathcal{D}(L_1)$, the new vector $Y = L_1X$ must be in the domain of L_2. Thus to form the product L_2L_1, we require that $\mathcal{R}(L_1) \subset \mathcal{D}(L_2)$.

Look at our machine again.

A FIGURE GOES HERE

The multiplication L_2L_1 means sending the output from L_1 as input into L_2. In order to join the machines in this way, surely one necessary requirement is that L_2 is equipped to act on the output from $\mathcal{R}(L_1)$, that is, $\mathcal{R}(L_1) \subset \mathcal{D}(L_2)$. Of course the L_2 machine might be able to digest input other than what L_1 sends out. But all we care is that L_2 can digest *at least* what L_1 sends it.

Definition: (multiplication). Let $L_1 \colon V_1 \to V_2$ and $L_2 \colon V_3 \to V_4$. If the range of L_1 is contained in the domain of L_2, $\mathcal{R}(L_1) \subset \mathcal{D}(L_2)$, then the *product* L_2L_1 is definable by the composition rule

$$L_2L_1X = L_2(L_1X), \text{ where } X \in V_1 = \mathcal{D}(L_1).$$

The product L_2L_1 maps the input V_1 for L_1 into the output V_4 for L_2, $L_2L_1 \colon V_1 \to V_3 \to V_4$.

We exhibit a little diagram (cf. p. ???).

A FIGURE GOES HERE

The way to get from V_1 to V_4 using L_2L_1 is to first use L_1 to reach V_2. Then use L_2 to get to V_4.

Remarks: If L_2L_1 is defined, it is *not* necessarily true that L_1L_2 is defined (Example 1 below). Furthermore, even if L_1L_2 is also defined, it is only a rare coincidence that multiplication is commutative. Usually $L_2L_1 \neq L_1L_2$ when both products are defined. Thus the *order* L_2L_1 is *important.*

Examples:

(1) Let $L_1 \colon \mathbb{R}^2 \to \mathbb{R}^3$ be defined as

$$L_1X = (x_1 - x_2, x_2, -x_1 - 2x_2), \text{ where } X = (x_1, x_2) \in \mathbb{R}^2,$$

and let $L_2 \colon \mathbb{R}^3 \to \mathbb{R}^1$ be defined as

$$L_2Y = (y_1 + 2y_2 - y_3), \text{ where } Y = (y_1, y_2, y_3) \in \mathbb{R}^3.$$

Then $\mathcal{R}(L_1) \subset \mathbb{R}^3 = \mathcal{D}(L_2)$ so that the product L_2L_1 is definable and $L_2L_1 \colon \mathbb{R}^2 \xrightarrow{L_1} \mathbb{R}^3 \xrightarrow{L_2} \mathbb{R}^1$. Consider what L_2L_1 does to the particular vector $X_0 = (-1, 2) \in \mathbb{R}^2$.

$$L_2L_1X_0 = L_2(L_1X_0) = L_2(-3, 2, -3) = (-3 + 4 + 3 = 4)$$

Thus L_2L_1 maps $(-1, 2) \in \mathbb{R}^2$ into $4 \in \mathbb{R}^1$. More generally, if X is any vector in $\mathbb{R}^2$,

$$\begin{aligned}
L_2L_1X = L_2(L_1X) &= L_2(x_1 - x_2, x_2, -x_1 - 2x_2) \\
&= (x_1 - x_2 + 2x_2 + x_1 + 2x_2) = 2x_1 + 3x_2 \in \mathbb{R}^1.
\end{aligned} \tag{4-8}$$

Thus L_2L_1 maps $(x_1, x_2) \in \mathbb{R}^2$ into $2x_1 + 3x_2 \in \mathbb{R}^1$.

Since $\mathcal{R}(L_2) = \mathbb{R}^1$ and $\mathcal{D}(L_1) = \mathbb{R}^2$, $\mathcal{R}(L_2)$ not $\subset \mathcal{D}(L_1)$ so that the product L_1L_2 is *not* defined. You might be thinking that $\mathbb{R}^1$ is part of $\mathbb{R}^2$. What you mean is that $\mathbb{R}^2$ has one dimensional subspaces. It certainly does—an infinite number of them, all of the straight lines through the origin. Because there are so many subspaces of $\mathbb{R}^2$ which are one dimensional, there is no natural way of regarding $\mathbb{R}^1$ as being contained in $\mathbb{R}^2$. [On the other hand, there is a natural way in which C^1 can be regarded as contained in C. We used this above in our second example for addition of linear operators].

(2) Define $L_1 \colon \mathbb{R}^2 \to \mathbb{R}^2$ by the rule $L_1X = (2x_1 - 3x_2, -x_1 + x_2)$ and $L_2 \colon \mathbb{R}^2 \to \mathbb{R}^2$ by the rule $L_2X = (2x_2, x_1 + x_2)$ Then $\mathcal{R}(L_1) = \mathbb{R}^2 = \mathcal{D}(L_2)$ so that L_2L_1 is defined. It is given by

$$L_2L_1X = L_2(2x_1 - 3x_2, -x_1 + x_2) = (-2x_1 + 2x_2, x_1 - 2x_2)$$

In particular, L_2L_1 maps $X_0 = (1,2)$ into $(2,-3)$. Now $\mathcal{R}(L_2) = \mathbb{R}^2 = \mathcal{D}(L_1)$, so that L_1L_2 is also definable. It is given by

$$\begin{aligned} L_1L_2X &= L_1(2x_2, x_1 + x_2) \\ &= (2 \cdot 2x_2 - 3 \cdot (x_1 + x_2), -2x_2 + (x_1 + x_2)) \\ &= (-3x_1 + x_2, x_1 - x_2). \end{aligned} \tag{4-9}$$

In particular, L_1L_2 maps $X_0 = (1,2)$ into $(-1,-1)$. Since L_1L_2 and L_2L_1 map the point $X_0 = (1,2)$ into two different points, it is clear that $L_1L_2 \neq L_2L_1$, the operators do not commute.

(3) Let A be the subspace of $\mathbb{R}^2$ spanned by some unit vector e_1 and B be the subspace spanned by another unit vector e_2. Consider the projection operators P_A and P_B. They are linear since, for example,

$$\begin{aligned} P_A(aX_1 + bX_2) &= \langle aX_1 + bX_2,\, e_1\rangle e_1 \\ &= a\langle X_1,\, e_1\rangle e_1 + b\langle X_2,\, e_1\rangle e_1 \\ &= aP_AX_1 + bP_AX_2. \end{aligned} \tag{4-10}$$

Because $P_A \colon \mathbb{R}^2 \to \mathbb{R}^2$ and $P_B \colon \mathbb{R}^2 \to \mathbb{R}^2$, both products P_AP_B and P_BP_A are defined. We have

$$\begin{aligned} P_AP_BX &= P_A(P_BX) = P_A(\langle X,\, e_2\rangle e_2) \\ &= \langle X,\, e_2\rangle P_Ae_2 = \langle X,\, e_2\rangle \langle e_2,\, e_1\rangle e_1. \end{aligned} \tag{4-11}$$

Also,

$$\begin{aligned} P_BP_AX &= P_B(P_AX) = P_B(\langle X,\, e_1\rangle e_1) \\ &= \langle X,\, e_1\rangle P_Be_1 = \langle X,\, e_1\rangle \langle e_1,\, e_2\rangle e_2. \end{aligned} \tag{4-12}$$

Since $P_AP_BX \in A \subset \mathbb{R}^2$, while $P_BP_AX \in B \subset \mathbb{R}^2$, it is clear that usually $P_AP_B \neq P_BP_A$. They will happen to be equal if $A = B$, or if $A \perp B$ (for then $P_AP_B = P_BP_A = 0$). See the figure at the beginning of this example—and draw some more special cases for yourself.

(4) Let $L \colon C^\infty \to C^\infty$ (C^∞ is the space of infinitely differentiable functions) be defined by

$$(Lu)(x) = xu(x), \quad u \in C^\infty,$$

and $D \colon C^\infty \to C^\infty$ be defined by

$$(Du)(x) = \frac{du}{dx}(x), \quad u \in C^\infty.$$

Then $\mathcal{R}(L) = \mathcal{D}(D)$ so that the product DL is definable by

$$DLu = D(Lu) = D(xu) = \frac{d}{dx}(xu(x)) = xu' + u.$$

Also, $\mathcal{R}(D) = \mathcal{D}(L)$ so LD is definable by

$$LDu = L(Du) = L(u') = xu'.$$

Notice that $LD \neq DL$ unless $u = 0$.

We collect some properties of multiplication.

Theorem 4.4 . *If $L_1 \colon V_1 \to V_2$, $L_2 \colon V_3 \to V_4$, and $L_3 \colon V_5 \to V_6$, where $V_1 \subset V_3$ and $V_4 \subset V_5$, then*

0. The operator $L = L_2L_1$ is a linear operator.

1. $L_3(L_2L_1) = (L_3L_2)L_1$ —Associative law.

PROOF: 0.

$$\begin{aligned} L(aX_1 + bX_2) &= L_2\big(L_1(aX_1 + bX_2)\big) \\ &= L_2\big(aL_1X_1 + bL_1X_2\big) \\ &= L_2(aL_1X_1) + L_2(bL_1X_2) \\ &= aL_2L_1X_1 + bL_2L_1X_2 \\ &= aLX_1 + bLX_2. \end{aligned} \tag{4-13}$$

(1) By definition of the product,

$$[L_3(L_2L_1)]X = L_3[(L_2L_1)X] = L_3[L_2(L_1X)]$$

and

$$[(L_3L_2)L_1]X = (L_3L_2)(L_1X) = L_3[L_2(L_1X)].$$

Now match the ends.

Notice that the commonly occurring special case $V_1 = V_2 = V_3 = V_4 = V_5 = V_6$ is included in this theorem. In this special case, even more can be proved. For then the identity operator I, defined by $IX = X$ for all $X \in V$ can be used to multiply any other operator. Moreover, addition, $L_1 + L_2$ also makes sense.

Theorem 4.5 . *If the linear operators L_1, L_2, L_3 all map V into V, then representing any one of these by L,*

(1) $LI = IL = L$.

(2) *For any positive integer* n*, we define* L^n *inductively by the rule* $L^{n+1} = LL^n$*, and* $L^0 = I$*. Then for any non-negative integers* m *and* n*,*

$$L^{m+1} = L^m L^n.$$

(3) $(L_1 + L_2)L_3 = L_1 L_3 + L_2 L_3$.

(4) $L_3(L_1 + L_2) = L_3 L_1 + L_3 L_2$ *(This is needed in addition to 3 because of the non-commutativity).*

Proof:

(1) If $X \in V$,

$$(LI)X = L(IX) = LX$$

$$(IL)X = I(LX) = LX$$

(2) We shall prove $L^{m+n} = L^m L^n$ by induction on m. The statement is true, by definition, for $m = 1$. Assume it is true for $m = k$, so $L^{k+n} = L^k L^n$. Our job is to prove the statement for $m = k+1$. By the definition and the induction hypothesis, we have

$$L^{k+n+1} = LL^{k+n} = L(L^k L^n).$$

Since multiplication is associative, we find that

$$L(L^k L^n) = (LL^k)L^n.$$

But, by definition,

$$LL^k = L^{k+1}.$$

Thus,

$$L^{k+n+1} = L^{k+1} L^n.$$

This completes the induction proof.

(3) If $X \in V$,

$$[(L_1 + L_2)L_3]X = (L_1 + L_2)(L_3 X)$$

Let $L_3 X = Y \in V$. Then $(L_1 + L_2)Y = L_1 Y + L_2 Y$. Thus

$$[(L_1 + L_2)L_3]X = L_1(L_3 X) + L_2(L_3 X) = (L_1, L_3)X + (L_2 L_3)X.$$

(4) Same proof as 3.

Remark: If V_1 and V_2 are two linear spaces, the set of all linear operators which map V_1 into V_2 is usually denoted by $\mathrm{Hom}(V_1, V_2)$—Hom rhymes with Mom and Tom. In this notation, the last theorem concerned $\mathrm{Hom}(V, V)$. The abbreviation Hom is for the impressive word "homomorphism". Tell your friends.

Examples: Consider $D\colon C^\infty \to C^\infty$ defined by $(Du)(x) = \frac{du}{dx}(x)$. Then $D^n = \dfrac{d^n}{dx^n}$.

Exercises

(1) Determine which of the following are linear operators.

(a) $T\colon \mathbb{R}^2 \to \mathbb{R}^2$

$$TX = (x_1 + x_2, x_1 - x_2),$$

where $X = (x_1, x_2) \in \mathbb{R}^2$.

(b) $T\colon \mathbb{R}^2 \to \mathbb{R}^2$,

$$T(X) = (x_1 + x_2 + 1, x_1 - x_2)$$

(c) $T\colon \mathbb{R}^3 \to \mathbb{R}^2$

$$T(X) = (x_1 + x_1 x_2, x_2)$$

(d) $T\colon \mathbb{R}^3 \to \mathbb{R}^1$

$$T(X) = (x_1 + x_2 - x_3)$$

(e) $T\colon \mathbb{R}^3 \to \mathbb{R}^1$

$$T(X) + (x_1 + x_2 - x_3 + 2)$$

(f) $D\colon \mathcal{P}_2 \to \mathcal{P}_1$. If $P(x) = a_2 x^2 + a_1 x + a_0 \in \mathcal{P}_2$ then

$$D(P) = 2a_2 + a_1 \in \mathcal{P}_1.$$

(g) $T\colon C^1[-1,1] \to \mathbb{R}^1$. If $u(x) \in C^1[-1,1]$, then

$$T(u) = u(0) + u'(0).$$

(h) $T\colon C[2,3] \to C[2,3]$. If $u \in C[2,3]$,

$$(Tu)(x) = \int_2^3 e^{x-t} u(t)\, dt$$

(i) $T\colon C[2,3] \to C[2,3]$,

$$(Tu)(x) = 1 + \int_2^3 e^{x-t} u(t)\, dt$$

(j) $T\colon C[2,3] \to C[2,3]$,

$$(Tu)(x) = \int_2^3 e^{x-t} u^2(t)\, dt$$

(k) $S_1\colon C[0,\infty] \to C[0,\infty]$

$$(S_1 u)(x) = u(x+1) - u(x)$$

(l) $L\colon A \to C[0,\infty]$,
where $A = \{\, u \in C[0,\infty]\colon \int_0^\infty |u(t)|\, dt < \infty \,\}$,

$$(Lu)(x) = \int_0^\infty e^{-xt} u(t)\, dt,$$

[Our restriction on A is just to insure that the integral exists. Lu is usually called the *Laplace transform* of u].

(m) $T\colon C[0,\infty] \to C[0,\infty]$

$$(Tu)(x) = a_2 u(x_2) + a_1 u(x+1) + a_0 u(x),$$

where the $a_k(x)$ are continuous functions.

(n) $T\colon C[0,1] \to C[0,1]$.

$$(Tu)(x) = 2xu(x).$$

(o) $T\colon \mathbb{R}^2 \to \mathbb{R}^1$

$$TX = |x_1 + x_2|\,, \text{ where } X = (x_1, x_2) \in \mathbb{R}^2.$$

[*Answers*: a,d,f,g,h,k,l,m,n are linear].

(2) (a) If $l(x)$ is a linear functional mapping $\mathbb{R}^1 \to \mathbb{R}^1$, prove that $l(x) = \alpha x$, where $\alpha = l(1)$.

(b) If $l(X)$ is a linear functional mapping $\mathbb{R}^n \to \mathbb{R}^1$, prove that $l(X) = \sum_{k=1}^{n} \alpha_k x_k$, where $X = (x_1, \ldots, x_n)$.

(3) Let $L_1\colon \mathbb{R}^1 \to \mathbb{R}^2$ be defined by

$$L_1 X = (x_1, 3x_1), \text{ where } X = (x_1) \in \mathbb{R}^1,$$

and $L_2\colon \mathbb{R}^2 \to \mathbb{R}^2$ be defined by

$$L_2 Y = (y_1 + y_2, y_1 + 2y_2), \text{ where } Y = (y_1, y_2) \in \mathbb{R}^2.$$

Compute $L_2 L_1 X_0$, where $X_0 = 2 \in \mathbb{R}^1$. Is $L_1 L_2$ defined?

(4) Let $A\colon \mathbb{R}^2 \to \mathbb{R}^2$ be defined by

$$AX = (x_1 + 3x_2, -x_1 - x_2), \qquad X = (x_1, x_2) \in \mathbb{R}^2$$

and $B\colon \mathbb{R}^2 \to \mathbb{R}^2$ by

$$BX = (-x_1 + x_2, 2x_1 + x_2).$$

a). Compute ABX, BAX, B^2X, A^2BX, and $(A+B)X$.

b). Find an operator C such that $CA = I$. [HINT: Let $CX = (c_{11}x_1 + c_{12}x_2, c_{21}x_1 + c_{22}x_2)$ and solve for c_{11}, c_{12}, etc.]

(5) Consider the operators $D\colon C^\infty \to C^\infty$, $(Du) = u'$ and $L\colon C^\infty \to C^\infty$, $(Lu)(x) = \int_0^x u(t)\, dt$.

(a) Show that $DL = I$, $LD = I - \delta$, where δ is the delta functional.

$$(L^2 u)(x) = \int_0^x \left(\int_0^2 u(t)\, dt \right) ds.$$

Integrate by parts to conclude that

$$(L^2 u)(x) + \int_0^x (x - t) u(t)\, dt.$$

(b) Observe that $D^2L^2 = D(DL)L = DIL = DL = I$. Use this observation to find a solution of the differential equation $D^2u = f$ for u, where $f \in C^\infty$. Solve the particular equation $(D^2u)(x) = \frac{1}{1+x^2}$

(6) Let $A\colon \mathbb{R}^2 \to \mathbb{R}^2$ be defined by

$$AX = (a_{11}x_1 + a_{12}x_2, a_{21}x_1 + a_{22}x_2),$$

and $B\colon \mathbb{R}^2 \to \mathbb{R}^2$ be defined by

$$BX = (b_{11}x_1 + b_{12}x_2, b_{21}x_1 + b_{22}x_2).$$

(a) Compute AB.

(b) Find a matrix B such that $AB = I$, that is, determine $b_{11}, b_{12}, \ldots$ in terms of $a_{11}, a_{12}, \ldots$ such that $AB = I$. [In the course of your computation, I suggest introducing a symbol, say Δ, for $a_{11}a_{12} - a_{12}a_{21}$ when that algebraic combination crops up.]

(7) In the plane $\mathbb{R}^2$, consider the operator R which rotates a vector by 90^o and the operator P projecting onto the subspace spanned by e (see fig). (a) Prove that R is linear. (b). Let $X = (x_1, x_2)$ be any point on $\mathbb{R}^2$. Compute PRX and RPX. Draw a sketch for the special case $X = (1,1)$.

(8) In $\mathbb{R}^3$, let A denote the operator of rotation through 90^o about the x_1-axis (so $A\colon (0,1,0) \to (0,0,1)$), B the operator of rotation through 90^o about the x_2-axis and C the operator of rotation through 90^o about the x_3-axis (see fig.) Prove these operators are linear (just do it for A). Show that $A^4 = B^4 = C^4 = I$, $AB \neq BA$, and that $A^2B^2 = B^2A^2$. Is it true that $ABAB = A^2B^2$?

(9) Let $\mathcal{P}$ denote the linear space of all polynomials in x. For $p \in \mathcal{P}$, consider the operators $Dp = \frac{dp}{dx}$ and $Lp = xp$. Show that $DL - LD = I$.

(10) (a) If $L_1L_2 = L_2L_1$, prove that

$$(L_1 + L_2)^2 = L_1^2 + 2L_1L_2 + L_2^2.$$

(b) If $L_1L_2 \neq L_2L_1$, then $(L_1 + L_2)^2 = ?$

(11) If L_1 and L_2 are operators such that $L_1L_2 - L_2L_1 = I$, prove the formula $L_1^nL_2 - L_2L_1^n = nL_1^{n-1}$, where $n = 1, 2, 3, \ldots$.

(12) If L_1 is a linear operator, $L_1 \colon V_1 \to V_2$ [or $L_1 \in \text{Hom}(V_1, V_2)$], and a is any scalar, define the operator $L = aL_1$ by the rule $LX = (aL_1)X = a(L_1X)$, where $X \in V_1$. Prove

(0). $L = aL_1$ is a linear operator, $L \colon V_1 \to V_1$.

(5). $a(bL_1) = (ab)L_1$, where a, b are any scalars.

(6). $1 \cdot L_1 = L_1$.

(7). $(a + b)L_1 = aL_1 + bL_1$.

(8). $a(L_1 + L_2) = aL_1 + aL_2$, where $L_2 \in \text{Hom}(V_1, V_2)$.

Coupled with Theorem 3, this exercise proves that the *set of all linear operators mapping one linear space in to another linear is itself a linear space*, that is, Hom(V_1, V_2) *is a linear space*.

(13) (a). In $\mathbb{R}^2$, let L denote the operator which rotates a vector by 90^o. Then $L \colon \mathbb{R}^2 \to \mathbb{R}^2$. If $X = (x_1, x_2) = x_1e_1 + x_2e_2$, where $e_1 = (1, 0)$ and $e_2 = (0, 1)$, write L as

$$LX = (a_{11}x_1 + a_{12}x_2, a_{21}x_1 + a_{22}x_2),$$

That is, find the coefficients $a_{11}, a_{12}, \ldots$. This gives two ways to represent L, as a rotation (geometrically), and by linear equations in terms of a particular basis (algebraically).

(b). In $\mathbb{R}^2$, consider the operator L of rotation through an angle α. Show that

$$Le_1 = (\cos\alpha, \sin\alpha), \quad Le_2 = (-\sin\alpha, \cos\alpha),$$

and then deduce that if $X = (x_1, x_2) = x_1e_1 + x_2e_2$,

$$LX = (x_1\cos\alpha - x_2\sin\alpha,\ x_1\sin\alpha + x_2\cos\alpha).$$

(14) Consider the space $\mathcal{P}_n$ of all polynomial of degree n. Define $L \colon \mathcal{P}_n \to \mathcal{P}_n$ as the translation operator $(Lp)(x) = p(x + 1)$, and $D \colon \mathcal{P}_n \to \mathcal{P}_n$ as the differentiation operator, $(Dp)(x) = \frac{dp}{dx}(x)$. Show that

$$L = I + D + \frac{D^2}{2!} + \cdots + \frac{D^{n-1}}{(n-1)!} + \frac{D^n}{n!}$$

(15) Consider the linear operators $L_1 = a_1D^2 + b_1D + c_1I$, and $L_2 = a_2D^2 + b_2D + e_2I$. Both L_1 and L_2 map the linear space of infinitely differentiable function into itself, $L_j \colon C^\infty \to C^\infty$. If the coefficients $a_1, a_2, b_1, \ldots$ are *constants*, prove that $L_1L_2 = L_2L_1$.

4.2 A Digression to Consider $au'' + bu' + cu = f$.

Essentially the only linear equation *you* can solve explicitly are linear algebraic equations, like two equations in two unknowns. Since our theory applies to much more general situations, we shall develop a different example for you to keep in the back of your minds along with that of linear algebraic equations. The example we have chosen has the additional virtue that it contains most of the solvable differential equations which arise anywhere. Watch closely because we shall be brief and with a high density of valuable ideas.

Problems concerning vibration or oscillatory phenomena are among the most important and significant ones which arise in applications. The simplest case is that of a *simple harmonic oscillator*. We have

A FIGURE GOES HERE

a mass m attached to a spring. Pull the mass back a little and watch it move back and forth, back and forth. These are oscillations. To make the situation simple, we assume that the spring has no mass and that the surface upon which the mass rests is frictionless. Let $u(t)$ denote the displacement of the center of gravity of the mass from the equilibrium position. Two *experimental* results are needed from physics.

1. *Newton's Second Law*: $m\ddot{u} = \sum F$, where $\sum F$ means the resultant of all the forces on the center of gravity of the mass (we assume all forces are acting horizontally).

2. *Hooke's Law*: If a spring is not stretched too far, then the force it exerts is proportional to the displacement,

$$F = -ku, \qquad k > 0.$$

We chose the minus sign since if a spring is displaced, the force it exerts is in the direction *opposite* to the displacement. [Under larger displacements, actually

$$F(u) = a_1u + a_2u^2 + a_3u^3 + \ldots$$

-where $a_0 = F(0) = 0$. If the displacement u is small, the lowest term in the Taylor series for $F(u)$ gives an adequate approximation. This is a more precise statement of Hooke's Law].

Putting these two results together, we find that

$$m\ddot{u} = -ku + F_1, \qquad (\text{notation} : \ddot{u} = \frac{d^2u}{dt^2})$$

where F_1 represents all of the remaining forces on the mass. One possible force (so far incorporated into F_1) is a so-called *viscous damping force*. It is of the form $F_v = -\mu\dot{u}$ where $\mu > 0$; at low velocities, this force is *experimentally* found to account for air resistance. It is directed opposite to the velocity, and increases as the speed does (speed $= \|\text{velocity}\|$). [Again, $F_v = b_1\dot{u} + b_2\dot{u}^2 + \ldots$, that is $F_v(\dot{u})$ is given by a Taylor series with $F_v(0) = 0$. At low speeds, the higher order terms can be neglected to yield a reasonable approximation.]

Thus, to our approximation,

$$m\ddot{u} = -ku - \mu\dot{u} + F_2,$$

where F_2 represents the forces yet unaccounted for. Let us assume that these remaining forces do not depend on the motion and are applied by the outside world. Then the force F_2 depends only on time, $F_2 = f(t)$. It is called the *applied* or *external force*. Newton's law gives

$$m\ddot{u} = -ku - \mu\dot{u} + f(t),$$

or

$$Lu\colon = a\ddot{u} = b\dot{u} + cu = f(t),$$

where $a = m$, $b = \mu$, and $c = k$. For the purposes of our discussion, we shall assume that k and μ do not depend on time. Then a, b and c are non negative constants.

In order to determine the motion of the mass, we must solve the ordinary differential equation $Lu = f$ for u. Have we given enough information to determine the solution? In other words, is the solution unique? For any physically reasonable problem, we expect the mathematical model has a unique solution since (neglecting quantum mechanical effects) once we let the mass go, it will certainly move in one particular way, the same way every time we perform the same experiment. It is clear that the motion will depend on the initial position $u(t_0)$. But if two masses have the same initial position, the resulting motion will still be different if their initial velocities $\dot{u}(t_0)$ are different. Thus we must also specify the initial velocity $\dot{u}(t_0)$ as well as the initial position $u(t_0)$. Are these sufficient to determine the motion? Yes, however that requires proof. What must be proved is that if we have two solutions $u_1(t)$ and $u_2(t)$ of the same ordinary differential equation (1) and if their initial positions and velocities coincide, then the solutions coincide, $u_1 = u_2$ for all later time, $t \geq t_0$.

Theorem 4.6 *(Uniqueness). Let $u_1(t)$ and $u_2(t)$ be two solutions of the ordinary differential equation*

$$Lu\colon = a\ddot{u} + b\dot{u} + cu = f(t),$$

where a, b, and c are constants, $a > 0, b \geq 0, c \geq 0$. If $u_1(t_0) = u_2(t_0)$, and $\dot{u}_1(t_0) = \dot{u}_2(t_0)$, then $u_1(t) = u_2(t)$ for all $t \geq 0$, in other words, the solution is uniquely determined by the initial position and velocity.

REMARK: The theorem is true under much more general conditions - as we shall prove in Chapter 6.

PROOF: Let $w(t) = u_2(t) - u_1(t)$. We shall show that $w(t) \equiv 0$ for all $t \geq t_0$. Now

$$Lw = L(u_2 - u_1) = Lu_2 - Lu_1 = f - f = 0,$$

that is,

$$a\ddot{w} + b\dot{w} + cw = 0 \tag{4-14}$$

Furthermore

$$w(t_0) = 0 \text{ and } \dot{w}(t_0) = 0, \tag{4-15}$$

since $w(t_0) = u_2(t_0) - u_1(t_0) = 0$, and $\dot{w}(t_0) = \dot{u}_2(t_0) - \dot{u}_1(t_0) = 0$. This reduces the question to showing that if $Lw = 0$, and if w has zero initial position and velocity, then in fact $w \equiv 0$.

The trick is to introduce a new function, $E(t)$, associated with (2) (which happens to be the total energy of the system)

$$E(t) = \frac{1}{2}a\dot{w}^2 + \frac{1}{2}cw^2.$$

How does this function change with time? We compute its derivative.

$$\dot{E}(t) = a\dot{w}\ddot{w} + cw\dot{w} = \dot{w}(a\ddot{w} + cw).$$

Using (2) we know that $a\ddot{w} + cw = -b\dot{w}$. Therefore

$$\dot{E}(t) = -b\dot{w}^2 \leq 0 \qquad (\text{since } b \geq 0)$$

[Thus energy is dissipated $(b > 0)$ - or conserved $\dot{E} = 0$ in the special case of no damping $(b = 0)$.] Consequently

$$E(t) \leq E(t_0) \quad \text{for all } t \geq t_0 \tag{4-16}$$

Now observe that for the mechanical system associated with w, we have $E(t_0) = \frac{a}{w}\dot{w}^2(t_0) + \frac{c}{2}w^2(t_0) = 0$. Furthermore, it is obvious from the definition of $E(t)$ (since a and c are positive) that $0 \leq E(t)$. Substitution of this information into (4) reveals

$$0 \leq E(t) \leq 0 \qquad \text{for all } t \geq t_0.$$

This proves $E(t) \equiv 0$ for all $t \geq t_0$, which in turn implies $w(t) \equiv 0$—again from the definition of $E(t)$. Our proof is completed. We have taken some care since all of our uniqueness proofs will use essentially no additional ideas. A more general case (a, b and c still constants but not necessarily positive) will be treated in Exercise 9.

Having proved that there is *at most one* solution of the initial value problem

$$\begin{aligned} Lu\colon = a\ddot{u} + b\dot{u} + cu = f(t) &\qquad \text{(differential equation)} \\ u(t_0) = \alpha \text{ and } \dot{u}(t_0) = \beta &\qquad \text{(initial conditions)} \end{aligned}$$

we must now prove there is *at least one* solution. This is the question of existence. For the special equation (5), the solution is shown to exist by explicitly exhibiting it. In the case of more complicated equations we are not as fortunate and must content ourselves with just showing that a unique solution exists but cannot exhibit it in closed form.

It is easiest to begin with the homogeneous equation $Lu = 0$, that is, find a solution of

$$a\ddot{u} + b\dot{u} + cu = 0 \quad \text{with } u(t_0) = \alpha, \text{ and } \dot{u}(t_0) = \beta.$$

Without motivation, let us see what the substitution $u(t) = e^{\lambda t}$ yields. Here λ is a constant. We must compute $Le^{\lambda t}$.

$$Le^{\lambda t} = (a\lambda^2 + b\lambda + c)e^{\lambda t}.$$

Can λ be chosen so that $e^{\lambda t}$ is a solution of $Lu = 0$? Since $e^{\lambda t} \neq 0$ for any t, this means, is it possible to pick λ so that $a\lambda^2 + b\lambda + c = 0$? Yes. In fact that "quadratic equation formula" yields two roots

$$\lambda_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \quad \lambda_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

of the *characteristic polynomial* $p(\lambda) = a\lambda^2 + b\lambda + c$. Notice that we have assumed $a \neq 0$. Thus, two solutions of the homogeneous equation are

$$u_1(t) = e^{\lambda_1 t} \quad \text{and} \quad u_2(t) = e^{\lambda_2 t}.$$

Since the operator L is linear, every linear combination of solutions is also a solution, $L(Au_1 + Bu_2) = ALu_1 + BLu_2 = 0$. Therefore $u(t) = Au_1(t) + Bu_2(t)$ is a solution of the homogeneous equation $Lu = 0$ for any choice of the scalars A and B.

What about the initial conditions $u(t_0) = \alpha$, $\dot{u}(t_0) = \beta$; can they be satisfied by picking the constants A and B suitably? Let us try. We want to pick A and B so that

$$\begin{aligned} &Ae^{\lambda_1 t_0} + Be^{\lambda_2 t_0} = \alpha \qquad (u(t_0) = \alpha) \\ &A\lambda_1 e^{\lambda_1 t_0} + B\lambda_2 e^{\lambda_2 t_0} = \beta \quad (\dot{u}(t_0) = \beta). \end{aligned}$$

These equations can be solved as long as

$$0 \neq \lambda_2 e^{(\lambda_1+\lambda_2)t_0} - \lambda_1 e^{(\lambda_2+\lambda_1)t_0} = (\lambda_2 - \lambda_1)e^{(\lambda_1+\lambda_2)t_0},$$

which means $\lambda_1 \neq \lambda_2$ or $b^2 - 4ac \neq 0$. [The linear equations $Ar_1 + Bs_1 = \alpha$, $Ar_2 + Bs_2 = \beta$ can be solved for A and B if $r_1 s_2 - r_2 s_1 \neq 0$]. Before dealing with the degenerate case $b^2 - 4ac = 0$, let us consider an

Example: Solve $\ddot{u} + 3\dot{u} + 2u = 0$ with the initial conditions $u(0) = 1$ and $\dot{u}(0) = 0$. If we seek a solution of the form $u(t) = e^{\lambda t}$, the characteristic polynomial is $\lambda^2 + 3\lambda + 2 = 0$, and has roots $\lambda_1 = -1$, $\lambda_2 = -2$. Therefore $u(t) = Ae^{-1} + Be^{-2t}$ is a solution. Since $\lambda_1 \neq \lambda_2$, we can solve for A and B by using the initial conditions. We find

$$\begin{aligned} A + B &= 1 \qquad (u(0) = 1), \\ -A - 2B &= 0 \qquad (\dot{u}(0) = 0). \end{aligned}$$

These two equations yield $A = 1$, $B = -1$. Thus

$$u(t) = 2e^{-t} - e^{-2t}$$

is the unique solution of our initial value problem.

The degenerate case $b^2 - 4ac = 0$ must be discussed separately. In this case $\lambda_1 = \lambda_2 = -\frac{b}{2a}$, so the two solutions $e^{\lambda_1 t}$ and $e^{\lambda_2 t}$ are really the same solution. Without motivation (but see Exercise 12) we claim that $te^{\lambda_1 t}$ is also a solution. This is easy to verify by a calculation.

$$\begin{aligned} L(te^{\lambda_1 t}) &= a(t\lambda_1^2 e^{\lambda_1 t} + 2\lambda_1 e^{\lambda_1 t}) + b(e^{\lambda_1 t} + \lambda_1 te^{\lambda_1 t}) + cte^{\lambda t} \\ &= (a\lambda_1^2 + b\lambda_1 + c)te^{\lambda t} + (2a\lambda_1 + b)e^{\lambda_1 t} \end{aligned} \tag{4-17}$$

Since $(a\lambda_1^2 + b\lambda_1 + c) = 0$ by definition of λ_1, and $\lambda_1 = -\frac{b}{2a}$ in our special case, both terms on the right vanish. Hence both $u_1(t) = e^{\lambda_1 t}$ and $u_2(t) = te^{\lambda_1 t}$ are solutions of $Lu = 0$ (if $b^2 - 4ac = 0$), so $u(t) = Ae^{\lambda_1 t} + Bte^{\lambda_1 t}$ is a solution for any choice of A and B. It is possible to pick A and B to satisfy arbitrary initial conditions $u(t_0) = \alpha$, $\dot{u}(t_0) = \beta$.

$$\begin{aligned} &Ae^{\lambda_1 t_0} + Bt_0 e^{\lambda_1 t_0} = \alpha, \qquad (u(t_0) = \alpha) \\ &A\lambda_1 e^{\lambda_1 t_0} + B(1 + \lambda_1 t_0)e^{\lambda_1 t_0} = \beta \qquad (\dot{u}(t_0) = \beta). \end{aligned}$$

These can be solved for A and B since

$$0 \neq (1 + \lambda_1 t_0)e^{2\lambda_1 t_0} - \lambda_1 t_0 e^{2\lambda_1 t_0} = e^{2\lambda_1 t_0}.$$

Example: Solve $\ddot{u} + 6\dot{u} + 9u = 0$ with the initial conditions $u(1) = 2$, $\dot{u}(1) = -1$. Seeking a solution in the form $e^{\lambda t}$, we are led to the characteristic equation $\lambda^2 + 6\lambda + 9 = 0$, which has $\lambda_1 = -3$, $\lambda_2 = -3$, as roots. Therefore $u_1(t) = e^{-3t}$ is a solution of $Lu = 0$. Since $\lambda_1 = \lambda_2$, another solution is $u_2(t) = te^{-3t}$. Thus $u(t) = Ae^{-35} + Bte^{-3t}$ is a solution for any A and B. To solve for A and B in terms of the initial conditions, we must solve the algebraic equations

$$\begin{aligned} &Ae^{-3} + B \cdot 1 \cdot e^{-3} = 2, \qquad (u(1) = 2), \\ &-3Ae^{-3} + B(1 - 3)e^{-3} = -1, \qquad (\dot{u}(1) = -1). \end{aligned}$$

We find that $A = -3e^3$ and $B = 5e^3$. Thus

$$u(t) = -3e^3 e^{-3t} + 5e^3 te^{-3t},$$

or, equivalently,

$$u(t) = -3e^{-3(t-1)} + 5te^{-3(t-1)}.$$

Our results will now be collected as

Theorem 4.7 . *The initial value problem*

$$a\ddot{u} + b\dot{u} + cu = 0,\ a \neq 0, \quad \textit{with} \quad u(t_0) = \alpha,\ \dot{u}(t_0) = \beta,$$

where a, b, and c are constants, has a unique solution.

i) If $b^2 - 4ac \neq 0$, it is of the form

$$u(t) = Ae^{\lambda_1 t} + Be^{\lambda_2 t}.$$

ii) If $b^2 - 4ac = 0$, so $\lambda_1 = \lambda_2$, it is of the form

$$u(t) = Ae^{\lambda_1 t} + Bte^{\lambda_1 t}.$$

Here λ_1 and λ_2 are the roots of the characteristic equation $a\lambda^2 + b\lambda + c = 0$, and the constants A and B are determined from the initial conditions.

REMARK: We have omitted the condition $a > 0$, $b \geq 0$, $c \geq 0$ from our theorem since the construction presented to find a solution did not depend on this. Uniqueness for that case is treated as exercise 9, as we mentioned earlier.

Another

EXAMPLE: Solve $\ddot{u} - 2\dot{u} + 2u = 0$, with the initial conditions $u(0) = 1$, $\dot{u}(0) = 1$. The characteristic polynomial is $\lambda^2 - 2\lambda + 2 = 0$. Its roots are $\lambda_1 = 1 + i$, and $\lambda_2 = 1 - i$. Since $\lambda_1 \neq \lambda_2$, the solution is of the form $u(t) = Ae^{(1_i)t} + Be^{(1-i)t}$. From the initial conditions, we find that

$$\begin{aligned} &A + B = 1, \qquad (u(0) = 1), \\ &(1+i)A + (1-i)B = 1, \qquad (\dot{u}(0) = 1). \end{aligned}$$

Thus $A = \frac{1}{2}$, $B = \frac{1}{2}$, so

$$u(t) = \frac{1}{2}e^{(1+i)t} + \frac{1}{2}e^{(1-i)t}.$$

Recalling that $e^{x+iy} = e^x(\cos t + i \sin y)$, this solution may be written in a more familiar form:

$$u(t) = \frac{1}{2}e^t(\cos t + i \sin t) + \frac{1}{2}e^t(\cos t - i \sin t),$$

that is,

$$u(t) = e^t \cos t.$$

What has been done can be summarized elegantly in the language of linear spaces. We have sought a solution of a second order linear O.D.E., which we write as $Lu = 0$. It was found that every solution of this equation could be expressed as a linear combination of two specific solutions u_1 and u_2, $u(t) = Au_1(t) + Bu_2(t)$, where the constants A and B are uniquely determined from $u(t_0)$ and $\dot{u}(t_0)$. Thus, *the set of functions u which satisfy $Lu = 0$ form a two dimensional subspace of $\mathcal{D}(L) = C^2$*. The functions u_1 and u_2 span that subspace. If we call the set of all solutions of $Lu = 0$ the nullspace of L, $\mathcal{N}(L)$, then our result simply reads "dim $\mathcal{N}(L) = 2$". A particular solution of $Lu = 0$ is found by specifying $u(t_0)$ and $\dot{u}(t_0)$.

The inhomogeneous equation $Lu = f$ is treated by finding a coset of the nullspace of L. For if u_0 is a particular solution of the inhomogeneous equation $Lu_0 = f$, then $u = \tilde{u} + u_0$;

where $\tilde{u} \in \mathcal{N}(L)$, is also a solution since $Lu = L(\tilde{u} + u_0) = L\tilde{u} + Lu_0 = 0 + f = f$. Therefore, if one solution u_0 of the inhomogeneous equation $Lu = f$ is found, the general solution is $u = \tilde{u} + u_0$ where $\tilde{u} \in \mathcal{N}(L)$. In particular, the solution $\tilde{u} \in \mathcal{N}(L)$ can be chosen so that arbitrary initial conditions for u, $u(t_0) = \alpha$, $\dot{u}(t_0) = \beta$, can be met. We shall defer (until our systematic treatment of linear O.D.E.'s) presenting a general method for finding a solution u_0 of the inhomogeneous equation. In our example, the particular solution will be found by guessing.

EXAMPLE: Solve $Lu := \ddot{u} - u = 2t$, with the initial conditions $u(0) = -1$, $\dot{u}(0) = 3$. The homogeneous equation $Lu = 0$ has the general solution $\tilde{u}(t) = Ae^t + Be^{-t}$. We observe that the function $u_0(t) = -2t$ is a particular solution of the inhomogeneous equation, $Lu = 2t$. Thus $u(t) = Ae^t + Be^{-t}$. The initial conditions lead us to solve the following equations for A and B,

$$A + B = -1 \tag{4-18}$$
$$A - B - 2 = 3. \tag{4-19}$$

A computation gives $A = 2$, $B = -3$. Thus the solution of our problem is

$$u(t) = 2e^t - 3e^{-t} - 2t$$

It is routine to verify that this function $u(t)$ does satisfy the O.D.E. and initial conditions (you should verify the solutions to check for algebraic mistakes).

Exercises

(1) Solve the following homogeneous initial value problems,
   - (a). $\ddot{u} - u = 0, \quad u(0) = 0, \quad \dot{u}(0) = 1.$
   - (b). $\ddot{u} + u = 0, \quad u(0) = 1, \quad \dot{u}(0) = 0.$
   - (c). $\ddot{u} - 4\dot{u} + 5u = 0, \quad u(0) = -1, \quad \dot{u}(0) = 2.$
   - (d). $\ddot{u} + 2\dot{u} - 8u = 0, \quad u(2) = 3, \quad \dot{u}(2) = 0.$
   - (e). $\ddot{u} = 0, \quad u(0) = 7, \quad \dot{u}(0) = 3.$

(2) Solve the following inhomogeneous initial value problems by guessing a particular solution of the inhomogeneous equation. Check your answers.
   - (a) $\ddot{u} - u = t^2, \quad u(0) = 0, \quad \dot{u}(0) = 0$

     HINT: Try $u_0(t) = a_1 t^2 + a_2 t + a_3$ and solve for a_1, a_2, a_3.]
   - (b) $\ddot{u} - 4\dot{u} + 5u = \sin t, \quad u(0) = 1, \quad \dot{u}(0) = 0$ [HINT: Try $u_0(t) = a_1 \sin t + a_2 \cos t$.]

(3) Consider an undamped harmonic oscillator with a sinusoidal forcing term, $\ddot{u} + n^2 u = \sin \gamma t$. Find the general solution if $\gamma^2 \neq n^2$ [try $u_0(t) = a_1 \sin \gamma t + a_2 \cos \gamma t$ for a particular solution]. What happens if $\gamma \overset{+}{\to} - n$? This is called *resonance*.

(4) You shall discuss damping in this problem. Consider the equation $\ddot{u} + 2\mu\dot{u} + ku = 0$, where $\mu > 0$, and $k > 0$. We shall let $\gamma = \sqrt{|\mu^2 - k|}$.

(a) *Light damping* ($\mu^2 < k$). Show that the solution is

$$u(t) = e^{-\mu t}(A\cos\gamma t + B\sin\gamma t),$$

and sketch a rough graph for the case $A = 1$, $B = 0$. This is the kind of oscillation you want for a pendulum clock, with μ small.

(b) *Heavy damping* ($\mu^2 > k$). Show that the solution is

$$u(t) = e^{-\mu t}(Ae^{\gamma t} + Be^{-\gamma t}).$$

Show that $u(t)$ vanishes at most once. Sketch a graph for the two cases $A = B = 1$ and $A = -1, B = 3$. The first describes the oscillation of an ideal screen door, while the second describes the ideal oscillation of a slammed car door.

(5) It is often useful to study the oscillations described by $\ddot{u} + 2\mu\dot{u} + ku = 0$ by sketching the solution in the u, $\dot{u}$ plane - or *phase space* as it is called. Investigate the curves for heavily and lightly damped oscillators. Show that the curve for a heavily damped oscillator will be a straight line through the origin for special initial conditions. What does the phase space curve look like for an undamped oscillator ($\mu = 0$, $k > 0$)?

(6) Consider the linear operator $Lu = a\ddot{u} + b\dot{u} + cu$, where a, b, c are constants. We have seen that $Le^{rt} = p(r)e^{rt}$ where $p(r)$ is the characteristic polynomial.

(a) If r is *not* one of the roots of the characteristic polynomial, observe that you can find a particular solution of $Lu = e^{rt}$. What is it?

(b) If neither r_1 nor r_2 is a root of the characteristic polynomial, find a particular solution of $Lu = a_1e^{r_1t} + a_2e^{r_2t}$, where a_1 and a_2 are specified constants.

(c) Use this procedure to find a particular solution of

$$i)\ddot{u} - 4u = \cosh t, \qquad ii)\ddot{u} + 4u = \sin t$$

(7) (a) Imitate our procedure and develop a theory for the first order homogeneous O.D.E. $Lu := \dot{u} + bu = 0$, where b is a constant. In particular, you should prove that there *exists* a *unique* solution satisfying the initial condition $u(t_0) = \alpha$, and give a recipe for finding it. Use your recipe to solve $\dot{u} + 2u = 0$, $u(0) = 3$.

(b) And now you will show us how to find a particular solution of the inhomogeneous equation $Lu = f$, where $f(t)$ is some given continuous function and $Lu := \dot{u} + bu$. [HINT: Try to find a function $\mu(t)$ such that $\mu(\dot{u} + bu) = \frac{d}{dt}(\mu u)$. Then integrate $\frac{d}{dt}(\mu u) = \mu f$, and solve for u]. Use your method to find a particular solution for $\dot{u} + 2u = x$, and then a solution of the same equation which satisfies the initial condition $u(0) = 1$.

(8) Find a solution of $u''' - 2u'' - u' + 2u = 0$ which satisfies the initial conditions $u(0) = u'(0) = 0$, $u''(0) = 1$. [HINT: The cubic equation $\gamma^3 - 2\gamma^2 - \gamma + 2$ has roots $+1, -1$ and 2].

(9) You will prove the uniqueness theorem for the equation $\ddot{u} + b\dot{u} + cu = 0$, where b and c are any constants (we have let $a = 1$, because if it is not 1, just divide the whole equation by a). The trick is to reduce this to the special case $b \geq 0$, $c \geq 0$, already done.

(a) Show that in order to prove the solution of

$$\ddot{u} + b\dot{u} + cu = f, \text{ where } u(t_0) = \alpha,\ \dot{u}(t_0) = \beta$$

is unique, it is sufficient to prove that the only solution of

$$\ddot{w} + b\dot{w} + cw = 0,\ w(t_0) = 0,\ \dot{w}(t_0) = 0$$

is $w(t) \equiv 0$.

(b) Define $\varphi(t)$ by $w(t) = e^{\gamma t}\varphi(t)$. Observe: to prove $w = 0$, it is sufficient to prove $\varphi \equiv 0$ (here γ is any constant). Use the differential equation and initial conditions for w to find the differential equation and initial conditions for φ. Show that γ can be picked so that the D.E. for φ is

$$\ddot{\varphi} + \tilde{b}\dot{\varphi} + \tilde{0}\varphi = 0,$$

where $\tilde{b}$ and $\tilde{c}$ are positive. Deduce that $\varphi \equiv 0$, and from that, that $w \equiv 0$, completing the proof.

(10) A *boundary value problem* for the equation

$$u'' + bu' + cu = 0$$

is to find a solution of the equation with given boundary values, say $u(0) = \alpha$ and $u(1) = \beta$. Assume b and c are real numbers.

(a) Show that a solution of the boundary value problem always exists if $b^2 - 4c \geq 0$ (the case $b^2 - 4c = 0$ will have to be done separately).

(b) Prove that if $b^2 - 4c \geq 0$, the solution is unique too. [I suggest letting $u(t) = e^{\gamma t}v(t)$, and then choosing γ so that the equation satisfied by v is of the form $v'' + \tilde{c}v = 0$, where $\tilde{c} \leq 0$. The case $\tilde{c} = 0$ is trivial. If $\tilde{(c)} < 0$, can the solution have a positive maximum or negative minimum?]

(11) If a spring is hung vertically and a mass m placed at its end, an external force of magnitude mg due to gravity is placed on the system. Assume there are no dissipative forces of any kind.

(a) Set up the differential equation of motion. Remember that you must specify which is the positive direction.

(b) If the tip of the spring is displaced a distance d by placing the mass on it (no motion yet), so the *equilibrium position* is d below the unstretched end of the spring, show that the spring constant k is given by $k = mg/d$.

(c) Let the body *weigh* 32 pounds, and d be 2 feet. Find the subsequent motion if the body is initially displaced from rest one foot below its equilibrium position. [Take $|g| = 32$ ft/sec^2].

(12) * Consider $au'' + bu' + cu = 0$. If $\gamma_1 \neq \gamma_2$, are the roots of the characteristic equation, observe that the function

$$\tilde{u}(t) = \frac{e^{\gamma_1 t} - e^{\gamma_2 t}}{\gamma_1 - \gamma_2}$$

is also a solution (it is a linear combination of $e^{\gamma_1 t}$ and $e^{\gamma_2 t}$). Now pass to the limit $\gamma_2 \to \gamma_1$ (leave γ_1 fixed and let γ_2 move) by using the Taylor series for $e^{\gamma t}$. The function you get is then a "guess" for a second solution in the degenerate case $\gamma_1 = \gamma_2$. This supplies some motivation for the guess made earlier.

(13) * Consider $Lu \colon = u'' + 2u = f$, where f is given. You know how to solve $Lu = A \sin nx$ (Exercise 6). Find a particular solution to the general inhomogeneous equation in the interval $[-\pi, \pi]$ by expanding f in a Fourier series and then use superposition. Apply this to solve $u'' + 2u = x$.

(14) Consider an undamped harmonic oscillator, whose motion is specified by $u(t)$, where $mu'' + ku = 0, k > 0$. Show that the solution $u(t) = A_1 \cos \sqrt{\frac{k}{m}} t + B_1 \sin \sqrt{\frac{k}{m}} t$ may be written in the form

$$u(t) = A \sin(wt + \theta),$$

where A is the *amplitude* of the oscillation, $w = 2\pi v$, v is the *frequency*, and θ is the *phase*. Show that $u(t)$ is periodic, $u(t+T) = u(t)$, where the *period* $T = 1/v$. Interpret the amplitude and phase and determine A, w, and θ in terms of A_1, B_1, k and m. [I suggest looking at a specific example and its graph first].

4.3 Generalities on $LX = Y$.

Undoubtedly the fundamental problem in the theory of linear (and nonlinear) operators is to determine the nature of the range of an operator L. One particular aspect of this is the vast problem of solving the equation

$$LX = Y$$

for X when Y is given to you. The question here is, "is a given Y in the range of L?", or "can we find some X such that $LX = Y$?" If one can solve the problem uniquely for any

Y, then the solution is written as

$$X = L^{-1}(Y),$$

where L^{-1} is the operator inverse to L, in the sense that $L^{-1}L = I$ (so to solve $LX = Y$, apply L^{-1}, $X = L^{-1}LX = L^{-1}Y$).

Let us give some examples, familiar and unfamiliar, of problems of the form $LX = Y$, where Y is given.

1. $LX = (2x_1 + 3x_2, x_1 + 2x_2), \quad X \in \mathbb{R}^2$,

$$L \colon \mathbb{R}^2 \to \mathbb{R}^2.$$

The problem of solving $LX = Y$ where $Y = (-1, 2) \in \mathbb{R}^2$ is that of solving the two equations

$$\begin{aligned} 2x_1 + 3x_2 &= -1 \\ x_1 + 2x_2 &= 2 \end{aligned}$$

for two unknowns $(x_1, x_2) = X$.

2. $Lu = u'' + 2u' + 3u$, where $u \in C^2$,

$$L \colon C^2 \to C.$$

The problem of solving $L(u) = x$ is that of solving the inhomogeneous ordinary differential equation

$$Lu \colon = u'' + 2u' + 3u = x$$

for $u(x)$.

3. $Lu = \int_0^\pi \cos(x - t) u(t)\, dt, \qquad u \in C[0, \pi]$.

You should check that L is a linear operator. The problem of solving $L(u) = \sin x$ is that of solving the integral equation

$$Lu \colon = \int_0^\pi \cos(x - t) u(t)\, dt = \sin x$$

for the function u. In this example, it is instructive to examine the range more closely. Since $\cos(x - t) = \cos x \cos t + \sin x \sin t$ and since functions of x are constant with respect to t integration, we see that Lu may be written as

$$Lu \colon = \cos x \int_0^\pi \cos t\, u(t)\, dt + \sin x \int_0^\pi \sin t\, u(t)\, dt,$$

or

$$Lu \colon = \alpha_1 \cos x + \alpha_2 \sin x,$$

where the *numbers* α_1 and α_2 are

$$\alpha_1 = \int_0^\pi (\cos t) u(t)\, dt; \quad \alpha_2 = \int_0^\pi (\sin t) u(t)\, dt.$$

Thus, the range of L is the linear space spanned by $\cos x$ and $\sin x$, which has dimension two. This linear operator L therefore maps the infinite dimensional space $C[0,\pi]$ into a finite (two) dimensional space. In order to even have a chance of solving $Lu = f$ for this operator L, we first check to see if f even lies in this two dimensional subspace (for if it doesn't, it is futile to go further). The particular function $\sin x$ does, so it is reasonable to look for a solution - which we shall not do right now (however there are infinitely many solutions, among them $u(x) = \frac{2}{\pi}\sin x$).

One particularly important equation which arises frequently is the *homogeneous equation*

$$LX = 0,$$

which is the special case $Y = 0$ of the *inhomogeneous equation*,

$$LX = Y.$$

Since L is a linear operator, there is no problem of our finding *one* solution of $LX = 0$ for $X = 0$ is a solution, the so-called *trivial solution* of the homogeneous equation. The problem is to find a non-trivial solution, or better yet, all solutions. In the previous section, this question was answered fully for the particular operator $Lu = au'' + bu' + cu$, where a, b, and c are constants. Many of our results there generalize immediately, as we shall see now.

DEFINITION: The set of all solutions of the homogeneous equation $LX = 0$ where L is a linear operator is called the *nullspace* of L. This nullspace of L, $\mathcal{N}(L)$, consists of all X in the domain of L which are mapped into zero by L,

$$L\colon \mathcal{N}(L) \to 0.\ \mathcal{N}(L) \subset \mathcal{D}(L).$$

We have called $\mathcal{N}(L)$ the null*space* of L, not the null *set* because of

Theorem 4.8 . *The nullspace of a linear operator $L\colon \mathcal{V}_1 \to \mathcal{V}_2$ is a linear space, a subspace of the domain of L.*

PROOF: Since the domain of L, $\mathcal{D}(L) = \mathcal{V}_1$, is a linear space and $\mathcal{N}(L) \subset \mathcal{D}(L)$, by Theorem 2, p.142 all we need show is that the set $\mathcal{N}(L)$ is closed under multiplication by scalars and under addition of vectors. Say X_1 and $X_2 \in \mathcal{N}(L)$. Then $LX_1 = 0$ and $LX_2 = 0$. We must show that $L(aX_1) = 0$ for any scalar a, and that $L(X_1 + X_2) = 0$. But $L(aX_1) = aL(X_1) = a \cdot 0 = 0$, and $L(X_1 + X_2) = LX_1 + LX_2 = 0 + 0 = 0$. Thus $\mathcal{N}(L)$ is a subspace of $\mathcal{D}(L) = \mathcal{V}_1$.

One important reason for examining the null space of a linear operator is because if $\mathcal{N}(L)$ is known, and if any *one* solution of the inhomogeneous equation is known, say $LX_1 = Y$ (where Y was given and X_1 is the solution we know), then *every* solution of the inhomogeneous equation is of the form $\tilde{X} + X_1$, where $\tilde{X} \in \mathcal{N}(L)$. In other words every solution of $LX = Y$ is in $\mathcal{N}(L) + X_1$, the X_1 coset of the subspace $\mathcal{N}(L)$.

Theorem 4.9 . *Let $L\colon \mathcal{V}_1 \to \mathcal{V}_2$ be a linear operator. If X_1 and X_2 are any two solutions of the inhomogeneous equation $LX = Y$, where Y is given, then $X_2 - X_1 \in \mathcal{N}(L)$, or $X_2 = \tilde{X} + X_1$ where $\tilde{X} \in \mathcal{N}(L)$.*

PROOF: Let $\tilde{X} = X_2 - X_1$. We shall show that $\tilde{X} \in \mathcal{N}(L)$.

$$L\tilde{X} = L(X_2 - X_1) = LX_2 - LX_1 = Y - Y = 0.$$

By using this theorem, we see that if *all* solutions of the homogeneous equation $LX = 0$ are known - the nullspace of L—and if *one* solution of the inhomogeneous equation $LX_1 = Y$ is known, then *all* of the solutions of the inhomogeneous equation are known. This *solution set of the inhomogeneous equation is the X_1 coset of $\mathcal{N}(L)$.*

EXAMPLE: 1 Let $L\colon \mathbb{R}^2 \to \mathbb{R}^2$ be defined by

$$LX = (x_1 + x_2, x_1 - x_2) \in \mathbb{R}^2$$

Then $\mathcal{N}(L)$ is the set of all points in $\mathbb{R}^2$ such that $LX = 0$, that is, which satisfy the equations

$$\begin{aligned} x_1 + x_2 &= 0 \\ x_1 - x_2 &= 0 \end{aligned}$$

Thus the nullspace of L consists of the intersection of the two lines $x_1 + x_2 = 0$, $x_1 - x_2 = 0$. The only point on both lines is 0. Thus $\mathcal{N}(L)$ is just the point 0. To solve the inhomogeneous equation $LX = Y$, where $Y = (1,1)$.

$$x_1 + x_2 = 1,\ x_1 - x_2 = 1,$$

we find one solution of it, $X_1 = (1,0)$. Then every solution of the inhomogeneous equation is of the form $X = \tilde{X} + X_1$, where $\tilde{X} \in \mathcal{N}(L)$. But since $\tilde{X} + 0$ is the only point in $\mathcal{N}(L)$, every solution is of the form $X = 0 + X_1 = X_1$. Thus every solution is exactly X_1, which is the *unique* solution of $LX = Y$. This situation is a general one. Again, we also saw this for $Lu = au'' + bu' + cu$.

Theorem 4.10 . *If the nullspace $\mathcal{N}(L)$ of the linear operator consists only of 0, then the solution of the inhomogeneous equation $LX = Y$ (if a solution exists) is unique. (Thus, if the nullspace contains only 0, then L is injective).*

PROOF: Say there were two solution X_1 and X_2. Then $LX_1 = Y$ and $LX_2 = Y$, which implies $L(X_2 - X_1) = LX_2 - LX_1 = Y - Y = 0$. Therefore $(X_2 - X_1) \in \mathcal{N}(L)$. Since the only element of $\mathcal{N}(L)$ is 0, $X_2 - X_1 = 0$, or, $X_1 = X_2$. In other words, the two solutions are the same.

EXAMPLE: 2 Let $L\colon C^2 \to C$ be defined on functions $u \in C^2$ by

$$Lu\colon = a(x)u'' + b(x)u' + c(x)u.$$

The nullspace of L consists of all solutions of the homogeneous equation $Lu = 0$. It turns out (see chapter 6) - as in the constant coefficient case - that every solution of this homogeneous O.D.E. has the form $u = Au_1 + Bu_2$, where u_1 and u_2 are any two linearly independent solutions of the equation, and where A and B are constants. Thus $\mathcal{N}(L)$ is a two dimensional space spanned by u_1 and u_2. If u_1 is a particular solution of the inhomogeneous equation $Lu_1 = f$, then *all* the solutions of $Lu = f$ are just the elements of the u_1 coset of $\mathcal{N}(L)$, that is, functions of the form $u = \tilde{u} + u_1$, where $\tilde{u} \in \mathcal{N}(L)$.

With every linear operator $L\colon V_1 \to V_2$, $V_1 = \mathcal{D}(L)$, we have associated two other linear spaces, the nullspace $\mathcal{N}(L) \subset \mathcal{D}(L) = V_1$ and range $\mathbb{R}(L) \subset V_2$. There is a valuable and elegant way to connect $\mathcal{D}(L), \mathcal{N}(L)$ and $\mathcal{R}(L)$. The result we are aiming at is certainly the most important theorem of this section.

We know that $\mathcal{R}(L) \subset V_2$. The space V_2 may be of arbitrarily high dimension. However, since $\mathcal{R}(L)$ is the image of $\mathcal{D}(L)$, we suspect that $\mathcal{R}(L)$ can take up "no more room" then $\mathcal{D}(L)$. To be more precise,

$$\dim \mathcal{R}(L) \leq \dim \mathcal{D}(L).$$

Thus, for example, if $L\colon \mathbb{R}^2 \to \mathbb{R}^{17}$, we expect that the range of L is a subspace of dimension no more than two in $\mathbb{R}^{17}$. Not only is this a justifiable expectation, but even more is true.

If Dim $\mathcal{R}(L) =$Dim $\mathcal{D}(L)$, essentially all of $\mathcal{D}(L)$ is carried over under the mapping. But if Dim $\mathcal{R}(L) <$ Dim $\mathcal{D}(L)$, what has happened to the remainder of $\mathcal{D}(L)$? Let us look at $\mathcal{N}(L) \subset \mathcal{D}(L)$. The elements of $\mathcal{N}(L)$ are all squashed into the zero element of V_2. In other words, a set of dim $\mathcal{N}(L)$ in $V_1 = \mathcal{D}(L)$ is mapped into a set of dimension zero in V_2. Does L decompose $\mathcal{D}(L) + V_1$ into two parts, $\mathcal{N}(L)$ and a complement $\mathcal{N}(L)'$ such that L maps $\mathcal{N}(L)$ into zero and the dimension of the remainder, $\mathcal{N}(L)'$, is preserved under L (so dim $\mathcal{N}(L)' = \dim \mathcal{R}(L)$). Of course,

A FIGURE GOES HERE

Theorem 4.11 . *Let the linear operator L map $V_1 = \mathcal{D}(L)$ into V_2. If $\mathcal{D}(L)$ has finite dimension, then*

$$\dim \mathcal{D}(L) = \dim \mathcal{R}(L) + \dim \mathcal{N}(L).$$

PROOF: Let $\mathcal{N}(L)'$ be a complement of $\mathcal{N}(L)$ (cf. pp. 163a-d). Since $\dim \mathcal{N}(L) + \dim \mathcal{N}(L)' = \dim \mathcal{D}(L)$, it is sufficient to prove that $\dim \mathcal{N}(L)' = \dim \mathcal{R}(L)$.

For $X \in V_1$, we can write $X = X_1 + X_2$, where $X_1 \in \mathcal{N}(L)$ and $X_2 \in N(L)'$. Now $LX = LX_1 + LX_2$, so the image of $\mathcal{D}(L)$ is the same as the image of $\mathcal{N}(L)'$. In addition, if $X_2 \in \mathcal{N}(L)'$, then $LX_2 = 0$ if and only if $X_2 = 0$, merely because $\mathcal{N}(L)'$ is a complement of the nullspace. Let $\{\theta_1, \ldots, \theta_k\}$ be a basis for $\mathcal{N}(L')$. If $X_2 \in \mathcal{N}(L)'$, we can write $X_2 = \sum_{j=1}^{k} a_j\theta_j$, and $LX_2 = \sum_{j=1}^{k} a_j L\theta_j$. Let $L\theta_1 = Y_1, L\theta_2 = Y_2, \ldots, L\theta_k = Y_k$. Since the image of $\mathcal{N}(L)'$ is $\mathcal{R}(L)$, the vectors $Y_1, \ldots, Y_k$ span $\mathcal{R}(L)$. Thus, $\dim \mathcal{R}(L) \leq k = \dim \mathcal{N}(L)'$.

To show that there is equality, $\dim \mathbb{R}(L) = \dim \mathcal{N}(L)'$, we prove that $Y_1, \ldots, Y_k$ are linearly independent. If $c_1Y_1 + \cdots + c_kY_k = 0$, then $0 = c_1L\theta_1 + \cdots + c_kL\theta_k = L(c_1\theta_1 + \cdots + c_k\theta_k) = L\tilde{X}$ where $\tilde{x} = c_1\theta_1 + \cdots + c_k\theta_k \in \mathcal{N}(L)'$. However for any $\tilde{X} \in \mathcal{N}(L)'$, we know $L\tilde{x} = 0$ implies that $\tilde{X} = 0$. The linear independence of $\theta_1, \ldots, \theta_k$ further shows that $c_1 = c_2 = \cdots = c_k = 0$. The hypothesis $c_1Y_1 + \cdots + c_kY_k = 0$ has led us to conclude that the c_j's are all zero, that is, the Y_j's are linearly independent. Therefore $\dim \mathcal{R}(L) = \dim \mathcal{N}(L)'$. Coupled with our first relationship, this proves the result.

Corollary 4.12 *:* $\dim \mathbb{R}(L) \leq \dim \mathcal{D}(L)$.

PROOF: $\dim \mathcal{N}(L) \geq 0$.

Two examples, one an illustration, the other an application.

EXAMPLE: 1 Consider a projection operator, P_A, mapping vectors from E^n into a subspace A of $\mathbb{R}^n$, where the $\dim A = m < n$. Let us first show that P_A is a linear operator. If $e_1, \ldots e_m$ is an orthonormal basis for A, then for any X and Y in $\mathbb{R}^n$,

$$\begin{aligned} P_A(X + Y) &= \sum_{k=1}^{m} \langle X + Y, e_k\rangle e_k = \sum_{k=1}^{m} (\langle X, e_k\rangle + \langle Y, e_k\rangle) e_k \\ &= \sum_{k=1}^{m} \langle X, e_k\rangle e_k + \sum_{k=1}^{m} \langle Y, e_k\rangle e_k = P_A X + P_A Y. \end{aligned}$$

Similarly, $P_A(aX) = aP_AX$ for every scalar a. Thus the projection operator is a linear operator. Since $\mathbb{R}(P_A) = A$ and $\dim A = m$, while $\dim \mathbb{R}^n = n$, we conclude that $\dim \mathcal{N}(P_A) = n - m$. This could have been arrived at immediately since P_A will certainly map everything perpendicular to A, that is $A^\perp$, into 0 (see fig. illustrating the case $\mathbb{R}^2 \to A$, where A is a line). Thus $\mathcal{N}(P_A) + A^\perp$, so $\dim \mathcal{N}(P_A) = \dim A^\perp = n - m$.

EXAMPLE: 2 Define $L\colon \mathbb{R}^n \to \mathbb{R}^k$ by,

$$LX = (a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n,\ a_{21}x_1 + \cdots + a_{2n}x_n, \cdots, a_{kl}x_1 + a_{k2}x_2 + \cdots + a_{kn}x_n)$$

where $X = (x_1, x_2, \ldots x_n) \in \mathbb{R}^n$. If we let $Y = (y_1, \ldots, y_k) \in \mathbb{R}^k$, then writing $Y = LX$, the linear operator L may be defined by the k equations (for $y_1, \ldots, y_k$) in n "unknowns" $(x_1, \ldots, x_n)$,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= y_2 \\ &\vdots \\ a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kn}x_n &= y_k. \end{aligned}$$

The problem of solving $LX = Y$, where Y is given, is that of solving k equations with n "unknowns".

Consider the *special case* $k < n$, when there are less equations than unknowns. Since the range of L is contained in $\mathbb{R}^k$, $\mathcal{R}(L) \subset \mathbb{R}^k$, then $\dim \mathcal{R}(L) \leq \dim \mathbb{R}^k = k$. Because $\mathcal{D}(L) = \mathbb{R}^n$, we also know that $\dim \mathcal{D}(L) = \dim \mathbb{R}^n = n$. Thus

$$\dim \mathcal{N}(L) = \dim \mathcal{D}(L) - \dim \mathcal{R}(L) \geq n - k > 0.$$

However *if* $\dim \mathcal{N}(L) > 0$, then $\mathcal{N}(L)$ must contain something other than zero. Thus *there is at least one non-trivial solution* $\tilde{X}$ *of the homogeneous equation,* $L\tilde{X} = 0$. Since $a\tilde{X}$ is also a solution, where a is any scalar, there are, in fact an *infinite number of solutions.*

Notice that the above was a *non-constructive* existence theorem. We proved that a solution does exist but never gave a recipe to obtain it. One consequence of this result is that, if $\dim \mathcal{N}(L) > 0$, and *if a solution of the inhomogeneous equation* $LX = Y$ *exists, it is not unique*; for if $LX_1 = Y$, then also $L(X_1 + \tilde{X}) = Y$, where $\tilde{X}$ is any solution of the homogeneous equation.

In the *special case* $n = k$, *and* $\dim \mathcal{N}(L) = 0$ a fascinating (and non-constructive) theorem falls out of Theorem 11: *the inhomogeneous equation* $LX = Y$ *always has a solution and the solution is unique.* Put in more conventional terms, *if there are the same number of equations as unknowns, and if the only solution of the homogeneous equation is zero, then the inhomogeneous equation always has a unique solution. Thus, if* $n = k$, *uniqueness implies existence.*

Since $\dim \mathcal{N}(L) = 0$, then $\dim \mathcal{R}(L) = \dim \mathcal{D}(L) = n$. However $L\colon \mathbb{R}^n \to \mathbb{R}^n$ in this case $(n = k)$. Since $\mathcal{R}(L) \subset \mathbb{R}^n$ and $\dim \mathcal{R}(L) = n$, we see that $\mathcal{R}(L)$ must be all of $\mathbb{R}^n$, that is, every $Y \in \mathbb{R}^n$ is in the range of L, which means that the inhomogeneous equation $LX = Y$ is solvable for every $Y \in \mathbb{R}^n$. Theorem 10 gives the uniqueness. We shall obtain a better theorem later.

REMARK: Some people refer to $\dim \mathcal{R}(L)$ as the *rank* of the linear operator L. We shall, however, refer to it as the dimension of the range of L.

If $L_1\colon V_1 \to V_2$ and $L_2\colon V_2 \to V_3$, it is easy to make a few statements about $\dim \mathcal{R}(L_2L_1)$. $\blacksquare$

Theorem 4.13 . *If* $L_1\colon V_1 \to V_2$ *and* $L_2\colon V_3 \to V_4$*, when* $V_2 \subset V_3$*, (so* L_2L_1*) is defined), then*

$$\dim \mathcal{R}(L_2L_1) \le \min(\dim \mathcal{R}(L_1), \dim \mathcal{R}(L_2)).$$

PROOF: The last corollary states that an operator is like a funnel with respect to dimension: the dimension can only get smaller or remain the same. After passing through two funnels, we obtain no more than the smallest allowed through. One might think that there should be equality in the formula. That this is not the case can be seen from the possibility illustrated in the figure. Only the shaded stuff gets through.

Exercises

(1) Let $L\colon \mathbb{R}^n \to \mathbb{R}^n$ be defined by

$$LX = (x_1, x_2, \ldots, x_k, 0, \ldots, 0),$$

where $X = (x_1, x_2, \ldots x_n) \in \mathbb{R}^n$. Describe $\mathcal{R}(L)$ and $\mathcal{N}(L)$. Compute $\dim \mathcal{R}(L)$ and $\dim \mathcal{N}(L)$.

(2) (a) Describe the range and nullspace of the linear operator $L\colon \mathbb{R}^3 \to \mathbb{R}^3$ defined by

$$LX = (x_1 + x_2 - x_3,\ 2x_1 - x_2 + x_3,\ x_2 - x_3),\ X = (x_1, x_2, x_3) \in \mathbb{R}^3.$$

(b) Compute $\dim \mathcal{R}(L)$ and $\dim \mathcal{N}(L)$.

(c) Is $(1, 2, 0) \in \mathcal{R}(L)$? Is $(1, 2, 1) \in \mathcal{N}(L)$?
Is $(1, 2, 2) \in \mathcal{N}(L)$? Is $(0, -1, -1) \in \mathcal{N}(L)$?

(3) Let $A = \{\, u \in C^2[0, 2]\colon u(0) = u(1) = 0 \,\}$, and define $L\colon A \to C[0, 1]$ by $Lu = u'' + b(x)u' - u$, where $b(x)$ is some continuous function. Prove $\mathcal{N}(L) = 0$. [HINT: If $u \in \mathcal{N}(L)$, can u have a positive maximum or negative minimum?]

(4) Consider the linear operator $L\colon C[0, 1] \to C[0, 1]$ defined by

$$(Lu)(x) = u(x) + 2\int_0^1 e^{x-t} u(t)\, dt$$

(a) Find the nullspace of L.

(b) Solve $Lu = 3e^x$. Is the solution unique?

(c) Show that the unique solution of $Lu = f$, where $f \in C[0, 1]$ is

$$u(x) = f(x) - ce^x, \text{ where } c = \frac{2}{3}\int_0^1 e^{-t} f(t)\, dt.$$

(5) Let $L\colon V \to V$ (so L^k is defined for $k = 0, 1, 2 \ldots$). Prove that

(a) $\mathcal{R}(L) \subset \mathcal{N}(L)$ if and only if $L^2 = 0$.

(b) $\mathcal{N}(L) \subset \mathcal{N}(L^2) \subset \mathcal{N}(L^3) \subset \ldots$

(c) $\mathcal{N}(L)' \supset \mathcal{N}(L^2)' \supset \mathcal{N}(L^3)' \supset \ldots$.

(6) If $L_1\colon V_1 \to V_2$ and $L_2\colon V_3 \to V_4$ where $V_2 \subset V_3$, Theorem 12 gives an upper bound for $\dim \mathcal{R}(L_2L_1)$.

(a) Prove the corresponding lower bound

$$\dim \mathcal{R}(L_2L_1) \ge \dim \mathcal{R}(L_1) + \dim \mathcal{R}(L_2) - \dim V_3.$$

[HINT: Prove the equivalent inequality $\dim \mathcal{R}(L_1) \le \dim \mathcal{R}(L_2L_1) + \dim \mathcal{N}(L_2)$ by letting $\tilde{V} = \mathcal{R}(L_1)$ and applying Theorem 11 to L_2 defined on $\tilde{V}$].

(b) Prove: if $\dim \mathcal{N}(L_2) = 0$, then

$$\dim \mathcal{R}(L_2L_1) = \dim \mathcal{R}(L_1).$$

(c) If $\dim \mathcal{N}(L_1) = 0$, is it then true that $\dim \mathcal{R}(L_2L_1) = \dim \mathcal{R}(L_2)$? Proof or counterexample.

(d) If $\dim V_1 = \dim V_2 = \dim V_3$ and $\dim \mathcal{N}(L_1) = 0$, is it true that $\dim \mathcal{R}(L_2L_1) = \dim \mathcal{R}(L_2)$? Proof or counterexample.

(7) If L_1 and L_2 both map $V_1 \to V_2$, prove

$$|\dim \mathcal{R}(L_1) - \dim \mathcal{R}(L_2)| \le \dim \mathcal{R}(L_1 + L_2).$$

(8) Consider the operator $L\colon C^2[0,\infty) \to C[0,\infty)$ defined by

$$Lu\colon = u'' + 3u' + 2u.$$

(a) Describe $\mathcal{N}(L)$. What is $\dim \mathcal{N}(L)$? Is $f(x) = \sin x \in \mathcal{R}(L)$?

(b) Consider the same operator L but mapping A into $C[0,\infty]$, where $A = \{\, u \in C^2[0,\infty)\colon u(0) + u'(0) = 0 \,\}$. Answer the same questions as part a).

(c) Same as b but $A = \{\, u \in C^2[0,\infty)\colon u(1) + u'(1) = 0 \,\}$ this time.

4.4 $L\colon \mathbb{R}^1 \to \mathbb{R}^n$. Parametrized Straight Lines.

Our study of particular linear operators begins with the most simple case: a linear operator which maps a one- dimensional space $\mathbb{R}^1$ into an n dimensional space $\mathbb{R}^n$. Since the dimension of the range of L is no greater than that of the domain R^1 and $\dim \mathbb{R}^1 = 1$, then

$$\dim \mathbb{R}(L) \le 1.$$

This proves

REMARK: If $L\colon \mathbb{R}^1 \to \mathbb{R}^n$, then the dimension of the range of L is either one or zero.

The case $\dim \mathcal{R}(L) = 0$ is trivial, for then L must map all of $\mathbb{R}^1$ into a single point, and that single point must be the origin since the range of L is a subspace. Thus, *if* $\dim \mathcal{R}(L) = 0$*, then* L *maps every point into 0.* Without change, the same holds if $L\colon V_1 \to V_2$ (where V_1 and V_2 are any linear spaces) and $\dim \mathcal{R}(L) = 0$. Not very profound.

If $\dim \mathcal{R}(L) = 1$, then the subspace $\mathcal{R}(L)$ in $\mathbb{R}^n$ is a one dimensional subspace in the n dimensional space $\mathbb{R}^n$, this is, $\mathcal{R}(L)$ is a "straight line" through the origin of $\mathbb{R}^n$. This straight line is determined if any one point $P \ne 0$ on it is known. Then there is a point $X_1 \in \mathbb{R}^1$ such that $LX_1 = P$. Since $\mathbb{R}^1$ is one dimensional it is spanned by any element other than zero, so every $X \in \mathbb{R}^1$ can be written as $X = sX_1$. Therefore, if X is any element of $\mathbb{R}^1$,

$$LX = L(sX_1) = sLX_1 = tP.$$

In other words, this last equation states that the range of L is a multiple of a particular vector P, that is, a straight line through the origin.

EXAMPLE:

If $L\colon \mathbb{R}^1 \to \mathbb{R}^2$ such that the point $X_1 = 2 \in \mathbb{R}^1$ is mapped into the point $P = (1,-2) \in \mathbb{R}^2$, then

$$L\colon X = s2 \to (s, -2s),$$

In particular, the point $X = 3 (s = \frac{3}{2})$ is mapped into the point $(\frac{3}{2}, -3)$.

A FIGURE GOES HERE

In applications, the domain $\mathbb{R}^1$ usually represents time, while the range represents the position of a particle. Then $L\colon \mathbb{R}^1 \to \mathbb{R}^n$ is an operator which specifies the position of a particle at a given time. Since L is linear and $L0 = 0$, the path of the particle must be a straight line which passes through the origin at $t = 0$. Later on in this section we shall show how to treat the situation of a straight line not through the origin, while in Chapter 7 we shall examine curved paths (non-linear operators).

EXAMPLE: This is the same example as before. $L\colon \mathbb{R}^1 \to \mathbb{R}^2$ is such that at time $t = 2 \in \mathbb{R}^1$ a particle is at the point $(1,-2)$ (while at $t = 0$ it is at the origin). At any time $t = s2$, the particle is at $(s, -2s)$. In particular, at $t = 3 (s = \frac{3}{2})$, the particle is at $(\frac{3}{2}, -3)$. It is also convenient to rewrite the position $(s, -2s)$ directly in terms of the time. Since $t = 2s$, the position at time t is $(\frac{t}{2}, -t)$. Thus we can write

$$L\colon t \to (\frac{t}{2}, -t),$$

which clearly indicates the position at a given time. If a point in the space $\mathbb{R}^2$ is specified by $Y = (y_1, y_2) \in \mathbb{R}^2$, then the operator can be written as

$$\begin{aligned} y_1 &= \frac{1}{2}t \\ y_2 &= -t. \end{aligned}$$

All of these are useful ways to write the operator L. In some situations, one might be more useful than another.

This brings us to an issue which perhaps seems a bit pedantic but can serve you well in times of need. How can we represent the operator in a picture? There are three distinct ways. Some clarity can be gained by distinguishing them carefully. The same ideas carry over immediately to nonlinear operators.

Our first picture has two parts. If $L\colon \mathbb{R}^1 \to \mathbb{R}^n$, then the first part is a diagram of $\mathbb{R}^1$, the second part a diagram of $\mathbb{R}^n$, and between them are arrows to indicate the image of each point in $\mathbb{R}^1$. The picture below the first example was of this type. All of the arrows get in the way, so a more convenient picture is needed. That comes next.

The second picture is the *graph* of an operator L. The *graph* $L\colon \mathbb{R}^1 \to \mathbb{R}^n$ is the set of points (X, LX) in the Cartesian product space $\mathbb{R}^1 \times \mathbb{R}^n$. Thus, if V^1 is time, and $\mathbb{R}^n$ space with L assigning a position to every time, then the points on the graph are points in time - space (X, LX). For the previous example, these are the points $(t, (\frac{t}{2}, -t))$ in $\mathbb{R}^1 \times \mathbb{R}^2$, a straight line in time-space (or space-time if you prefer). To each time, there is a unique point in space. In a sense, this second picture, the graph, associated with an operator results from gluing together the two pieces of the first picture. By using the graph of an operator, we avoid the arrow mess of the first picture.

The third picture just indicates the range of an operator (when thinking of pictures, the range is often referred to as the *path* of the operator). In terms of the time- position example, this picture only shows the *path* of a particle in space and ignores when a particle had a given position. Thus, this picture is the second half of the first picture. From our physical interpretation, it is clear that two *different operators* might have the *same path* (for two particles could travel the same path without having the same position at every time). Thus, this picture is an incomplete representation of an operator.

EXAMPLE: If $\tilde{L}\colon \mathbb{R}^1 \to \mathbb{R}^2$ such that the point $X_1 = 1 \in \mathbb{R}^1$ is mapped into the point $P = (1, -2) \in \mathbb{R}^2$, then

$$\tilde{L}\colon X = s \cdot 1 \to (s, -2s).$$

In particular, the point $X = 3 (s = 3)$ is mapped into the point $(3, -6)$. The *graph* of $\tilde{L}$ is the set of points $(s, s, -2s)$, which is a straight line in $\mathbb{R}^1 \times \mathbb{R}^2$. Compare this with the operator L considered previously (we remind you that $L\colon X = 2s \to (s, -2s)$). The graph of L was the set of point $(2s, s, -2s)$. These two sets of points the graphs of $\tilde{L}$ and L, respectively, do not coincide since the operators are the same. On the other hand, the *path* of $\tilde{L}$ is the set of points $(s, -2s)$, which is exactly the same set of points as the path of L. Shortly, we shall ask the question, how can we describe a straight line in $\mathbb{R}^n$? One way is to find an operator whose path is that straight line. Since many operators have the same path, there will be many possible ways to describe the straight line. All we need do is pick one, any one will do.

Let $L\colon \mathbb{R}^1 \to \mathbb{R}^n$ and Y_0 be some fixed point in $\mathbb{R}^n$. Consider the operator $MX := LX + Y_0$. Since $M0 = L0 + Y_0 = Y_0 \neq 0$, we see that M is not a linear operator; it is called an *affine operator* or *affine mapping*. The range of M is the subspace translated by the vector Y_0, a straight line which does not necessarily pass through the origin (it will if and only if $Y_0 \in \mathcal{R}(L)$). In other words, $\mathcal{R}(M)$ is the Y_0 coset of the subspace $\mathcal{R}(L)$.

EXAMPLE: Take L to be the same as before, so $L\colon X = 2s \to (s, -2s)$ or $L(2s) = (s, -2s)$. Let $Y_0 = (-3, 2)$. Then $MX := LX + Y_0 = (s, -2s) + (-3, 2) = (s - 3, -2s + 2)$, where $X = 2s$. In particular, M maps the point $X = 3 \in V^1 (s = \frac{e}{2})$ into $(-\frac{3}{2}, -1) \in \mathbb{R}^2$. The figure shows the path of L and M. Since $X = 2s$, we can eliminate s from the above formula and write

$$MX = (\frac{1}{2}X - 3, -X + 2), \qquad X \in \mathbb{R}^1.$$

If we denote by $Y = (y_1, y_2)$ a general point in $\mathbb{R}^2$, then M may be written in the standard form

$$y - 1 = \frac{1}{2}X - 3$$
$$y - 2 = -X + 2.$$

Of course, one could eliminate X from these too and be left with $2y_1 + y_2 = -4$, which is the equation of the path and could come from any mapping with the same path.

It is instructive to investigate the reverse question, given two points P and Q in $\mathbb{R}^n$, find an equation for the straight line passing through them. Any mapping whose path is the desired line will do. We have learned that $MX = LX + Y_0$ is the general equation

of a straight line through Y_0. There is complete freedom in specifying which points are mapped into P and Q, so we would be foolish not to pick the most simple case. Let $M\colon 0 \to P$ and $M\colon 1 \to Q$. Then $P = M(0) = L(0) + Y_0 = Y_0$, so $Y_0 = P$, and $Q = M(1) = L(1) + Y_0 = L(1) + P$, so $L\colon 1 \to P - Q$. This completely determines M (since L is determined once the image of one point is known, $L\colon 1 \to P - Q$, and the vector Y_0 is also determined, $Y_0 = P$).

EXAMPLE: Find an equation for the straight line passing through the two points $P = (1, 2, -3, -4)$, $Q = (-1, 3, 2, -2)$ in $\mathbb{R}^4$. Say P is the image of 0 and Q is the image of 1, so $M\colon 0 \to P$ and $M\colon 1 \to Q$. Then since $MX = LX + Y_0 \Rightarrow P = M(0) = Y_0$ so $Y_0 = (1, 2, -3, -4)$. Also $Q = L(1) + Y_0 \Rightarrow L(1) = Q - Y_0 = Q - P = (-2, 1, 5, 2)$. Because every $X \in \mathbb{R}^1$ can be written as $X = s \cdot 1 \Rightarrow LX = L(s \cdot 1) = sL(1) = s(-2, 1, 5, 2)$, or $LX = (-2s, s, 5s, 2s)$, where $X = s \cdot 1 \in \mathbb{R}^1$. Thus $MX = LX + Y_0 = (-2s, s, 5s, 2s) + (1, 2, -3, -4)$, or

$$MX = (-2s + 1, s + 2, 5s - 3, 2s - 4), \text{ where } X = s \cdot \in \mathbb{R}^1.$$

If we use $Y = (y_1, y_2, y_3, y_4)$ to indicate a general point in $\mathbb{R}^4$, then $M\colon \mathbb{R}^1 \to \mathbb{R}^4$ can be written as four equations.

$$y_1 = -2s + 1$$
$$y_2 = s + 2$$
$$y_3 = 5s - 3$$
$$y_4 = 2s - 4$$

where $X = s \cdot 1 \in \mathbb{R}^1$. For example, the image of $X = 2 (s = 2)$ in $\mathbb{R}^1$ is the point $Y = (-3, 4, 7, 4) \in \mathbb{R}^4$.

The discussion before the example contained most of the proof of *Theorem 13.* Let P and Q be two points in $\mathbb{R}^n$. Then the affine mapping

$$MX = P + s(Q - P),$$

has as its path the straight line passing through P and Q.

REMARK: 1 The affine mapping $\tilde{M}X = P + ks(Q - P)$, where $k \neq 0$ is some constant, has the same path too. The only change is that while $M\colon 0 \to P$ and $M\colon 1 \to Q$ this mapping $\tilde{M}\colon 0 \to P$ and $\tilde{M}\colon ks \to Q$. In other words for $\tilde{M}$ we have chosen to take ks (not s) as the pre-image of Q. This pre-image of Q was entirely arbitrary anyway.

REMARK: 2 The equation $MX = P + s(Q - P)$ of $M\colon \mathbb{R}^1 \to \mathbb{R}^n$, where $X = s \cdot 1 \in \mathbb{R}^1$ is called a *parametric equation* of the straight line which passes through P and Q in $\mathbb{R}^n$, and s is called the *parameter*. Other parametric equations of the *same* line arise if $X = ks \cdot 1 \in \mathbb{R}^1$ (cf. Remark 1), where k is some non-zero constant.

In order to introduce the slope of a straight line, let us paraphrase the last few paragraphs in terms of particle motion. If P and Q are two points in $\mathbb{R}^n$, then $Mt = P + t(Q - P)$

$M\colon \mathbb{R}^1 \to \mathbb{R}^n$, where $t \in \mathbb{R}^1$ describes the position of the particle at time t. At $t = 0$ the particle is at P, while at $t = 1$ the particle is at Q. Another particle moving k times as fast has the position $\tilde{M}t = P + kt(Q - P)$. This other particle is also at P when $t = 0$, but takes time $t = \frac{1}{k}$ to reach the point Q. It still has the same path as the first particle. If we denote by $Y = (y_1, y_2, \ldots, y_n)$ an arbitrary point in $\mathbb{R}^n$, then the position Y at time t is

$$\begin{aligned} y_1 &= p_1 + kt(q_1 - p_1) \\ y_2 &= p_2 + kt(q_2 - p_2) \qquad \vdots \\ y_n &= p_n + kt(q_n - p_n). \end{aligned}$$

Now consider the mapping $Mt = P + kt(Q - P)$. The *derivative* at $t = t_1$ is

$$\frac{dM}{dt}\Big|_{t=t_1} = \lim_{t_2 \to t_1} \frac{M(t_2) - M(t_1)}{t_2 - t_1}$$

It represents the velocity at $t = t_1$. To have this make sense, we must introduce a norm in $\mathbb{R}^n$ so that the limit can be defined. Use the Euclidean norm (although any other one could be used, for it turns out that there is no need for a limit in the case of a straight line). Since $M(t_2) - M(t_1) = P + kt_2(Q - P) - [P + kt_1(Q - P)] = k(t_2 - t_1)(Q - P)$, we have

$$\frac{M(t_2) - M(t_1)}{t_2 - t_1} = k(Q - P),$$

so

$$\frac{dM}{dt}(t) = k(Q - P).$$

Because this is independent of t, it is the derivative at *any* time t. Thus, the derivative is a vector, $k(Q - P)$. The derivative represents the *velocity* of a particle moving on the line. The *speed* is the length of the velocity vector, speed $= \|k(Q - P)\|$. What is the slope of the line? Since the line is the path of a mapping, it should not depend on which mapping is used. In terms of mechanics, the slope should not depend on the speed of the particle moving along the line, but only that it moved along the straight line, that is its velocity vector was along the line. Thus we define the *slope* as a unit vector in the direction of the velocity. In our case, slope $= Q - P/\|Q - P\|$. This is a unit vector from P to Q and only depends upon the mapping to specify a positive direction (orientation) for the line.

EXAMPLE: A particle moves on a straight line from $P = (1, -2, 1)$ at $t = 0$ to $Q = (3, 1, -5)$ at $t = 2$. Find the position of the particle as a function of time, the velocity and speed of the particle, and slope of the path.

The equation of the path is $Mt = P + kt(Q - P)$, where k is determined from $Q = M(2) = P + 2k(Q - P)$, so $k = \frac{1}{2}$. Thus $M(t) = (1, -2, 1) + \frac{1}{2}t(2, 3, -6) = (1 + t, 2 + \frac{3}{2}t, 1 - 3t)$. Velocity $= \frac{1}{2}(Q - P) = (1, \frac{3}{2}, -3)$. Speed $= \|\text{velocity}\| = \frac{7}{2}$. Slope $=$ velocity /speed $= (\frac{2}{7}, \frac{3}{7}, -\frac{6}{7})$.

A glance at the formulas which precede the example reveals that the position of a particle which moves along a straight line through P can be written in any of the forms

$$1. \quad M(t) = P + kt(Q - P).$$

where Q is another point on the path and the particle is at Q when $t = \frac{1}{k}$,

$$2. \quad M(t) = P + \frac{dM}{dt}t$$

or

$$3. \quad M(t) = P + Vt,$$

where V is the velocity. See Exercise 5 too.

Exercises

(1) (a) If $L\colon \mathbb{R}^1 \to \mathbb{R}^2$ such that the point $X_1 = 3 \in \mathbb{R}^1$ is mapped into $P = (1, 0)$, which of the following points are in $\mathcal{R}(L)$ i) $(2, 0)$, ii) $(1, 2)$, iii) $(-1, 0)$?
  (b) Sketch two pictures, one of the graph of L, the other of the path of L.
  (c) Find another operator $\tilde{L}\colon \mathbb{R}^1 \to \mathbb{R}^2$ whose path is the same as that for L.

(2) Find a mapping whose path is the straight line passing through the points $(2, -1, 3)$ and $(1, -3, -5)$. Find its slope too.

(3) If a point is at $(1, -1, 0)$ at $t = 0$ and at $(2, 3, 8)$ at $t = 3$, find the position as a function of time if the particle moves along a straight line. What is the velocity and speed of the particle?

(4) If a particle is initially at $(0, 1, 0, 1)$ and has constant velocity $(1, -2, 3, -1)$, find its position as a function of time. Where is it at $t = 3$?

(5) A particle moves along a straight line in such a way that at $t = t_0$ it is at $\tilde{P}$, while at $t = t_1$ it is at $\tilde{Q}$.
  (a) Show that its position $M(t)$ as a function of time is
$$M(t) = \tilde{P} + (t - t_0)\frac{\tilde{Q} - \tilde{P}}{t_1 - t_0}$$
  (b) What is the velocity?
  (c) Show that
$$M(t) = M(t_0) + \frac{dM}{dt}(t - t_0).$$

(6) Two straight lines are *parallel* if they have the same slope. If $M(t) = P + t(Q - P)$ is a parametric equation of one line, find an equation for the parallel line which passes through the point $\tilde{P}$.

4.5 $L\colon \mathbb{R}^n \to \mathbb{R}^1$. Hyperplanes.

Whereas in the previous section we examined linear mappings from a one-dimensional linear space into an n dimensional space, now we shall look at the opposite extreme, linear mappings from an n dimensional space into a one- dimensional space.

Let $L\colon \mathbb{R}^n \to \mathbb{R}^1$. We would like to find a representation theorem for this linear operator. The most natural way to do this is to work with a basis $\{e_1, \ldots, e_n\}$ for $\mathbb{R}^n$. Then every $X \in \mathbb{R}^n$ can be written as $X = \sum_1^n x_k e_k$. Consequently,

$$LX = L\left(\sum_1^n x_k e_k\right) = \sum_1^n L(x_k e_k) = \sum_1^n x_k L(e_k).$$

It is clear that LX is determined once we know all the numbers Le_k. In other words, the linear mapping L is determined by the effect of the mapping on a basis for the domain of the operator. This proves

Theorem 4.14 . *Let $L\colon \mathbb{R}^n \to \mathbb{R}^1$ linearly. If $\{e_k\}$ is a basis for the domain of L, $\mathbb{R}^n$, then*

$$LX = a_1x_1 + a_2x_2 + \ldots + a_nx_n = \sum_{k=a}^n a_k x_k,$$

where $X = \sum_1^n x_k e_k$ and $a_k = Le_k$. Notice that the a_k are scalars since they are in the range of L—and the range of L is $\mathbb{R}^1$ by hypothesis.

Examples:

(1) Consider the linear operator $L\colon \mathbb{R}^3 \to \mathbb{R}^1$, which maps $L\colon e_1 = (1,0,0) \to 1$, $L\colon e_2 = (0,1,0) \to 0$, and $L\colon e_3 = (0,0,1) \to 0$. Since the e_k constitute a basis for $\mathbb{R}^3$, the mapping L is completely determined by using Theorem 14. If $X = (x_1, x_2, x_3) \in \mathbb{R}^3$, then $X = x_1e_1 + x_2e_2 + x_3e_3$. Thus

$$LX = x_1Le_1 + x_2Le_2 + x_3Le_3 = x_1 - x_2$$

or

$$LX = x_1.$$

For example, $L\colon (2,1,7) \to 2$. The nullspace of L—those points $X \in \mathbb{R}^3$ such that $LX = 0$—are the points $X = (x_1, x_2, x_3) \in \mathbb{R}^3$ such that $x_1 = 0$ which is the x_2x_3 plane.

(2) Let $L\colon \mathbb{R}^4 \to \mathbb{R}^1$ such that $Le_1 = 1$, $Le_2 = -2$, $Le_3 = 5$, $Le_4 = -3$, where $e_1 = (1,0,0,0)$, $e_2 =$ etc. Then if $X = (x_1, x_2, x_3, x_4) \in \mathbb{R}^4$, we have

$$LX = x_1 - 2x_2 + 5x_3 - 3x_4.$$

The nullspace of L is again a hyperplane, the hyperplane $x_1 - 2x_2 + 5x_3 - 3x_4 = 0$ in $\mathbb{R}^4$.

So far we have not given any attention to the range of L, all of our pictures being in the domain of L. Since the range is $\mathbb{R}^1$, its picture is a simple straight line which is not very interesting. However the graph of L is interesting. Let $L\colon \mathbb{R}^n \to \mathbb{R}^1$ and $Y = (y) \in \mathbb{R}^1$. Then

$$y = a_1x_1 + \ldots + a_nx_n.$$

The graph of L is the set of points (X, LX) [or (X, Y) where $Y = LX$] in $\mathbb{R}^n \times \mathbb{R}^1 \cong \mathbb{R}^{n+1}$. A point $(X, Y) = (x_1, \ldots, x_n, y) \in \mathbb{R}^n \times \mathbb{R}^1$ is on the graph if the coordinates satisfy the equation $y = a_1x_1 + \ldots + a_nx_n$. This equation can be written as $0 = a_1x_1 + \ldots + a_nx_n + (-1)y$ which is a hyperplane in $\mathbb{R}^{n+1}$.

Thus we have found two ways to associate a hyperplane with $L\colon \mathbb{R}^n \to \mathbb{R}^1$,

i) All X such that $LX = 0$, which is the nullspace of L, a linear space of dimension $n-1$ (since $\dim \mathcal{N}(L) = \dim \mathcal{D}(L) - \dim \mathcal{R}(L) = n-1$).

ii) The graph of L, that is, all points of the form (X, LX), is a linear space of dimension $n+1$.

Although this is confusing, both ways are used in practice, whichever is most convenient for the problem at hand. For the remainder of this section, we shall confine our attention to hyperplanes defined in the first way.

Since linear mappings $L\colon \mathbb{R}^n \to \mathbb{R}^1$ all have the form $LX = a_1x_1 + \ldots + a_nx_n$, and since it is natural to think of the sum as the scalar product of the vectors $N = (a_1, \ldots, a_n)$ and $X = (x_1, \ldots, x_n)$. Theorem 14 may be rephrased as

Theorem 4.15 . *If $L\colon \mathbb{R}^n \to \mathbb{R}^1$, then $LX = \langle N, X\rangle$, where N is the vector $N = (Le_1, \ldots, Le_n)$ and $\{e_k\}$ form a basis for $\mathbb{R}^n$.*

Remark: The vector N is an element of the so-called *dual space* of $\mathbb{R}^n$. From the above, it is clear that the dual space of $\mathbb{R}^n$ also has dimension n.

Theorem 14' is a "representation theorem". It states that every linear mapping $L\colon \mathbb{R}^n \to \mathbb{R}^1$ may be represented in the form $LX := \langle N, X\rangle$ for some vector N which depends on L. You may wish to think of N as a vector perpendicular to the hyperplane $LX = 0$ (cf. Ex. 8, p. 225).

Example: Consider the operator L of Example 2 in this section. For it, $LX = \langle N, X\rangle$ where N is the particular vector $N = (1, -2, 5, 3)$.

Recall that a linear functional is a linear operator l whose range is $\mathbb{R}^1$. Since the operators $L\colon \mathbb{R}^n \to \mathbb{R}^1$ we are considering have range $\mathbb{R}^1$, they are all linear functionals. We may again rephrase Theorem 14 in this language. It states that every linear functional defined on $\mathbb{R}^n$ may be represented in the form $l(X) = \langle N, X\rangle$, where N depends on the functional l at hand. This is just a restatement of Theorem 14 with the realization that our L's are linear functionals. Don't let the excess language bewilder you.

So far in this section, we have concentrated our attention on the *algebraic* representation of a linear operator (functional) $L\colon \mathbb{R}^n \to \mathbb{R}^1$. Let us turn to geometry for a bit. In passing we observed that the nullspace of the operator was a hyperplane in the domain of L (a hyperplane in a linear space V is a "flat" subset of V whose dimension is one less than V, that is, of codimension one). These hyperplanes, $\{X \in \mathbb{R}^n : LX = 0\}$, all passed through the origin of $\mathbb{R}^n$. A plane parallel to this one which passes through the particular point $X^0 \in \mathbb{R}^n$ has the form

$$L(X - X^0) = 0.$$

It is clear that the point $X = X^0$ does satisfy the equation. From the representation theorem,

$$L(X - X^0) = a_1(x_1 - x_1^0) + a_2(x_2 - x_2^0) + \ldots + a_n(x_n - x_n^0) = 0,$$

is the equation of this hyperplane, where $X = (x_1, x_2, \ldots, x_n)$ and $X^0 = (x_1^0, x_2^0, \ldots, x_n^0)$. If we again write $N = (a_1, a_2, \ldots, a_n)$, then the equation of the hyperplane is

$$\langle N,\, X - X^0 \rangle = 0,$$

all vectors X such that $X - X_0$ is perpendicular to N.

EXAMPLES:

(1) Find the equation of a plane which passes through the point $X^0 = (1, 2, -5)$ and is parallel to the plane $-2x_1 + 7x_2 + 4x_3 = 0$.

Solution: Here $N = (-2, 7, 4)$, $X = (x_1, x_2, x_3)$, so the plane has the equation

$$0 = \langle N,\, X - X^0 \rangle = -2(x_1 - 1) + 7(x_2 - 2) + 4(x_3 + 5),$$

which may be written as

$$-2x_1 + 7x_2 + 4x_3 = -8.$$

The equation has been cooked up so that $X^0 = (1, 2, -5)$ does satisfy it.

(2) Find the equation of a plane which passes through the point $X^0 = (1, 2, -5)$ and is parallel to the plane $-2x_1 + 7x_2 + 4x_3 = 37$.

Solution: Since this plane is also parallel to the plane $-2x_1 + 7x_2 + 4x_3 = 0$, the solution is that of Example 1.

(3) Find the equation of the plane in $\mathbb{R}^4$ which is perpendicular to the vector $N = (1, -2, 3, 1)$ and passes through the point $X^0 = (1, 0, 1, -1)$. Easy. The plane is all points X such that

$$\langle N,\, X - X^0 \rangle = 0,$$

that is

$$(x_1 - 1) - 2(x_2 - 0) + 3(x_3 - 1) + (x_4 + 1) = 0,$$

or

$$x_1 - 2x_2 + 3x_3 + x_4 = 3.$$

(4) Find the equation of the plane in $\mathbb{R}^3$ which passes through the three points

$$X^1 = (7, 0, 0), \quad X^2 = (1, 0, -2), \quad X^3 = (0, 5, 1).$$

We shall find this by using the general equation of a plane,

$$a_1(x_1 - x_1^0) + a_2(x_2 - x_2^0) + a_3(x_3 - x_3^0) = 0.$$

Here $X^0 = (x_1^0, x_2^0, x_3^0)$ is a particular point on the plane. We may use any of X^1, X^2, or X^3 for it. Since X^1 is simplest, we take $X^0 = (7, 0, 0)$. All that remains is to find the coefficients a_1, a_2, and a_3 in

$$a_1(x_1 - 7) + a_2x_2 + a + 3x_3 = 0.$$

Since X^2 and X^3 are in the plane (and so must satisfy its equation), the substitution $X = X^2$ and $X = X^3$ yields two equations for the coefficients,

$$a_1(1 - 7) + a_2 0 + a_3(-2) = 0$$
$$a_1(0 - 7) + a_2(5) + a_3(1) = 0.$$

These two equations in three unknowns may be solved for any two in terms of the third. We find $a_3 = -3a_1$ and $a_2 = 2a_1$, so the equation is

$$a_1(x_1 - 7) + 2a_1x_2 - 3a_1x_3 = 0.$$

Factoring out the coefficient a_1, we obtain the desired equation

$$x_1 - 7 + 2x_2 - 3x_3 = 0.$$

(It is clear from the general equation of a plane that the coefficients are determined only to within a constant multiple).

Exercises

(1) Let $L\colon \mathbb{R}^2 \to \mathbb{R}^1$ map

$$L\colon (1, 0) \to 3, \quad L\colon (0, 1) \to -2.$$

Write LX in the form $Lx = a_1x_1 + a_2x_2$. $L\colon (7, 3) \to ?$

(2) Let $L\colon \mathbb{R}^2 \to \mathbb{R}^1$ map

$$L\colon (2, 1) \to 1, \quad L\colon (0, 3) \to -2.$$

Write LX in the form $LX = a_1x_1 + a_2x_2$. $L\colon (7, 3) \to\ ?$

(3) Find the equation of a plane in $\mathbb{R}^3$ which passes through the point $(3, -1, 2)$ and is parallel to the plane $x_1 - x_2 - 2x_3 = 7$.

(4) Find the equation of a plane in $\mathbb{R}^5$ which is perpendicular to the vector $N = (6, 2, -3, 1, -1)$ and contains the point $(1, 1, 1, 1, 4)$.

(5) Find the equation of a plane in $\mathbb{R}^4$ which contains the four points $X_1 = (2, 0, 0, 0)$, $X_2 = (1, 0, 2, 0)$, $X_3 = (0, -1, 0, -1)$, $X_4 = (3, 0, 1, 1)$.

(6) In this problem, you will have to use the norm induced by the scalar product.

a). Show that the distance between the point $Y \in \mathbb{R}^n$ and the plane $A = \{ X \in \mathbb{R}^n : \langle N,\, X - X^0 \rangle = 0 \}$ is

$$d(Y, A) = \frac{\left|\langle N,\, Y - X^0 \rangle\right|}{\|N\|}.$$

b). Prove that the distance between the parallel planes $A = \{ X \in \mathbb{R}^n : \langle N,\, X - X^1 \rangle = 0 \}$ and $B = \{ X \in \mathbb{R}^n : \langle N,\, X - X^2 \rangle = 0 \}$ is

$$d(A, B) = \frac{\left|\langle N,\, X^2 - X^1 \rangle\right|}{\|N\|}.$$

Chapter 5

Matrices and the Matrix Representation of a Linear Operator

5.1 $L\colon \mathbb{R}^m \to \mathbb{R}^n$.

The simplest example of a linear operator L which maps $\mathbb{R}^m$ into $\mathbb{R}^n$ is supplied by n linear algebraic equations with m variables. Let $X = (x_1, \ldots, x_m) \in \mathbb{R}^m$. Then we define

$$LX = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + \quad \cdots\cdots \quad + a_{nm}x_m \end{pmatrix} \tag{5-1}$$

Notice the right side of this equation is a (column) vector with n components. If we let $Y = (y_1, \ldots, y_n)$, then the equation $LX = Y$ or

$$\sum_{j=1}^{m} a_{ij}x_j = y_i, \qquad i = 1, 2, \ldots, n,$$

determines a vector Y in $\mathbb{R}^n$ for every X in $\mathbb{R}^m$. Since the operator L is essentially specified by the coefficients $a_{11}, a_{12}, \ldots, a_{nm}$, it is convenient to represent it by the notation

$$L = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{lm} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{nl} & a_{n2} & \cdots & a_{nm} \end{pmatrix},$$

and use the notation

$$LX = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{lm} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{nl} & \cdots & \cdots & a_{nm} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ x_m \end{pmatrix} \tag{5-2}$$

The ordered array of $m \times n$ coefficients is called a *matrix* associated with L, and the numbers a_{ij} are called the *elements* of the matrix. The first index i refers to the *row* while the second index j refers to the *column*. We may also write $L = ((a_{ij}))$ as a shorthand to refer to the whole matrix. Since we shall only use *linear* operators in this chapter it is convenient to drop the letter L for the operator and use $A = ((a_{ij}))$ instead. This will facilitate the notation when referring to other matrices $B = ((b_{in})$, etc. since there will be enough subscripts without adding to the confusion by using L_1, L_2, etc. for linear operators.

In this section we shall work out the meaning of operator algebra applied to the special case of operators $L\colon \mathbb{R}^m \to \mathbb{R}^n$ which are represented by matrices. It turns out that every operator $L\colon \mathbb{R}^m \to \mathbb{R}^n$ can be represented by a matrix (proved later in this very section).

Let us first i) define equality, ii) exhibit the matrices for the zero operator $O(X) = 0$ (additive identity). If $A = ((a_{ij}))$ and $B = ((b_{ij}))$ both map $\mathbb{R}^m \to \mathbb{R}^n$, then by definition, $A = B$ if and only if $AX = BX$ for every $X \in \mathbb{R}^m$, that is, for all $X = (x_1, x_2, \ldots, x_m)$,

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{im}x_m = b_{il}x_l + \cdots + b_{im}x_m, \quad i = 1, 2, \ldots, n$$

or

$$\sum_{j=1}^{m} a_{ij}x_j = \sum_{j=1}^{m} b_{ij}x_j, \qquad i = 1, 2, \ldots, n.$$

Subtracting, we find that

$$\sum_{j=1}^{m} (a_{ij} - b_{ij})x_j = 0, \qquad i = 1, 2, \ldots, n$$

must hold for any choice of $X = (x_1, x_2, \ldots, x_m)$. From the particular choice $X = (1, 0, 0, \ldots 0)$, we see that

$$a_{i1} - b_{i1} = 0, \qquad i = 1, 2, \ldots, n,$$

that is,

$$a_{11} = b_{11}, a_{21} = b_{21}, \ldots, a_{nl} = b_{nl}.$$

Similarly, by using other vectors X, we conclude

Theorem 5.1 *1 (*EQUALITY*). If $A = ((a_{ij}))$ and $B = ((b_{ij}))$ both map $\mathbb{R}^m \to \mathbb{R}^n$, then $A = B$ if and only if the corresponding elements of their matrices are equal,*

$$a_{ij} = b_{ij}, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

It is clear that the $n \times m$ matrix all of whose elements are zero

$$0 = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

has the property that it maps every $X \in \mathbb{R}^m$ into zero, and thus satisfies the conditions for the zero matrix. That this is the only such matrix follows from Theorem 1, since any other matrix which acts the same way on every vector $X \in \mathbb{R}^m$ must have the same elements - all zeroes.

Theorem 5.2 *2. The* ZERO MATRIX *$0\colon \mathbb{R}^m \to \mathbb{R}^n$ is uniquely represented by a matrix with n rows and m columns, all of whose elements are zero.*

How is the identity matrix I defined? Since $I\colon \mathbb{R}^n \to \mathbb{R}^n$ maps every vector into itself, $IX = X$, the linear equations (1) must have the property that given any vector $X = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, then

$$\sum_{j=1}^{n} \delta_{ij}x_j = x_i, \qquad i = 1, 2, \ldots, n$$

If $a_{ij} = \delta_{ij}$ (the Kronecker delta), so $a_{11} = a_{22} = \cdots = a_{nn} = 1$ while $a_{ij} = 0, \quad i \neq j$, then indeed

$$\sum_{j=1}^{n} \delta_{ij}x_j = x_i, \qquad i = 1, 2, \ldots n$$

is satisfied. Thus, the coefficients of the identity matrix are $I = ((\delta_{ij}))$. This is a square $(n^x m)$ matrix,

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

with ones along the main *diagonal* and zeroes elsewhere.

Theorem 5.3 *3. The* IDENTITY MATRIX *$I\colon \mathbb{R}^n \to \mathbb{R}^n$ is uniquely represented by a square $(n \times n)$ matrix whose elements are $I = ((\delta_{ij}))$.*

We turn to addition. Let $A = ((a_{ij}))$ and $B = ((b_{ij}))$ be two $n \times m$ matrices, so they both represent operators mapping $\mathbb{R}^m$ into $\mathbb{R}^n$. Their sum $C = A + B$ is defined as the operator which acts upon X according to the rule (p. 268)

$$CX = AX + BX, \quad X \in \mathbb{R}^m.$$

The elements c_{ij} of the matrix C consequently satisfy

$$\sum_{j=1}^{m} c_{ij}x_j = \sum_{j=1}^{m} a_{ij}x_j + \sum_{j=m}^{m} b_{ij}x_j, \quad j = 1, 2, \ldots, n.$$

or

$$= \sum_{j=1}^{m} (a_{ij} + b_{ij})x_j, \qquad j = 1, 2, \ldots, n$$

for all $X = (x_1, x_2, \ldots, x_m)$. Thus, the c_{ij} are in fact $a_{ij} + b_{ij}$ (by Theorem 1)

Theorem 5.4 *4. If $A = ((a_{ij}))$ and $B = ((b_{ij}))$ both map $\mathbb{R}^m$ into $\mathbb{R}^n$, then their sum $C = A + B$ has elements*
$$c_{ij} = a_{ij} + b_{ij}.$$

REMARK: From this it follows that the zero matrix is actually the additive identity, for if $A = ((a_{ij}))$, then $C = A + 0$ has elements $c_{ij} = a_{ij} + 0 = a_{ij}$, that is, $A + 0 = A$.

EXAMPLE: 1 Let A and B which map $\mathbb{R}^3 \to \mathbb{R}^4$ be represented by the matrices

$$A = \begin{pmatrix} -3 & 0 & 1 \\ 7 & 2 & -1 \\ 5 & 4 & -3 \\ 0 & 1 & 1 \end{pmatrix}; \quad B = \begin{pmatrix} 2 & 2 & 2 \\ -3 & 0 & 0 \\ -4 & -2 & 2 \\ 0 & -1 & -1 \end{pmatrix}.$$

Then
$$A + B = \begin{pmatrix} -1 & 2 & 3 \\ 4 & 2 & -1 \\ 1 & 2 & -1 \\ 0 & 0 & 0 \end{pmatrix}.$$

EXAMPLE: 2. Let A and B be the operators on p. 268 (called L_1 and L_2 there) which map $\mathbb{R}^2 \to \mathbb{R}^3$. Then

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 0 & -1 \end{pmatrix}, \qquad B = \begin{pmatrix} -3 & 1 \\ 1 & -1 \\ 1 & 0 \end{pmatrix},$$

so
$$A + B = \begin{pmatrix} -2 & 2 \\ 2 & 1 \\ 1 & -1 \end{pmatrix}$$

which agrees with the sum obtained there.

If $A = ((a_{ij}))$, is there a matrix $\tilde{A}$ such that $A + \tilde{A} = 0$? Clearly the matrix $\tilde{A}$ defined by $\tilde{A} = ((-a_{ij}))$ does the job since

$$A + \tilde{A} = ((a_{ij})) + ((-a_{ij})) = ((0))$$

by definition of addition. We shall denote the matrix with elements $((-a_{ij}))$ by "$-A$" since $A + (-A) = 0$. This matrix "$-A$" is the additive inverse to A.

EXAMPLE: If
$$A = \begin{pmatrix} 1 & -1 \\ -\pi & 2 \\ 0 & -1 \end{pmatrix} \quad \text{then} \quad -A = \begin{pmatrix} -1 & 1 \\ \pi & -2 \\ 0 & 1 \end{pmatrix}.$$

Since a linear operator which is represented by a matrix is still a linear operator, Theorem 3 (p. 269) certainly holds for matrix addition. We shall rewrite it.

Theorem 5.5 *Let $A, B, C, \ldots$ be matrices which map $\mathbb{R}^m$ into $\mathbb{R}^n$ (so they are $n \times m$ matrices). The set of all such matrices forms an abelian (commutative) group under addition, that is,*

1. $A + (B + C) = (A + B) + C$
2. $A + B = B + A$
3. $A + 0 = A$
4. For every A, there is a matrix $(-A)$ such that

$$A + (-A) = 0.$$

PROOF: No need to do this again since it was carried out in even greater generality on p. 270. For practice, you might want to write out the proof in the special case of 2×3 matrices and see how much more awkward the formulas become when you use the specific elements instead of proceeding more abstractly as we did in the proof on p. 270.

If α is a scalar and $A = ((a_{ij}))$ is an $n \times m$ matrix which represents a linear operator mapping $\mathbb{R}^m \to \mathbb{R}^n$, the operator αA is *defined* by the rule

$$(\alpha A)X = A(\alpha X)$$

where X is any vector in $\mathbb{R}^m$. In terms of the elements $((a_{ij}))$, this means that the elements $((\tilde{a}_{ij}))$ of αA are given by

$$\begin{aligned} \sum_{j=1}^{m} \tilde{a}_{ij} x_j &= \sum_{j=1}^{m} a_{ij}(\alpha x_j), \qquad i = 1, 2, \ldots, n \\ &= \sum_{j=1}^{m} (\alpha a_{ij}) x_j, \qquad i = 1, 2, \ldots, n \end{aligned}$$

so $\tilde{a}_{ij} = \alpha a_{ij}$. Thus, the matrix αA is found by multiplying each of the elements of A by α,

$$\alpha \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & \cdots & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{nl} & \cdots & \cdots & a_{nm} \end{pmatrix} = \begin{pmatrix} \alpha a_{11} & \alpha a_{12} & \cdots & \alpha a_{1m} \\ \alpha a_{21} & \alpha a_{22} & \cdots & \alpha a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha a_{nl} & \cdots & \cdots & \alpha a_{nm} \end{pmatrix}$$

EXAMPLE:
$$-1 \begin{pmatrix} 7 & 1 & 3 \\ -2 & -1 & 4 \\ 9 & 6 & 5 \\ -3 & 1 & -1 \end{pmatrix} = \begin{pmatrix} -14 & -2 & -6 \\ 4 & 2 & -8 \\ -18 & -12 & -10 \\ 6 & -2 & 2 \end{pmatrix}.$$

The following theorem concerns multiplication of matrices by scalars. It is proved either by direct computation - or more simply by realizing that it is a special case of Exercise 12, p. 284.

Theorem 5.6 . *If A and B are matrices which map $\mathbb{R}^m \to \mathbb{R}^n$, and if α, β are any scalars, then*

1. $\alpha(\beta A) = (\alpha\beta)A$
2. $1 \cdot A = A$
3. $(\alpha + \beta)A = \alpha A + \beta A$
4. $\alpha(A + B) = \alpha A + \alpha B$.

Remark: Theorems 5 and 6 together state that the *set of all matrices which map $\mathbb{R}^m$ into $\mathbb{R}^n$ forms a linear space.* It is easy to show that the dimension of this space is $m \cdot n$ (by exhibiting $m \cdot n$ linearly independent matrices which span the whole space).

Now we get more algebraic structure and see how to multiply. Let A map $\mathbb{R}^?$ into $\mathbb{R}^n$ and $B = ((b_{ij}))$ map $\mathbb{R}^r$ into $\mathbb{R}^s$. By definition of operator multiplication (p. 271-2), the product AB is defined on an element $X \in \mathbb{R}^r = \mathcal{D}(B)$ by the rule

$$ABX = A(BX).$$

Since the vector $BX \in \mathbb{R}^s$ must be fed into A, we find that $BX \in \mathbb{R}^m$ too. Thus, *in order for the product AB of an $n \times m$ matrix A with a $s \times r$ matrix B to make sense, we must have* $? = m$, that is, the range of B must be contained in the domain of A,

A FIGURE GOES HERE

If $C = ((c_{ij})) = AB$, then for every $X \in \mathbb{R}^r$

$$CX = A(BX)$$

or

$$\sum_{k=1}^{r} c_{ij} x_k = \sum_{j=1}^{s} a_{ij} \left(\sum_{k=1}^{r} b_{jk} x_k \right), \quad i = 1, 2, \ldots, n$$

so

$$= \sum_{k=1}^{r} \left(\sum_{j=1}^{s} a_{ij} b_{jk} \right) x_k, \quad i = 1, 2, \ldots, n.$$

Therefore, the elements c_{ik} of the product AB are given by the formula

$$c_{ik} = \sum_{j=1}^{s} a_{ij} b_{jk} \qquad \begin{array}{l} i = 1, 2, \ldots, n \\ k = 1, 2, \ldots, r \end{array}.$$

Since the summation signs have probably overwhelmed you, we repeat it in a special case. Let B be determined by the linear equations

$$\begin{aligned} b_{11}x_1 + b_{12}x_2 + b_{13}x_3 &= y_1 \\ b_{21}x_1 + b_{22}x_2 + b_{23}x_3 &= y_2. \end{aligned}$$

Then $B\colon \mathbb{R}^3 \to \mathbb{R}^2$. Also let $A\colon \mathbb{R}^2 \to \mathbb{R}^2$ be determined by

$$\begin{aligned} a_{11}y_1 + a_{12}y_2 &= z_1 \\ a_{21}y_1 + a_{22}y_2 &= z_2. \end{aligned}$$

The product AB maps a vector $X \in \mathbb{R}^3$ first into $Y = BX \in \mathbb{R}^2$ and then into $Z = ABX \in \mathbb{R}^2$.

A FIGURE GOES HERE

Ordinary substitution yields $Z = ABX$ as a function of X:

$$\begin{aligned} a_{11}(b_{11}x_1 + b_{12}x_2 + b_{13}x_3) + a_{12}(b_{21}x_1 + b_{22}x_2 + b_{21}x_3) &= z_1 \\ a_{21}(b_{11}x_1 + b_{12}x_2 + b_{13}x_3) + a_{22}(b_{21}x_1 + b_{22}x_2 + b_{23}x_3) &= z_2, \end{aligned}$$

or

$$\begin{aligned} (a_{11}b_{11} + a_{12}b_{21})x_1 + (a_{11}b_{12} + a_{12}b_{22})x_2 + (a_{11}b_{13} + a_{12}b_{23})x_3 &= z_1 \\ (a_{21}b_{11} + a_{22}b_{21})x_1 + (a_{21}b_{12} + a_{22}b_{22})x_2 + (a_{21}b_{13} + a_{22}b_{23})x_3 &= z_2. \end{aligned}$$

If we write this in the matrix form

$$\begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix},$$

we find

$$c_{11} = a_{11}b_{11} + a_{12}b_{21}, \quad c_{12} = a_{11}b_{12} + a_{12}b_{22}$$

etc., just as was dictated by the general formula for the multiplication of matrices.

Theorem 5.7 . *If $A = ((a_{ij}))$ and $B = ((b_{ij}))$ are matrices with $B\colon \mathbb{R}^r \to \mathbb{R}^s$ and $A\colon \mathbb{R}^s \to \mathbb{R}^n$, then the product $C = AB$ is defined and the elements of the product $C = ((c_{ij}))$ are given by the formula*

$$c_{ik} = \sum_{j=a}^{s} a_{ij} b_{jk}, \quad i = 1, 2, \ldots, n; \quad k = 1, 2, \ldots, r.$$

Remark: Since this formula for matrix multiplication is impossible to remember as it stands, it is fortunate that there is an easy way to remember it. We shall work with the example of matrices $A\colon \mathbb{R}^2 \to \mathbb{R}^2$ and $B\colon \mathbb{R}^3 \to \mathbb{R}^2$ discussed earlier. Then

$$AB = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{pmatrix}.$$

To compute the element c_{ik}, we merely observe that

$$c_{ik} = \sum_{j=1}^{2} a_{ij} b_{jk} = a_{i1} b_{1k} + a_{12} b_{2k}$$

c_{ik} is the *scalar product* of the ith row in A with the kth column in B (see fig.). Thus, the element c_{21} in $C = AB$ is the scalar product of the 2nd row of A with the 1st column of B. Do not be embarrassed to use two hands to multiply matrices. Everybody does.

EXAMPLES:

(1) (cf. p. 274 where this was done without matrices). If

$$A = \begin{pmatrix} 2 & -3 \\ -1 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 2 \\ 1 & 1 \end{pmatrix},$$

then

$$AB = \begin{pmatrix} 2 & -3 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} -3 & 1 \\ 1 & -1 \end{pmatrix}$$

and

$$BA = \begin{pmatrix} 0 & 2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 2 & -3 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} -2 & 2 \\ 1 & -2 \end{pmatrix}.$$

Notice that even though AB and BA are both defined, we have $AB \neq BA$—the expected noncommutativity in operator multiplication.

(2) (cf. p. 272 bottom where this was done without matrices). If

$$A = \begin{pmatrix} 1 & -1 \\ 0 & 1 \\ -1 & -2 \end{pmatrix}, \quad B = (1, 2, -1),$$

then

$$BA = (1, 2, -1) \begin{pmatrix} 1 & -1 \\ 0 & 1 \\ -1 & -2 \end{pmatrix} = (2, 3).$$

However the product AB does not make sense.

From the general theory of linear operators (Theorem 4, p. 276) we can conclude

Theorem 5.8 . *Matrix multiplication is associative, that is, if*

$$\mathbb{R}^k \xrightarrow{A} \mathbb{R}^l \xrightarrow{B} \mathbb{R}^m \xrightarrow{C} \mathbb{R}^n,$$

so the products $C(BA)$ and $(CB)A$ are defined, then

$$C(BA) = (CB)A.$$

Thus the parenthesis can be omitted without risking chaos.

REMARK: Returning to linear algebraic equations, you will observe that the matrix notation AX there [eq(2)] can now be viewed as matrix multiplication of the $n \times m$ matrix $A = ((a_{ij}))$ with the $m \times 1$ matrix (column vector) X.

In developing the algebra of matrices - and operators in general - we have been neglecting one important issue, that of an inverse operator. If $L\colon V_1 \to V_2$, can we find an operator $\tilde{L}\colon V_2 \to V_1$ which reverses the effect of L, that is, if $LX = Y$, where $X \in V_1$ and $Y \in V_2$, is there an operator $\tilde{L}$ such that $\tilde{L}Y = X$? If so, then

$$\tilde{L}LX = \tilde{L}Y = X,$$

and we write

$$\tilde{L}L = I.$$

This operator $\tilde{L}$ is the *left* (multiplicative) *inverse of* L. Similarly, an operator $\hat{L}$ such that $L\hat{L} = I$ is the *right* (multiplicative) *inverse of* L. We shall shortly prove that *if* an operator L has both a left inverse L and a right inverse $\hat{L}$, then they are equal, $\hat{L} = \tilde{L}$, so without ambiguity one can write L^{-1} for *the* inverse.

A FIGURE GOES HERE

To begin, we compute the inverse of the matrix

$$A = \begin{pmatrix} 5 & -2 \\ 3 & -1 \end{pmatrix}$$

associated with the system of linear equations

$$\begin{aligned} 5x_1 - 2x_2 &= y_1 \\ 3x_1 - x_2 &= y_2. \end{aligned}$$

These equations specify a mapping from $\mathbb{R}^2$ into $\mathbb{R}^2$. They map a point X into Y. Finding the inverse of A is equivalent to answering the question, if we are given a point Y, can we find the X whence it came?

$$AX = Y, \quad X = A^{-1}Y.$$

Finding the X in terms of Y means solving these two equations, a routine task. The answer is

$$\begin{aligned} x_1 &= -y_1 + 2y_2 \\ x_2 &= -3y_1 + 5y_2. \end{aligned} \quad \text{so} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -1 & 2 \\ -3 & 5 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

Thus,

$$X = A^{-1}Y,$$

where

$$A^{-1} = \begin{pmatrix} -1 & 2 \\ -3 & 5 \end{pmatrix}.$$

The matrix A^{-1} is the matrix inverse to A. It is easy to check that

$$AA^{-1} = \begin{pmatrix} 5 & -2 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} -1 & 2 \\ -3 & 5 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I$$

and

$$A^{-1}A = \begin{pmatrix} -1 & 2 \\ -3 & 5 \end{pmatrix} \begin{pmatrix} 5 & -2 \\ 3 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I.$$

Thus, this matrix A^{-1} is both the right and left inverse of A.

Our second example is of a more geometric nature. We shall consider a matrix R which represents rotation of a vector in $\mathbb{R}^2$ through an angle α.

A FIGURE GOES HERE

R is represented by the matrix (cf. Ex. 13b p. 285)

$$R = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}.$$

It is geometrically clear that in inverse of this operator R is an operator which rotates through an angle $-\alpha$, unwinding the effect of R. Thus, immediately from the formula for R, we find

$$R^{-1} = \begin{pmatrix} \cos(-\alpha) & -\sin(-\alpha) \\ \sin(-\alpha) & \cos(-\alpha) \end{pmatrix} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}.$$

To check that geometry has not deceived us, we should multiply out RR^{-1} and $R^{-1}R$. Do it. You will find $RR^{-1} = R^{-1}R = I$. One could also have found R^{-1} by solving linear algebraic equations as was done in the first example.

The problem of finding the matrix inverse to any *square* matrix,

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

is equivalent to the dull problem of solving n linear algebraic equations in n unknowns

$$\begin{array}{llll} a_{11}x_1+ & \cdots & +a_{1n}x_n & = y_1 \\ a_{21}x_1+ & \cdots & +a_{2n}x_n & = y_2 \\ \vdots & & \vdots & \\ a_{n1}x_1+ & \cdots & +a_{nn}x_n & = y_n \end{array}$$

for X in terms of Y, $X = A^{-1}Y$. For $n = 2$ the computation is not too grotesque, and yields the formulas

$$x_1 = \frac{a_{22}}{\Delta}y_1 - \frac{a_{12}}{\Delta}y_2$$
$$x_2 = \frac{-a_{21}}{\Delta}y_1 + \frac{a_{11}}{\Delta}y_1$$

where $\Delta = a_{11}a_{22} - a_{12}a_{21}$ (= determinant of A, for those who have seen this before). From this formula we read off that the inverse of the 2×2 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{is} \quad A^{-1} = \frac{1}{\Delta}\begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}.$$

As a check, one computes that

$$AA^{-1} = A^{-1}A = I.$$

Thus the 2×2 matrix A has an inverse if and only if $\Delta := a_{11}a_{22} - a_{12}a_{21} \neq 0$.

A FIGURE GOES HERE

Fortunately, one rarely needs the explicit formula for the inverse of a square $n \times n$ matrix other than the reasonable cases $n = 2$ and $n = 3$. The inverse of a matrix has greater *conceptual* use as the inverse of an operator.

Having relegated the computation of the inverse of a matrix to the future, let us see what can be said about the inverse without computation. This will necessarily be a bit more abstract. Since the issues involve solving systems of linear algebraic equations, we shall invoke the theory concerning that which was developed in Chapter 4 Section 3. For this discussion, it is convenient to use the following definition (cf. p. 6).

Definition: An operator $A\colon V_1 \to V_2$ is *invertible* if it has the two properties

i) If $X_1 \neq X_2$ then $AX_1 \neq AX_2$ (injective, 1-1)

ii) To every $Y \in V_2$, there is at least one $X \in V_1$ such that $AX = Y$ (surjective, onto).

Thus, an operator is invertible if and only if it is bijective. An invertible matrix is usually called *non-singular*, while a matrix which is not invertible is called *singular*.

To show that this definition is identical with the previous one, we must show that every invertible linear operator A has a right and left inverse. A more pressing matter though, is

Theorem 5.9 . *If the linear operator $A\colon V_1 \to V_2$ where V_1 and V_2 are finite dimensional, is invertible, then $\dim V_1 = \dim V_2$, so a matrix must necessarily be square for an inverse to exist (but being square is not sufficient, as was seen in the 2×2 case where the additional condition $a_{11}a_{22} - a_{21}a_{12} \neq 0$ we needed). In other words, you haven't got a chance to invert a matrix unless it is square, but being square is not enough.*

Proof: Condition i) states that $\mathcal{N}(A) = 0$, for if $X_1 \neq 0$, then $AX_1 \neq 0$. Therefore

$$\dim \mathcal{R}(A) = \dim \mathcal{D}(A) - \dim \mathcal{N}(A) = \dim V_1 - 0 = \dim V_1.$$

On the other hand, condition ii) states that $V_2 \subset \mathcal{R}(A)$. Since $A\colon V_1 \to V_2$, we know that $\mathcal{R}(A) \subset V_2$. Therefore $\mathcal{R}(A) = V_2$. Coupled with the first part, we have

$$\dim V_1 = \dim \mathcal{R}(A) = \dim V_2.$$

Theorem 5.10 . *Given an operator A which is invertible, there is a linear operator A^{-1} such that $AA^{-1} = A^{-1}A = I$.*

PROOF: If $\tilde{Y} \in V_2$, there is an $\tilde{X} \in V_1$ such that $A\tilde{X} = \tilde{Y}$ (by property ii), and that $\tilde{X}$ is unique (property i). Therefore without ambiguity we can define $A^{-1}\tilde{Y} = \tilde{X}$. A similar process defines the operator A^{-1} for every $Y \in V_2$. From our construction, it is clear (or should be) that
$$AA^{-1} = A^{-1}A = I.$$
All that remains is to show A^{-1} is linear. If $A\tilde{X} = \tilde{Y}$ and $A\hat{X} = \tilde{Y}$, then since A is linear, $A(a\tilde{X}+b\hat{X}) = aA\tilde{X}+bA\hat{X} = a\tilde{Y}+b\hat{Y}$. Thus $A^{-1}(a\tilde{Y}+b\hat{Y}) = a\tilde{X}+b\hat{X} = aA^{-1}\tilde{Y}+bA^{-1}\hat{Y}$.

REMARK: Glancing over this proof, it should be observed that finite dimensionality (or even the concept of dimension) never entered - so the result is true for infinite dimensional spaces. Furthermore, linearity was only used to show that A^{-1} was linear. Thus the theorem (except for the claim that A^{-1} is linear) is true for nonlinear operators as well. Needless to say, this construction of A^{-1} one point at a time is useless as a method for finding A^{-1} (since even in the simplest case $A\colon \mathbb{R}^1 \to \mathbb{R}^1$ it involves an infinite number of points).

This theorem shows that if an operator A is invertible, then there are right and left inverses which are equal $AA^{-1} = A^{-1}A = I$. We can reverse the theorem and prove

Theorem 5.11 . *Given the linear operator $A\colon V_1 \to V_2$, if there are linear operators $\hat{A}$ (right inverse) and $\tilde{A}$ (left inverse) such that*
$$A\hat{A} = \tilde{A}A = I,$$
then A is invertible and $A^{-1} = \hat{A} = \tilde{A}$.

PROOF: Verify condition i: If $AX_1 = AX_2$, then $\tilde{A}AX_1 = \tilde{A}AX_2$. Since $\tilde{A}A = I$, this implies $X_1 = X_2$.

Verify condition ii. If Y is any element in V_2, let $X = \hat{A}Y$. Then $AX = A\hat{A}Y = Y$, so that Y is the image of X under the mapping.

The proof that $A^{-1} = \tilde{A} = \hat{A}$ is delightfully easy. Only the associative property of multiplication is used:
$$\hat{A} = (A^{-1}A)\hat{A} = A^{-1}(A\hat{A}) = A^{-1} = (\tilde{A}A)A^{-1} = \tilde{A}(AA^{-1}) = \tilde{A}.$$

EXAMPLES:

(1) The identity operator I on every linear space is invertible, for it trivially satisfies both criteria. Not only that, but it is its own inverse for $II = I$.

(2) The zero operator is never invertible, for even though $X_1 \neq X_2$, we always have $0(X_1) = 0 = 0(X_2)$.

(3) The 2×2 matrix
$$A = \begin{pmatrix} 1 & 3 \\ -2 & -6 \end{pmatrix}$$
is not invertible since, from the formula
$$AX = \begin{pmatrix} 1 & 3 \\ -2 & -6 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 & + & 3x_2 \\ -2x_1 & - & 6x_2 \end{pmatrix},$$
we see that the vector $(-3, 1) \neq 0$ is mapped into zero by A (whereas criterion i). states that only 0 can be mapped into 0 by an invertible linear operator). Another way to see that A is not invertible is to observe that $\Delta = a_{11}a_{22} - a_{12}a_{21} = 0$. thus violating the explicit condition for 2×2 matrices found earlier.

In this last example, we observed that if a linear operator A is invertible, then by property i) the equation $AX = 0$ has exactly one solution $X = 0$. If $A\colon V_1 \to V_2$ on a *finite* dimensional space, and $\dim V_1 = \dim V_2$ the converse is true also.

Theorem 5.12 *If the linear operator A maps the linear space V_1 into V_2 and $\dim V_1 = \dim V_2 < \infty$, then*
$$A \textit{ is invertible } \iff AX = 0 \textit{ implies } X = 0.$$

PROOF: $\Rightarrow$ A restatement of condition i) in the definition.
$\Leftarrow$ A restatement of lines 7-10 on page 316.

Corollary 5.13 . *A square matrix $A = ((a_{ij}))$ is invertible if and only if its columns*
$$\mathcal{A}_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ \cdot \\ \cdot \\ \cdot \\ a_{n1} \end{pmatrix}, \quad \mathcal{A}_2 = \begin{pmatrix} a_{12} \\ a_{22} \\ \cdot \\ \cdot \\ \cdot \\ a_{n2} \end{pmatrix}, \ldots, \mathcal{A}_n = \begin{pmatrix} a_{1n} \\ a_{2n} \\ \cdot \\ \cdot \\ \cdot \\ a_{nn} \end{pmatrix}$$
are linearly independent vectors.

PROOF: To test for linear independence, we examine
$$x_a\mathcal{A}_1 + x_2\mathcal{A}_2 + \cdots + x_n\mathcal{A}_n = 0,$$
and try to prove that $x_1 = x_2 = \cdots = x_n = 0$. But writing the equation in full, it reads
$$\begin{array}{cccc} a_{11}x_1+ & a_{12}x_2+ & \cdots & +a_{1n}x_n & = 0 \\ a_{21}x_1+ & a_{22}x_2+ & \cdots & +a_{2n}x_n & = 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1}x_1+ & a_{n2}x_2+ & \cdots & +a_{nn}x_n & = 0, \end{array}$$

or

$$AX = 0.$$

By the theorem, A is invertible if and only if the equation $AX = 0$ has only the solution $X = 0$. Thus A is invertible if and only if the only solution of

$$x_1\mathcal{A}_1 + x_2\mathcal{A}_2 + \cdots + x_n\mathcal{A}_n = 0$$

is $x_1 = x_2 = \cdots = x_n = 0$.

We close our discussion of invertible operators with

Theorem 5.14 . *The set of all invertible linear operators which map a space into itself constitutes a (non- commutative) group under multiplication; that is, if $L_1, L_2, \ldots$ are invertible operators which map V into itself then they satisfy*

0. Closed under multiplication (L_1L_2 is an invertible linear operator which maps V into itself).

(1) $L_1(L_2L_3) = (L_1L_2)L_3$ - *Associative*

(2) *There is an identity I such that*

$$IL = LI = L.$$

(3) *For every operator L in the set, there is another operator L^{-1} for which*

$$LL^{-1} = L^{-1}L = I.$$

PROOF: 0) L_1L_2 is a linear operator which maps V into itself by part 0. of Theorem 4 (p. 276). It is invertible since its inverse can be written in the explicit form (an *important* formula)

$$(L_1L_2)^{-1} = L_2^{-1}LL_1^{-1},$$

as we will verify:

$$(L_1L_2)(L_2^{-1}L_1^{-1}) = L_1(L_2L_2^{-1})L_1^{-1} = L_1IL_1^{-1} = L_1L_1^{-1} = I$$

connect these?? $(L_2^{-1}L_1^{-1})(L_1L_2) = L_2^{-1}(L_1^{-1}L_1)L_2 = L_2^{-1}IL_2 = L_2^{-1}IL_2 = L_2^{-1}L_2 = I.$

(1) Part 1 of Theorem 4 (p. 276).

(2) Part 1 of Theorem 5 (p. 277)

(3) A direct restatement of the fact that our set consists only of invertible operators.

Closely associated with a matrix $A\colon \mathbb{R}^n \to \mathbb{R}^m$

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & \cdots & \cdots & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

is another matrix A^*, the *transpose* or *adjoint* of A, which is obtained by interchanging the rows and columns of A, viz.

$$A^* = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ a_{1n} & \cdots & \cdots & a_{mn} \end{pmatrix}.$$

For example,

$$\text{if } A = \begin{pmatrix} 1 & 2 \\ 4 & -2 \\ 5 & -2 \end{pmatrix}, \text{ then } A^* = \begin{pmatrix} 1 & 4 & 5 \\ 2 & -2 & -1 \end{pmatrix}.$$

If $A = ((a_{ij}))$, then $A^* = ((a_{ij}))$. The adjoint of an $m \times n$ matrix is an $n \times m$ matrix. Thus, if $A\colon \mathbb{R}^n \to \mathbb{R}^m$ then $A^*\colon \mathbb{R}^m \to \mathbb{R}^n$, and for any $Z \in \mathbb{R}^m$, we have

$$A^*Z = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ a_{1n} & \cdots & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} z_1 \\ \cdot \\ \cdot \\ \cdot \\ z_m \end{pmatrix} = \begin{pmatrix} a_{11}z_1 & + & a_{21}z_2 & + & \cdots & + & a_{m1}z_m \\ a_{12}z_1 & + & \cdots & & \cdots & + & a_{m2}z_m \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ a_{1n}z_1 & + & \cdots & & \cdots & + & a_{mn}z_m \end{pmatrix},$$

so the jth component $(A^*Z)j$ of the vector $A^*Z \in \mathbb{R}^n$ is

$$(A^*Z)j = \sum_{i=1}^{m} a_{ij}z_i = a_{1j}z_1 + a_{2j}z_2 + \cdots + a_{mj}z_m.$$

Beware: The classical literature on matrices uses the term "adjoint of a matrix" for an entirely different object. Our nomenclature is now standard in the theory of linear operators.

A real square matrix A is called *symmetric* or *self-adjoint* if $A = A^*$. For example,

$$A = \begin{pmatrix} 7 & 2 & -3 \\ 2 & -1 & 5 \\ -3 & 5 & 4 \end{pmatrix} = A^*.$$

For a symmetric matrix A, we have $a_{ij} = a_{ji}$.

The significance of the adjoint of a matrix (as well as its relation to the more general conception of the adjoint of an arbitrary operator) arises in the following way. If $A\colon \mathbb{R}^n \to \mathbb{R}^m$, then for any X in $\mathbb{R}^n$ the vector $Y = AX$ is a vector in $\mathbb{R}^m$. We can form the scalar product of this vector $Y = AX$ with any other vector Z in $\mathbb{R}^m$ (because Y and Z are both in $\mathbb{R}^m$

$$\langle Z, Y\rangle = \langle Z, AX\rangle.$$

Since $A^*\colon \mathbb{R}^m \to \mathbb{R}^n$, and $Z \in \mathbb{R}^m$, then A^*Z makes sense, and is a vector in $\mathbb{R}^n$, so $\langle A^*Z, X\rangle$ is a real number for any $X \in \mathbb{R}^n$. *Claim*:

$$\langle Z, AX\rangle = \langle A^*Z, X\rangle.$$

This is easy to verify. Let $A = ((a_{ij}))$. Then

$$(AX)_i = \sum_{j=1}^{n} a_{ij}x_j \quad \text{and} \quad (A^*Z)j = \sum_{i=1}^{m} a_{ij}z_i,$$

so that

$$\begin{aligned}\langle Z, AX\rangle &= \sum_{i=1}^{m} z_i(AX)_i = \sum_{i=1}^{m} z_i(\sum_{j=1}^{n} a_{ij}x_j) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} z_i a_{ij}x_j.\end{aligned}$$

In the same way,

$$\begin{aligned}\langle A^*Z, X\rangle &= \sum_{j=1}^{n}(A^*Z)_j x_j = \sum_{j=1}^{n}(\sum_{i=1}^{m} a_{ij}z_i)x_j \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} z_i a_{ij}x_j.\end{aligned}$$

Comparison reveals we have proved

Theorem 5.15 . *If $A\colon \mathbb{R}^n \to \mathbb{R}^m$, then for any $X \in \mathbb{R}^n$ and any $Z \in \mathbb{R}^m$,*

$$\langle Z, AX\rangle = \langle A * Z, X\rangle,$$

where A^* is the adjoint of A.

REMARK: From a more abstract point of view, the operator A^* is usually *defined* as the operator which has the above property. If this definition is adopted, one must use it to prove the adjoint A^* of a matrix A is found by merely interchanging the rows and columns (try to do it!).

It is remarkably easy to obtain some properties of the adjoint by using Theorem 14. Our attention will be restricted to square matrices (although the results are still true with but minor modifications for a rectangular matrix).

Theorem 5.16 . *Let A and B be $n\times n$ matrices (so the products AB, BA, B^*A^*, $A+B$ etc. are all defined). Then*

0. $I^* = I$ (because I is symmetric)
1. $(A^*)^* = A$
2. $(AB)^* = B^*A^*$
3. $(A+B)^* = A^* + B^*$.
4. $(cA)^* = cA^*$, c is a real scalar.
5. A is invertible if and only if A^* is invertible, and

$$(A^*)^{-1} = (A^{-1})^*.$$

6. A is invertible if and only if the rows of A are linearly independent.

PROOF: We could use subscripts and the a_{ij} stuff - but it is clearer to use the result of Theorem 14. In order to do so, an important preliminary result is needed.

Theorem 5.17 . *If $C\colon \mathbb{R}^n \to \mathbb{R}^m$, then the equation*

$$\langle Cx, Y\rangle = 0 \quad \textit{for all} \quad X \quad \textit{in} \quad \mathbb{R}^n \quad \textit{and} \quad Y \quad \textit{in} \quad \mathbb{R}^m$$

$\Longleftrightarrow$ *C is the zero operator, $C = 0$. Thus if C_1 and C_2 map $\mathbb{R}^n$ into itself, the equation*

$$\langle C_1X, Y\rangle = \langle C_2X, Y\rangle \textit{ for all } X, Y \in \mathbb{R}^n \Longleftrightarrow C_1 = C_2.$$

PROOF: $\Rightarrow$ By contradiction, if $C \neq 0$ there is some X_0 such that $0 \neq CX_0 \in \mathbb{R}^n$. Now just pick $Y_0 = CX_0$. Then

$$0 = \langle CX_0, Y_0\rangle = \langle CX_0, CX_0\rangle = \|CX_0\|^2 > 0$$

because by assumption $CX_0 \neq 0$. A glance at this line reveals the desired contradiction.
$\Leftarrow$ Obvious.
The last assertion of the theorem follows by subtraction,

$$0 = \langle C_1X, Y\rangle - \langle C_2X, Y\rangle = \langle C_1X - C_2X, Y\rangle = \langle (C_1 - C_2)X, Y\rangle$$

and letting $C = C_1 - C_2$.

Now we return to the
Proof of Theorem 15: The vectors X, Z will be in $\mathbb{R}^n$.

(0) Particularly clear because I is symmetric. You should try constructing another proof patterned on those below.

(1) Two successive interchanges of the rows and columns of a matrix leave it unchanged. Again, try to construct another proof patterned on those below.

(2) $$\begin{aligned}\langle (AB)^*Z, X\rangle &= \langle Z, ABX\rangle = \langle Z, A(BX)\rangle = \langle A^*Z, BX\rangle \\ &= \langle B^*(A^*Z), X\rangle = \langle (B^*A^*)Z, X\rangle\end{aligned}$$

for all X, Z in $\mathbb{R}^n$. Application of Theorem 16 yields the result.

(3) $\langle (A+B)^* Z,\, X\rangle = \langle Z,\, (A+B)X\rangle = \langle Z,\, AX + BX\rangle$

$$= \langle Z,\, AX\rangle + \langle Z, BX, =\rangle \langle A^* Z,\, X\rangle + \langle B^* Z,\, X\rangle$$

$$= \langle A^* Z + B^* Z,\, X\rangle = \langle (A^* + B^*)Z,\, X\rangle.$$

And apply Theorem 16.

(4) $\langle (cA)^* Z,\, X\rangle = \langle Z,\, cAX\rangle = c\langle Z,\, AX\rangle$

$$= c\langle A^* Z,\, X\rangle = \langle (cA^*)Z,\, X\rangle.$$

Apply Theorem 16.

(5) If A is invertible, then $AA^{-1} = A^{-1}A = I$. An application of parts 0 and 2 shows

$$(A^{-1})^* A* = (AA^{-1})^* = I^* = I.$$

Similarly, $A^*(A^{-1})^* = I$. Thus A^* has a left and right inverse, so it is invertible by Theorem 11. The above formulas reveal $(A^*)^{-1} = (A^{-1})^*$.

In the other direction, assume A^* is invertible. Since $A^{**} = A$ (part 1) the matrix A is the adjoint of A^*. But we just saw that if a matrix is invertible then its adjoint is too. Thus the invertibility of $A*$ implies that of A.

(6) By the Corollary to Theorem 12, A^* is invertible if and only if its columns are linearly independent. Since the columns of A^* are the rows of A, we find that A^* is invertible if and only if the rows of A are linearly independent. Coupled with Part 5, the proof is completed.

In our later work we shall need an inequality. Why not insert it here for future reference.

Theorem 5.18 . *If $A = ((a_{ij}))$ is an $m \times n$ matrix, so $A\colon \mathbb{R}^n \to \mathbb{R}^m$, then for any X in $\mathbb{R}^n$ and Y in $\mathbb{R}^m$*

$$\|AX\| \le k\|X\|$$

and

$$|\langle Y,\, AX\rangle| \le k\|X\| \quad \|Y\|,$$

where

$$k^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2.$$

Proof: By definition

$$\|AX\|^2 = \sum_{i=1}^{m} (AX)_i^2 = \sum_{i=1}^{m} (\sum_{j=1}^{n} a_{ij}x_j)^2,$$

where $(AX)_i$ is the ith component of the vector AX. The Schwarz inequality shows

$$(\sum_{j=1}^{n} a_{ij}x_j)^2 \le \sum_{j=1}^{n} a_{ij}^2 \sum_{j=1}^{n} x_j^2 = \|X\|^2 \sum_{j=1}^{n} a_{ij}^2.$$

Thus,

$$\|AX\|^2 \le \|X\|^2 \sum_{i=1}^{m} (\sum_{j=1}^{n} a_{ij}^2) = k^2 \|X\|^2,$$

which proves the first part. The second part follows from this and one more application of Schwarz:

$$|\langle Y,\, AX\rangle| \le \|Y\|\, \|AX\| \le k\|X\|\|Y\|.$$

After all of this detailed discussion of matrices as an example of a linear operator L mapping one *finite* dimensional space into another, our next theorem will show why matrices are so ubiquitous. You see, we shall prove that *every* such linear operator $L\colon V_1 \to V_2$ can be represented as a matrix *after* bases for V_1 and V_2 have been selected.

Theorem 5.19 . *(Representation Theorem) Let L be a linear operator which maps one finite dimensional space into another*

$$L\colon V_1 \to V_2.$$

Let $\{ e_1, e_2, \dots, e_n \}$ be a basis for V_1, and $\{ \theta_1, \theta_2, \dots \theta_m \}$ be a basis for V_2. Then in terms of these bases L may be represented by the matrix ${}_\theta L_e$ whose jth column is the vector $(Le_j)_\theta$, that is, the vector Le_j (which is a vector in V_2) written in terms of the θ basis for V_2. Pictorially we have

$${}_\theta L_e = ((Le_1)_\theta \cdots (Le_n)_\theta)\,.$$

Proof: Finding the representation of L in terms of given bases for V_1 and V_2 means: given a vector X in V_1 which is represented in the e basis for V_1 (write it as X_e) to find a matrix ${}_\theta L_e$ such that the image vector ${}_\theta L_e X_e$ is the image $(LX)_\theta$ of X written in the θ basis for V_2. We have used the cumbersome notation ${}_\theta L_e$ to make explicit the fact that it maps vectors written in the e basis for V_1 into vectors written in the θ basis V_2.

To avoid even further notation, we shall carry out the details only for the particular case where the domain V_1 is two dimensional with basis $\{ e_1, e_2 \}$ and V_2 is three dimensional with basis $\{ \theta_1, \theta_2, \theta_3 \}$. The general case is proved in the same way.

Since the vectors Le_1 and Le_2 are in V_2, they can be written in the θ basis, say,

$$Le_1 = a_1\theta_1 + b_1\theta_2 + c_1\theta_3,\; Le_2 = a_2\theta_1 + b_2\theta_2 + c_2\theta_3,$$

so

$$(Le_1)_\theta = \begin{pmatrix} a_1 \\ b_1 \\ c_1 \end{pmatrix} \quad \text{and} \quad (Le_2)_\theta = \begin{pmatrix} a_2 \\ b_2 \\ c_2 \end{pmatrix}.$$

Given X in V_1, it can be written in the e basis for V_1,

$$X = x_1e_1 + x_2e_2, \quad \text{so} \quad X_e = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Then

$$\begin{aligned} LX = L(x_1e_1 + x_2e_2) &= x_1Le_1 + x_2Le_2 \\ &= x_1(a_1\theta_1 + b_1\theta_2 + c_1\theta_3) + x_2(a_2\theta_1 + b_2\theta_2 + c_2\theta_3) \\ &= (a_1x_1 + a_2x_2)\theta_1 + (b_1x_1 + b_2x_2)\theta_2 + (c_1x_1 + c_2x_2)\theta_3. \end{aligned}$$

If we write LX as a column vector in the θ basis it is

$$(LX)_\theta = \begin{pmatrix} a_1x_1 + a_2x_2 \\ b_1x_1 + b_2x_2 \\ c_1x_2 + c_2x_3 \end{pmatrix}$$

which is recognized as a product

$$\theta^L e = \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \\ c_1 & c_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \\ c_1 & c_2 \end{pmatrix} X_e$$

therefore, the matrix we want is

$${}_\theta L_e = \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \\ c_1 & c_2 \end{pmatrix} ((Le_1)_\theta\, (Le_2)_\theta),$$

a matrix whose jth column is the vector Le_j written in the θ basis for V_2.

EXAMPLE: Consider the integral operator $L\colon = \int_0^x$ as a map of the two dimensional space $\mathcal{P}_1$ into the three dimensional space $\mathcal{P}_2$. Any bases for $\mathcal{P}_1$ and $\mathcal{P}_2$ will do, however we must simply fix our attention to specific bases. Say
basis for $\mathcal{P}_1 := \{\, e_1(x) = 1, \quad e_2(x) = x\,\}$
basis for $\mathcal{P}_2 := \{\, \theta_1(x) = \frac{1+x}{2}, \quad \theta_2(x) = \frac{1-x}{2}, \quad \theta_3(x) = x^2\,\}$.
Then

$$Le_1 = \int_0^x 1\, dt = x = \theta_1 - \theta_2$$

and

$$Le_2 = \int_0^x t\, dt = \frac{x^2}{2} = \frac{1}{2}\theta_3.$$

Therefore

$$(Le_1)_\theta = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \quad \text{and} \quad (Le_2)_\theta = \begin{pmatrix} 0 \\ 0 \\ \frac{1}{2} \end{pmatrix},$$

so

$${}_\theta L_e = ((Le_1)_\theta\, (Le_2)_\theta) = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$$

is the matrix representing L in terms of the given e basis for $\mathcal{P}_1$ and θ basis for $\mathcal{P}_2$. To make you believe this, let us evaluate

$$L_P = \int_0^x P$$

for some polynomial $p \in \mathcal{P}_1$ by using the matrix. For example, $p(x) = 3 - x = 3e_1 - e_2$, so in the e basis for $\mathcal{P}_1$, $P_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$. Its image under L in terms of the θ basis for $\mathcal{P}_2$ is then

$$(L_p)_\theta =_\theta L_e^{(p_e)} = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 3 \\ -1 \end{pmatrix} = \begin{pmatrix} 3 \\ -3 \\ -\frac{1}{2} \end{pmatrix};$$

that is,

$$L_p = 3\theta_1 - 3\theta_2 - \frac{1}{2}\theta_3 = 3(\frac{1+x}{2}) - 3(\frac{1-x}{2}) - \frac{1}{2}(x^2) = 3x - \frac{1}{2}x^2$$

which, of course, agrees with

$$\int_0^x p(t)\, dt = \int_0^x (3-t)\, dt = 3x - \frac{1}{2}x^2.$$

WARNING: If we had used a different basis for either $\mathcal{P}_1$ or $\mathcal{P}_2$, the resulting matrix representing L would be different. For example, if the *same* basis were used for $\mathcal{P}_1$ but a *different* basis for $\mathcal{P}_2$,

$$\tilde\theta \quad \text{basis for} \quad \mathcal{P}_2 := \{\, \tilde\theta_1(x) = 1, \quad \tilde\theta_2(x) = x, \quad \tilde\theta_3(x) = x^2\,\},$$

then

$$Le_1 = x = \tilde\theta_2 \qquad \text{and} \qquad Le_2 = \frac{x^2}{2} = \frac{1}{2}\tilde\theta_3,$$

so

$$(Le_1)_{\tilde\theta} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad (Le_2)_{\tilde\theta} = \begin{pmatrix} 0 \\ 0 \\ \frac{1}{2} \end{pmatrix}.$$

Therefore the matrix $\tilde\theta^L e$ which represents L in terms of the e basis for $\mathcal{P}_1$ and the $\tilde\theta$ basis for $\mathcal{P}_2$ is

$${}_{\tilde\theta} L_e = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Again, if $p(x) = 3 - x = 3e_1 - e_2$, then in the $\tilde{\theta}$ basis

$$(L_p)_{\tilde{\theta}} = {}_{\tilde{\theta}} L_e P_e = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 3 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \\ -\frac{1}{2} \end{pmatrix};$$

that is,

$$Lp = 0\tilde{\theta}_1 + 3\tilde{\theta}_2 - \frac{1}{2}\tilde{\theta}_3 = 3x - \frac{1}{2}x^2,$$

to no one's surprise.

Observe that the matrices ${}_{\theta}L_e$ and ${}_{\tilde{\theta}}L_3$ both represent L—but with respect to different basis. The second matrix ${}_{\tilde{\theta}}L_e$ is somewhat simpler that the first since it has more zeroes. It is often useful to pick bases in order that the representing matrix be as simple as possible. We shall not discuss that issue right now.

There is a simple class of operators (transformations) which are not linear, but enjoy most of the properties which linear ones do. They are *affine* operators, or affine transformations. To define them, it is best to first define the *translation* operator.

Definition: If V is any linear space and Y_0 a particular element of V, then the operator $T\colon V \to V$ defined by

$$TY = Y + Y_0, \quad Y \in V,$$

is the *translation operator*. It translates a vector Y into the vector $Y + Y_0$.

Definition: An *affine transformation* A is a linear transformation L followed by a translation. if $L\colon V_1 \to V_2$ and $Y_0 \in V_2$, it has the form

$$AX := LX + Y_0. \qquad X \in V_1, \quad Y_0 \in V_2.$$

Affine transformations can be added and multiplied by the same definition which governed linear transformations. Thus, if A and B are affine transformations mapping V_1 into V_2,

$$(A + B)X := AX + BX.$$

In particular, if $AX = L_1X + Y_0$ and $BX = L_2X + Z_0$, where Y_0 and Z_0 are in V_2, then

$$\begin{aligned}(A + B)X &= AX + BX = L_1X + Y_0 + L_2X + Z_0 \\ &= (L_1 + L_2)X + (Y_0 + Z_0).\end{aligned}$$

Similarly, if $A\colon V_1 \to V_2$ and $B\colon V_3 \to V_4$, where $V_2 \subset V_3$, then

$$\begin{aligned}(BA)X := B(AX) &= B(L_1X + Y_0) = L_2(L_1X + Y_0) + Z_0 \\ &= L_2L_1X + L_2Y_0 + Z_0,\end{aligned}$$

where $Y_0 \in V_2$ and $Z_0 \in V_4$.

You will carry out the (straightforward) proofs of the algebraic properties for affine transformations in Exercise 23.

The curtain on this longest of sections will be brought down with a brief discussion of the operators which characterize rigid body motions, or Euclidean motions, as they are often called.

Definition: The transformation $R\colon \mathbb{R}^n \to \mathbb{R}^n$ is an isometric *transformation*, (or *Euclidean transformation* or rigid body transformation) if the distance between two points is preserved (invariant) under the transformation. Thus, R is an isometry if

$$\|RX - RY\| = \|X - Y\|$$

for all X and Y in $\mathbb{R}^n$.

It is interesting to think for a moment how all these names originated. The phrase rigid body transformation arises from the idea that any motion of a rigid body (such as a translation or rotation) does not alter the distance between any two points in the body. In the framework of Euclidean geometry the whole notion of *congruence* is defined to be just those properties of a figure which are invariant under isometries. By allowing deformations other than isometries, one obtains geometries, so affine geometry is the study of properties invariant under all affine motions.

The study of isometric transformations is mainly contained in that of a special case, *orthogonal transformations*. These are isometries which leave the origin fixed, $R0 = 0$. It should be clear from our next theorem (part 3) that the idea of an orthogonal transformation generalizes the idea of a rotation to higher dimensional space. Reflections (mirror images) are also orthogonal transformations. Theorem 20 states that every isometric transformation is the result of an orthogonal transformation followed by a translation.

Example: The matrix $R = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ defines an orthogonal transformation since if

$$X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \text{then} \quad RX = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ -x_2 \end{pmatrix},$$

and if

$$Y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad \text{then} \quad RY = \begin{pmatrix} y_1 \\ -y_2 \end{pmatrix}.$$

Consequently $\|RX - RY\| = \|X - Y\| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$, so R, being isometric and linear is an orthogonal transformation. It represents a reflection across the x_1 axis.

Our definition of an orthogonal transformation does not presume its linearity. This is because the linearity is a *consequence* of the given properties. A proof is outlined in Ex. 16, p. 390. For convenience, the linearity will be assumed in the following theorem where we collect the standard properties of orthogonal transformations.

Theorem 5.20 . *Let $R\colon \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation. The following properties of R are equivalent.*

(1) *R is an orthogonal transformation, that is*

$$\|RX - RY\| = \|X - Y\| \quad \textit{and} \quad R0 = 0.$$

(2) $\|RX\| = \|X\|$

(3) $\langle RX,\, RY\rangle = \langle X,\, Y\rangle$ *(so angles are preserved)*

(4) $R^*R = I$

(5) *R is invertible and $R^{-1} = R^*$. (Only in this part do we use the finite dimensionality of $\mathbb{R}^n$).*

PROOF: We shall prove the following chain of implications: $1 \Longrightarrow 2 \Longrightarrow 3 \Longrightarrow 4 \Longrightarrow 5 \Longrightarrow 4 \Longrightarrow 1$
$1 \Longrightarrow 2$. Trivial, for

$$\|RX\| = \|RX - R0\| = \|X - 0\| = \|X\|.$$

$2 \Longrightarrow 3$. By linearity and part 2) applied to the vector $X + Y$, we have

$$\|RX + RY\| = \|R(X + Y)\| = \|X + Y\|.$$

Now square both sides and express the norm as a scalar product:

$$\langle RX + RY,\, RX + RY\rangle = \langle X + Y,\, X + Y\rangle.$$

Upon expanding both sides, we find that

$$\|RX\|^2 + 2\langle RX,\, RY\rangle + \|RY\|^2 = \|X\|^2 + 2\langle X,\, Y\rangle + \|Y\|^2.$$

Since by part 2) $\|RX\| = \|X\|$ and $\|RY\| = \|Y\|$, we are done.
$3 \Longrightarrow 4$. By part 3) and Theorem 14 (p. 369),

$$\langle R^*RX,\, Y\rangle = \langle RX,\, RY\rangle = \langle X,\, Y\rangle.$$

Thus, an application of the second part of Theorem 16 (p. 371) gives us $R^*R = I$.
$4 \Longrightarrow 5$. Since $X = R^*RX$, we see that $RX = 0$ implies $X = 0$, consequently, R is invertible (Theorem 12, p. 364). Moreover $R^*R = I$ so $R^* = R^{-1}$.
$5 \Longrightarrow 4$. Clear, since $R^* = R^{-1}$.
$5 \Longrightarrow 1$. Because R is linear, $R0 = 0$. It remains to show that $\|RX - RY\| = \|X - Y\|$, an easy computation. $\|RX - RY\|^2 = \|R(X - Y)\|^2 = \langle R(X - Y),\, R(X - Y)\rangle$, so using 4)

$$= \langle R^*R(X - Y),\, X - Y\rangle = \langle (X - Y),\, (X - Y)\rangle = \|X - Y\|^2.$$

Done.

Earlier in this section (p. 357-8) we considered a matrix R which represented the operator which rotates a vector in $\mathbb{R}^2$ through an angle α. This matrix is the simplest (non-trivial) example of a rigid body transformation which leaves the origin fixed, that is, an orthogonal transformation.

$$R = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}.$$

To prove that R is an orthogonal matrix, by Theorem 19 part 3, it is sufficient to verify $\langle RX,\, RY\rangle = \langle X,\, Y\rangle$ for all X and Y in $\mathbb{R}^2$. A calculation is in order here.

$$RX = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1\cos\alpha - x_2\sin\alpha \\ x_1\sin\alpha + x_2\cos\alpha \end{pmatrix}.$$

Similarly for RY, just replace x_1 and x_2 by y_1 and y_2 respectively. Then

$$\begin{aligned}\langle RX,\, RY\rangle = (x_1\cos\alpha - x_2\sin\alpha)(y_1\cos\alpha - y_2\sin\alpha) + \text{ missing?}\\ (x_1\sin\alpha + x_2\cos\alpha)(y_1\sin\alpha + y_2\cos\alpha)\\ = x_1y_1\cos^2\alpha - (x_1y_2 + x_2y_1)\sin\alpha\cos\alpha + x_2y_2\sin^2\alpha\\ + x_1y_1\sin^2\alpha + (x_1y_2 + x_2y_1)\sin\alpha\cos\alpha + x_2y_2\cos^2\alpha\\ = x_1y_1 + x_2y_2 = \langle X,\, Y\rangle. \quad \text{Done.}\end{aligned}$$

We previously found an expression for R^{-1} (p. 358) by geometric reasoning. It is reassuring to notice $R^{-1} = R*$, just as part 5 of our theorem states.

The most general rotation in $\mathbb{R}^3$ may be decomposed into a product of these simple two dimensional rotations. For a brief discussion - complete with pictures - open Goldstein, *Classical Mechanics* to pp. 107-9.

Now to the last theorem of this section.

Theorem 5.21 . *If $R\colon \mathbb{R}^n \to \mathbb{R}^n$ is a rigid body transformation, then for every $X \in \mathbb{R}^n$*

$$RX = R_0X + X_0,$$

where R_0 is an orthogonal transformation (rotation) and X_0 is a fixed vector in $\mathbb{R}^n$. Thus, every rigid body motion is composed of a rotation (by R_0 and a translation (through X_0).

PROOF: Let $R_0X = RX - R0$. Since

$$R_00 = R0 - R0 = 0,$$

the operator R_0 has the property $R_00 = 0$. Furthermore, for any X and Y in $\mathbb{R}^n$,

$$\begin{aligned}\|R_0X - R_0Y\| &= \|RX - R0 - RY + R0\|\\ &= \|RX - RY\| = \|X - Y\|.\end{aligned}$$

Therefore R_0 satisfies the definition of an orthogonal transformation. The proof is completed by defining X_0 to be the image of the origin under R, $X_0 = R0$. Then

$$R_0X = RX - X_0,$$

or

$$RX = R_0X + X_0.$$

5.2 Supplement on Quadratic Forms

Quadratic polynomials of the form

$$Q(X) = \alpha x_1^2 + \beta x_1 x_2 + \gamma x_2^2, \quad X = (x_1, x_2)$$

and the generalization to n variables $X = (x_1, x_2, \dots, x_n)$

$$Q(X) = \sum_{i}^{n} \sum_{j=1}^{n} \alpha_{ij} x_i x_j$$

often arise in mathematics. They are called *quadratic forms* and can always be represented in the form $\langle X, SX \rangle$ where S is a self adjoint matrix. For example, the first quadratic form can be written as

$$Q(X) = (x_1, x_2) \begin{pmatrix} \alpha & \frac{\beta}{2} \\ \frac{\beta}{2} & \gamma \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \langle X, SX \rangle,$$

where S is the matrix indicated.

The procedure for finding the elements $((a_{ij}))$ of the matrix S is simple. First take care of the diagonal terms by letting a_{ii} be the coefficient of x_1^2 in $Q(X)$. Realizing that $x_i x_j = x_i x_j$, collect the terms $\alpha_{ij} x_i x_j$ and $\alpha_{ji} x_j x_i$ in $Q(X)$, getting $(\alpha_{ij} + \alpha_{ji}) x_i x_j$. Then let

$$a_{ij} = a_{ji} = \frac{1}{2}(\alpha_{ij} + \alpha_{ji}) \quad i \neq j.$$

Example: $Q(X) = x_1^2 - 2x_1x_3 - x_2^2 + 6x_1x_2 + 4x_3x_1$. Rewrite this as $Q(X) = x_1^2 - x_2^2 + 6x_1x_2 + 2x_1x_3$. Then

$$S = \begin{pmatrix} 1 & 3 & 1 \\ 3 & -1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

and

$$Q(X) = \langle X, SX \rangle.$$

as you can easily verify.

Definition: A quadratic form $Q(X)$ is *positive semi definite* if $Q(X) \geq 0$ for all X and *positive definite* if $Q(X) > 0$, $x \neq 0$. $Q(X)$ is *negative semi definite* or *negative definite* if, respectively, $Q(X) \leq 0$, or $Q(X) < 0$, $X \neq 0$. If S is the self adjoint matrix associated with the quadratic form $Q(X)$, then S is positive semi definite, positive definite, etc., if $Q(X)$ has the respective property.

We may think of $Q(X)$ as representing a quadratic surface. Thus, if S is diagonal, for example

$$S = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix},$$

with positive diagonal elements, then the equation $Q(X) = 1$, where $Q(X) = \langle X, SX \rangle = 2x_1^2 + x_2^2 + 3x_3^2$, represents an ellipsoid. This matrix S is positive definite since by inspection $Q(X) > 0, \quad X \neq 0$.

It is easy to see if a diagonal matrix S is positive semi definite, negative semi definite, positive definite, or negative definite.

Example: The diagonal matrix

$$S = \begin{pmatrix} \gamma_1 \dots 0 \\ \ddots \\ 0 \dots \gamma_n \end{pmatrix}$$

is

(a) positive semi definite if and only if $\gamma_1, \dots, \gamma_n$ are all non-negative,

(b) positive definite if and only if $\gamma_1, \dots, \gamma_n$ are all positive (not zero), and the obvious statements for negative semi definite and negative definite.

The problem of determining if a non diagonal symmetric matrix is positive etc. is more subtle. We shall find necessary and sufficient conditions for the two variable case, but only necessary conditions for the general case.

Consider the 2×2 self-adjoint matrix

$$S = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

and the associated quadratic form

$$Q(X) = ax^2 + 2bxy + cy^2.$$

There are several cases.

(i) *If* $a = 0$, then

$$Q(X) = (2bx + cy)y.$$

If $b \neq 0$, by choosing x and y appropriately, we can make $Q(X)$ assume *both* positive and negative values. Thus, for $a = 0$, $b \neq 0$, Q can be neither a positive nor a negative semi-definite form. On the other hand, if $a = 0$, and $b = 0$, then Q is positive (negative) semi definite if and only if $c \geq 0\,(c \leq 0)$. If $a = 0$, Q can never be positive definite or negative definite since if $X = (x, 0)$ where $x \neq 0$, then $Q(X) = 0$ but $X \neq 0$.

(ii) *If* $a \neq 0$, then Q can be written as

$$Q(X) = \frac{1}{a}[(ax + by)^2 = (ac - b^2)y^2].$$

We can immediately read off the conditions from this. Q is positive semi definite (definite) if and only if $a > 0$ and $ac - b^2 \geq 0\,(ac - b^2 > 0)$, and negative semi definite (definite) if and only if $a < 0$ and $ac - b^2 \geq 0\,(ac - b^2 > 0)$.

In summary, we have proved

Theorem 5.22 *A. Let $Q(X) = ax^2 + 2bxy + cy^2$, and S be the associated symmetric matrix. Then*

(a) Q is positive semi definite if and only if $a \geq 0$ and $ac - b^2 \geq 0$ (this implies $c \geq 0$ too).

(b) Q is positive definite if and only if $a > 0$ and $ac - b^2 > 0$ (this implies $c > 0$ too).

The general case of a quadratic form in n variables is much more difficult to treat. There are known necessary and sufficient conditions, but they are not too useful in practice, especially for a large number of variables. We shall only prove one necessary condition for a quadratic form to be positive semi-definite (or positive definite), a condition which is both transparent to verify in practice and even easier to prove.

THEOREM B. If the self adjoint matrix $S = ((a_{ij}))$ is positive definite, then the *diagonal* elements must all be positive, $a_{11}, a_{22}, \ldots, a_{nn} > 0$. Similarly, if S is negative definite then the diagonal elements must all be negative.

PROOF: $Q(X) = \langle X, SX\rangle = \sum_{i,j=1}^{n} a_{ij}x_ix_j$. Since Q is positive definite, $Q(X) > 0$ for all $X \neq 0$. In particular, $Q(e_k) > 0$, $k = 1, \ldots, n$, where e_k is the kth coordinate vector $e_k = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$. But $Q(e_k) = a_{kk}$. Thus $a_{kk} > 0$, $k = 1, \ldots, n$, just what we wanted to prove.

EXAMPLES: 1. The quadratic form $Q(X) = 3x^2 + 743xy - y^2 + 4z^2 + xz$ is positive definite or semi definite since the coefficient of y^2 is negative. It is not negative definite or semi definite since the coefficient of x^2 is positive.

2. The quadratic form $Q(X) = x^2 - 5xy + y^2 + 2z^2$ satisfies the necessary conditions of Theorem B, but the conditions of Theorem D were not sufficient conditions for positive definiteness. Thus, we cannot conclude this $Q(X)$ is positive definite. In fact, this $Q(X)$ is *not* positive definite or semi definite since, for example, if $X = (1, 1, 1)$, then $Q(X) = -1$. It is clearly not negative definite or semi definite.

Exercises

(1) Find the self-adjoint matrix S associated with the following quadratic forms:

(a) $Q(X) = x_1^2 - 2x_1x_2 + 4x_2^2$.

(b) $Q(X) = -x_1^2 + x_1x_2 - x_1x_3 + x_2^2 - 3x_2x_1 - 2x_3x_2 + 3x_3^2$

(c) $Q(X) = 2x_1x_2 - 3x_3x_2 + 4x_2x_4 + x_3x_4 + 7x_2^2$

[Answers: (a) $\begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix}$, (b) $\begin{pmatrix} -1 & -1 & -\frac{1}{2} \\ -1 & 1 & -1 \\ -\frac{1}{2} & -1 & 3 \end{pmatrix}$, (c) $\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 7 & -\frac{3}{2} & 2 \\ 0 & -\frac{3}{2} & 0 & \frac{1}{2} \\ 0 & 2 & \frac{1}{2} & 0 \end{pmatrix}$

(2) Use Theorem A or B to determine which of the following quadratic forms in two variables are positive or negative definite, or semi definite, or none of these.

(a) $Q(X) = x_1^2 - 2x_1x_2 + 4x_2^2$

(b) $Q(X) = -x_1^2 + x_1x_2 - 4x_2^2$

(c) $Q(X) = x_1^2 - 6x_1x_2 - 4x_2^2$

(d) $Q(X) = x_1^2 - 6x_1x_2 + 4x_2^2$

(e) $Q(X) = x_1^2 - 6x_1x_2 + 4x_2x_3 - x_2^2 + 4x_3^2$

(3) If the self-adjoint matrix S is positive definite, prove it is invertible. Give an example of an invertible self-adjoint matrix which is neither positive nor negative definite.

(4) Find all real values for λ for which the quadratic form

$$Q(X) = 2x^2 + y^2 + 3z^2 + 2\lambda xy + 2xz$$

is positive definite. [Hint: $Q(X) = (\frac{5}{3} - \lambda^2)x^2 + (\lambda x + y)^2 + (\sqrt{3}z + \frac{1}{\sqrt{3}}x)^2$]

(5) Let the integer n be ≥ 3. If the quadratic form

$$Q(X) = \sum_{i,j=1}^{n} a_{ij}x_ix_j, \qquad a_{ij} = a_{ji}$$

is the product of two linear forms

$$Q(X) = (\sum_{i=1}^{n} \lambda_i x_i)(\sum_{j=1}^{n} \mu_j x_j),$$

show that $\det A = \det((a_{ij})) = 0$.

(6) If the self-adjoint matrix S is positive definite or semi-definite, prove the *generalized Schwarz inequality*:

$$|\langle Y, SX\rangle|^2 \leq \langle Y, SY\rangle\langle X, SX\rangle$$

for all X and Y. [Hint: Observe $[X, Y] := \langle Y, SX\rangle$ satisfies all the axioms for a scalar product].

(7) If the self-adjoint matrix S is positive definite (so S^{-1} exists by Exercise 3), prove that S^{-1} is also positive definite. [Hint: Use the generalized Schwarz inequality, Exercise 6, with $Y = S^{-1}X$ and the inequality $\langle X, SX\rangle \leq k^2\|X\|^2$ of Theorem 17, p. 373].

(8) Proof or counterexample:

(a) If a matrix $A = ((a_{i_j}))$ is positive definite, then all of its elements are positive, $a_{ij} > 0$ for all i, j.

(b) If a matrix A is such that all of its elements are positive, $a_{ij} > 0$, then the matrix is positive definite.

Exercises

(1) Write out the matrices associated with the operators A and B in Exercise 4a, p. 281, and carry out the computation there using matrices.

(2) Write out the matrices R_A, R_B, and R_C for the rotation operators A, B, and C in Exercise 8 p. 281 and complete that problem using matrices. [Ans. $R_A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$ in terms of the basis $e_1 = (1,0,0)$, $e_2 = (0,1,0)$, $e_3 = (0,0,1)$].

(3) Prove Exercise 2b (p. 281) as a corollary of Theorem 18.

(4) If

$$A = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

compute AB, BA, and B^2.

(5) Compute A^{-1} if

(a). $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ [ans. $A^{-1} = \begin{pmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix}$]

(b). $A = \begin{pmatrix} 4 & 0 & 5 \\ 0 & 1 & -6 \\ 3 & 0 & 4 \end{pmatrix}$ [ans. $A^{-1} = \begin{pmatrix} 4 & 0 & -5 \\ -18 & 1 & 24 \\ -3 & 0 & 4 \end{pmatrix}$]

(c) $A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ [ans. $A^{-1} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$]

(6) If A is the matrix of 5a) above, from the definition compute *directly*,

(a) $-6A^{-1} + \frac{1}{2}A^*$ [ans. $\begin{pmatrix} \frac{25}{2} & -\frac{9}{2} \\ -8 & 5 \end{pmatrix}$].

(b) $(A^*)^{-1}$ and $(A^{-1})^*$. Compare them. State and prove a general theorem.

(c) AA^* and A^*A.

(7) If

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix},$$

compute $(AB)^*$, A^*B^*, and B^*A^*. Compare $(AB)^*$ and B^*A^* and explain the outcome.

(8) Prove that $I^* = I$ and $A^{**} = A$ using only Theorems 14 and 16 (cf. Parts 2-4 of Theorem 15).

(9) If $A\colon \mathbb{R}^n \to \mathbb{R}^m$ and $B\colon \mathbb{R}^m \to \mathbb{R}^n$ where $n > m$, prove that BA (an $n \times n$ matrix) is singular. Is AB necessarily singular? (Proof or counterexample).

(10) Given two square matrices A and B such that $AB = 0$, which of the following statements are *always* true. Proofs or counterexamples are called for. [I suggest you confine your search for counterexamples to the case of 2×2 matrices.]

(a). $A = 0$.

(b). $B = 0$.

(c). A and/or B are (is) singular (not invertible).

(d). A is singular.

(e). B^{-1} exists.

(f). If A^{-1} exists, then $B = 0$.

(g). If B is nonsingular, then $A = C$.

(h). $BA = 0$.

(i). If $A \neq 0$ and $B \neq 0$, then neither A nor B are invertible.

(11) (a). If A is a square matrix which satisfies

$$A^2 - 2A - I = 0,$$

find A^{-1} in terms of A. [Hint: Find a matrix B such that $AB = BA = I$.]

(b). If A is a square matrix which satisfies

$$A^n + a_{n-1}A^{n-1} + a_{n-2}A^{n-2} + \ldots + a_1 A + a_0 I = 0, \quad a_0 \neq 0,$$

where $a_0, a_1, \ldots, a_{n-1}$ are scalars, prove that A is invertible and find A^{-1} in terms of A.

(12) (a). If $L\colon \mathbb{R}^n \to \mathbb{R}^m$, prove $\mathcal{N}(L^*) = \mathcal{R}(L)^\perp$

[Hint: Show (in two lines) that $X \in \mathcal{N}(L^*) \iff \langle X, LZ\rangle = 0$ for all $Z \in \mathbb{R}^n$—from which the result is immediate.]

(b). Use part (a) to show that $\dim \mathcal{R}(L) = \dim \mathcal{R}(L^*)$.

(c). Do exercise 19, page 441.

(13) (a). If $T\colon \mathbb{R}^n \to \mathbb{R}^n$ is a translation, $TX = X + X_0$, prove T is invertible by explicitly finding T^{-1} (which is a trivial task). [Answer: $T^{-1}X = X - X_0$.]

(b). If $R\colon \mathbb{R}^n \to \mathbb{R}^n$ is a rigid body transformation, show that R is always invertible by exhibiting R^{-1}. [Answer: If $RX = R_0X + X_0$, then R can be written as $Rx = (TR_0)X$. $R^{-1} = R_0^* T^{-1}$.]

(14) If A is any $n \times n$ matrix, find matrices A_1 and A_2 such that A is decomposed into the two parts

$$A = A_1 + A_2$$

where A_1 is symmetric and A_2 is *anti-symmetric*, i.e., $A_2^* = -A_2$. [Hint: Assume there is such a decomposition and use it to find A_1 and A_2 in terms of A and A^*. Then verify that these work.]

(15) Consider the operator $D = \frac{d}{dx}$ on $\mathcal{P}_5$. Prove that D is not invertible (return to the definition p. 360) but exhibit an operator L which is a right inverse, $DL = I$.

(16) This problem proves that R is orthogonal if and only if R is linear and isometric.

(a) Prove that if R is linear and isometric, then it is orthogonal. (Trivial!).

(b) If R is orthogonal, prove that

i) $\|RX\| = \|X\|$

ii) $\langle RX, RY\rangle = \langle X, Y\rangle$ (Hint: Use $\|RX - RY\|^2 = \|X - Y\|^2$)

iii) $R(aX) = aRX$ (Hint: Prove $\|R(aX) - aRX\|^2 = 0$)

iv) $R(X + Y) = RX + RY$ (Hint: Prove $\|$ "something" $\|^2 = 0$)

v) R is linear and isometric

[Warning: If you assume linearity in b), you'll vitiate the whole problem].

(17) (a). Let A be a square matrix such that $A^5 = 0$. Verify that $(I + A)^{-1} = I - A + A^2 - A^3 + A^4$.

(b). If $A^7 = 0$, then $(I - A)^{-1} = ?$

(18) Consider the matrices

$$\text{(a).}\ \begin{pmatrix} \alpha & \frac{1}{2} \\ -\frac{1}{2} & \delta \end{pmatrix}, \qquad \text{(b).}\ \begin{pmatrix} \frac{1}{\sqrt{2}} & \beta \\ \gamma & \frac{1}{\sqrt{2}} \end{pmatrix},$$

$$\text{(c).}\ \begin{pmatrix} 0 & \beta \\ \gamma & 0 \end{pmatrix}, \qquad \text{(d).}\ \begin{pmatrix} 1 & \beta \\ 0 & 2 \end{pmatrix}.$$

For what value(s) of α, β, γ and δ do these matrices represent orthogonal transformations?

(19) If $A = ((a_{ij}))$ is a square $(n \times n)$ matrix, the *trace* of A is defined as the sum of the elements on the main diagonal, $\operatorname{tr} A\colon = a_{11} + a_{22} + \ldots + a_{nn}$. Prove

(a). $\operatorname{tr}(\alpha A) = \alpha A$, where α is a scalar.

(b). $\operatorname{tr}(A + B) = \operatorname{tr} A + \operatorname{tr} B$, where B is also an $n \times n$ matrix.

(c). $\operatorname{tr}(AB) = \operatorname{tr}(BA)$.

(d). $\operatorname{tr}(I) =?$

(20) Assume that $A\colon \mathbb{R}^n \to \mathbb{R}^n$ is anti-symmetric, $A^* = -A$.

(a). Prove $A - I$ is invertible. [By Theorem 12, it is sufficient to show $(A - I)X = 0 \Rightarrow X = 0$. Use the property of A to prove it $AX = X$, then $\langle X, AX\rangle = \|X\|^2$, $\langle A^*X, X\rangle = -\|X\|^2$, and $\langle X, AX\rangle = \langle A^*X, X\rangle$.]

(b). If $U = (A + I)(A - I)^{-1}$, then U is an orthogonal transformation.

(21) Let A_n be the orthogonal matrix which rotates vectors in $\mathbb{R}^2$ through an angle of $2\pi/n$.

(a). Find a matrix representing A_n (use the standard basis for $\mathbb{R}^2$).

(b). Let B denote the orthogonal matrix of reflection across the x_1 axis (p. 382). Show that $BA_b = A_n^{-1}B$. [The group of matrices generated by A_n and B and all possible products is the *dihedral group of order n*].

(22) Prove that the set of all orthogonal transformations of $\mathbb{R}^n$ into $\mathbb{R}^n$ forms a (non-commutative) group under multiplication.

(23) An affine transformation $AX = LX + X_0$ of a linear space into *itself* is called non-singular if the linear transformation L is non-singular. Prove that the set of all such non-singular affine transformations form a (non-commutative) group under multiplication.

(24) Let $A = ((a_{ij}))$ be a square matrix. Find all such matrices with the property that $\operatorname{tr}(AA^*) = 0$ (see Ex. 19 for the definition of the trace).

(25) Consider the linear space

$$\mathcal{S} = \{\, f(x)\colon f(x) = a + b\cos x + c\sin x \,\}.$$

with the scalar product

$$\langle f, g\rangle = a\tilde{a} + \frac{1}{2}(b\tilde{b} + c\tilde{c}),$$

where $g(x) = \tilde{a} + \tilde{b}\cos x + \tilde{c}\sin x$. Define the linear transformation $R\colon \mathcal{S} \to \mathcal{S}$ by the rule

$$(Rf)(x) = f(x + \alpha), \quad \alpha \text{ real}.$$

(a) Show that R is an orthogonal transformation by proving that $\langle Rf, Rg\rangle = \langle f, g\rangle$ for all f, g in $\mathcal{S}$.

(b) Choose a basis for $\mathcal{S}$ and exhibit a matrix ${}_eR_e$ which represents R with respect to that basis for both the domain and target.

(26) Let $A\colon \mathbb{R}^n \to \mathbb{R}^m$. Prove: A is surjective (= onto) if and only if A^* is injective (= one to one).

(27) Define $A\colon \mathcal{P}_3 \to \mathbb{R}^3$ by

$$A[p(x)] = (p(0),\, p(1),\, p(-1)) \quad \text{where } p \in \mathcal{P}_3.$$

Find the matrix for this transformation with respect to the basis $e_1 = 1 \quad e_2 = (x+1)^2$, $e_3 = (x-1)^2$, $e_4 = x^3$ for $\mathcal{P}_3$; and the standard basis for $\mathbb{R}^3$.

(b). Find the matrix representing A using the same basis for $\mathbb{R}^3$ but using the basis $\hat{e}_1 = 1$, $\hat{e}_2 = x$, $\hat{e}_3 = x^2$ and $\hat{e}_4 = x^3$ for $\mathcal{P}_3$.

(28) If A and B both map the linear space V into itself, and if B is the *only* right inverse of A, $AB = I$, prove A is invertible. [Hint: Consider $BA + B + I$].

(29) Let $A\colon \mathbb{R}^n \to \mathbb{R}^m$ be represented by the matrix $((a_{ij}))$, and $B\colon \mathbb{R}^m \to \mathbb{R}^n$ by $((b_{ij}))$. If

$$\langle Y,\, AX\rangle = \langle BY,\, X\rangle$$

for all $X \in \mathbb{R}^n$ and all $Y \in \mathbb{R}^m$, prove $B = A^*$. This proves the statement made in the remark following Theorem 14.

(30) Let $L\colon \mathbb{R}^4 \to \mathbb{R}^4$ be defined by $LX = (x_1, 0, x_3, 0)$, where $X = (x_1, x_2, x_3, x_4)$. Find a matrix representing L in terms of some basis. You may use the same basis for both the domain and the target.

5.3 Volume, Determinants, and Linear Algebraic Equations.

Often we have stated that thus and so is true if and only if a certain set of vectors are linearly independent. But we still have no adequate criteria for determining if a set of vectors is linearly independent. What would be an ideal criterion? One superb criteria would be as follows. Find a function which assigns to a set of n vectors $X_1, X_2, \ldots, X_n$ in $\mathbb{R}^n$ a real number, with the property that this number is zero if and only if the vectors are linearly dependent.

There is a geometric way of solving this problem. For clarity we shall work in two dimensions, $\mathbb{R}^2$. If X_1 and X_2 are any two vectors in $\mathbb{R}^2$, then intuition tells us X_1 and X_2 are linearly dependent if and only if the area of the parallelogram (see fig.) is zero. Thus, once we define the analogue of volume for n dimensional parallelepipeds in $\mathbb{R}^n$, the appropriate criterion appears to be that a set of n vectors $X_1, \ldots, X_n$ in $\mathbb{R}^m$ is linearly dependent if and only if the volume of the parallelepiped they span is zero.

The major hurdle is constructing a volume function which behaves in the manner dictated by two and three dimensional intuition. Our program is to state a few (four to be

exact) desirable properties of a volume function V for parallelepipeds, then construct a simpler related function - the determinant D, and observe that $V = |D|$ (absolute value of D) is a volume function. This determinant function will prove useful in the theory of linear algebraic equations.

Let X_1 and X_2 be any two vectors in $\mathbb{R}^2$. We define the *parallelogram spanned* by X_1 and X_2 to be the set of points X in $\mathbb{R}^2$ which have the form

$$X = t_1X_1 + t_2X_2, \qquad 0 \le t_1 \le 1,\ 0 \le t_2 \le 1$$

You can check that these points are precisely those in the parallelogram drawn above. The volume function (really area in this case) $V(X_1, X_2)$ which assigns to each parallelogram its volume should have the properties

1. $V(X_1, X_2) \ge 0$.
2. $V(\lambda X_1, X_2) = |\lambda|\, V(X_1, X_2), \quad \lambda$ scalar.
3. $V(X_1 + X_2, X_2) = V(X_1, X_2) = V(X_1, X_1 + X_2)$.
4. $V(e_1, e_2) = 1$. $e_1 = (1, 0)$, $e_2 = (0, 1)$.

The second property states that if one side is multiplied by λ_1 then the volume is multiplied by $|\lambda|$ (see fig.).

The third property is more subtle. It states that the volume of the parallelogram spanned by X_1 and X_2 is the same as the parallelogram spanned by X_1 and $X_1 + X_2$. This is clear from the figure since both parallelograms have the same base and height.

The last property merely normalizes the volume. It states that the unit square has volume 1.

Our first task is to define a parallelepiped in $\mathbb{R}^n$.

Definition: The n dimensional *parallelepiped* in $\mathbb{R}^n$ spanned by a linearly independent set of vectors $X_1, X_2, \ldots, X_n$ is the set of all points X in $\mathbb{R}^n$ of the form

$$X = t_1X_1 + t_2X_2 + \cdots + t_nX_n, \qquad 0 \le t_j \le 1.$$

It is a straightforward matter to write the axioms for the volume $V(X_1, X_2, \ldots, X_n)$ for the n dimensional parallelepiped in $\mathbb{R}^n$.

V-1. $V(X_1, X_2, \ldots, X_n) \ge 0$.

V-2. $V(X_1, X_2, \ldots, X_n)$ is multiplied by $|\lambda|$ if some X_j is replaced by λX_j where λ is real.

V-3. $V(X_1, X_2, \ldots, X_n)$ does not change if some X_j is replaced by $X_j + X_k$, where $j \ne k$.

V-4. $V(e_1, e_2, \ldots, e_n) = 1$, where $e_1 = (1, 0, 0, \ldots, 0)$, etc.

These axioms are amazingly simple. It is surprising that the volume function V in *uniquely* determined by them; that is, there is only one function which satisfies these axioms. You might wonder why we did not add the reasonable stipulation that volume remains unchanged if the parallelepiped is subjected to a rigid body transformation. The reason is that this axiom would be redundant, for this invariance of volume under rigid body transformation will be one of our theorems.

The most simple way to obtain the volume function is to first obtain the determinant function $D(X_1, X_2, \dots, X_n)$. We *define* the *determinant* function $D(X_1, X_2, \dots, X_n)$ of n vectors $X_1, X_2, \dots, X_n$ in $\mathbb{R}^n$ by the following axioms (selected from those for V).

D-1. $D(X_1, X_2, \dots, X_n)$ is a real number.

D-2. $D(X_1, X_2, \dots, X_n)$ is multiplied by λ if some X_j is replaced by λX_j where λ is real.

D-3. $D(X_1, X_2, \dots, X_n)$ does not change if some X_j is replaced by $X_j + X_k$, where $j \neq k$.

D-4. $D(e_1, e_2, \dots, e_n) = 1$, where $e_1 = (1, 0, 0, \dots, 0)$ etc.

Remarks:

(1) If $A = ((a_{ij}))$ is a (square) $n \times n$ matrix,

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

we can consider it as being composed of n column vectors $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n$, and define the determinant of the square matrix A in terms of the determinant of these vectors

$$\det A = D(\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n) = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & & & a_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ a_{n1} & \cdots & \cdots & a_{nn} \end{vmatrix}.$$

(2) Although we have written a set of axioms for D, it is not at all obvious that such a function exists. Rest assured that we will prove the existence of such a function.

(3) Observe: if we define

$$V(X_1, X_2, \dots, X_n) := |D(X_1, X_2, \dots, X_n)|, \qquad X_j \in \mathbb{R}^n,$$

then V does satisfy the axioms for volume.

Granting existence of D, we derive some algebraic consequences of the axioms.

Theorem 5.23 . *Let D be a function which satisfies axiom D-1 to D-3 (not necessarily D-4).*

(1) *If X_j is replaced by $X_j = \sum_{k \neq j} \lambda_k X_k$ then D does not change.*

(2) *If one of the vectors X_j is zero, then $D = 0$.*

(3) *If the vectors $X_1, X_2, \dots, X_n$ are linearly dependent then $D = 0$. In particular $D = 0$ if two vectors are equal.*

(4) *D is a linear function of each of its variables, that is*

$$D(\dots, \lambda Y + \mu Z, \dots) = \lambda D(\dots, Y, \dots) + \mu D(\dots, Z, \dots)$$

(so D is a multilinear function).

(5) *If any two vectors X_i and X_j are interchanged, then D is multiplied by -1.*

$$D(\dots, X_i, \dots, X_j, \dots) = -D(\dots, X_j, \dots, X_i, \dots)$$

PROOF: These proofs, like the statements above, are conceptually simple but notationally awkward. Notice that only Axioms 1-3 but not Axiom 4 will be used. We shall need this fact shortly.

(1) We prove this only if X_j is replaced by $X_j + \lambda X_k$, $\quad j \neq k$ and $\lambda \neq 0$. The general case is a simple repetition of this until the other X_k's are used up. It is simplest to work backward. By Axiom 2,

$$D(\dots, X_j + \lambda X_k, \dots, X_k \dots) = \frac{1}{\lambda} D(\dots, X_j + \lambda X_k, \dots, \lambda X_k, \dots)$$

so by axiom 3 (since λX_k is now a vector in D)

$$= \frac{1}{\lambda} D(\dots, X_j, \dots, \lambda X_k, \dots)$$

and axiom 2 again

$$= D(\dots, X_j, \dots, X_k, \dots).$$

(2) Write the vector $X_j = 0$ as $0X_j$ where 0 is now a scalar. This scalar may be brought outside D by axiom 2. Since D is a real number, $0 \cdot D = 0$.

(3) Let $X_j = \sum_{k \neq j} a_k X_k$. By part 1, D does not change if X_j is replaced by $X_j + \sum_{k \neq j} \lambda_k X_k$. Choose $\lambda_k = -a_k$. This gives a D with one vector zero, $X_j - \sum_{k \neq j} a_k X_k = 0$. Thus D is zero by part 2.

(4) The trickiest part. Axiom 2 immediately reduced this to the special case $\lambda = \mu = 1$. For notational convenience, let $Y + Z$ by in the last slot. We have to prove

$$D(X_1, X_2, \ldots, Y + Z) = D(X_1, X_2, \ldots, Y) + D(X_1, X_2, \ldots, Z).$$

If $X_1, X_2, \ldots, X_{n-1}$ (which appear in all three terms above) are linearly dependent, we are done by part 3. Thus assume they are linearly independent. Since our linear space $\mathbb{R}^n$ has dimension n, these $n-1$ vectors can be extended to a basis for $\mathbb{R}^n$ by adding one more, $\tilde{X}_n$. Now we can write Y and Z as a linear combination of these basis vectors

$$Y = a_1X_1 + \cdots + a_{n-1}X_{n-1} + a_n\tilde{X}_n, \quad Z = b_1X_1 + \cdots + b_{n-1}X_{n-1} + b_n\tilde{X}_n.$$

Substituting this into D we obtain

$$D(X_1, \ldots, Y + Z) = D(X_1, \ldots, \ldots, \quad \sum_{1}^{n-1}(a_j + b_j)X_j + (a_n + b_n)\tilde{X}_n).$$

But by part 1,

$$= D(X_1, \ldots, (a_n + b_n)\tilde{X}_n)$$

and axiom 1 results in

$$= (a_n + b_n)D(X_1, \ldots, \tilde{X}_n).$$

However, again by part 1,

$$\begin{aligned} D(X_1, \ldots, Y) &= D(X_1, \ldots, \sum_{1}^{n-1} a_jX_j + a_n\tilde{X}_n) \\ &= D(X_1, \ldots, \ldots, a_n\tilde{X}_n) = a_nD(X_1, \ldots, \tilde{X}_n). \end{aligned}$$

Similarly

$$D(X_1, \ldots, Z) = b_nD(X_1, \ldots, \tilde{X}_n).$$

Adding these two expressions and comparing them with the above, we obtain the result.

(5) To avoid a mess, indicate only the ith and jth vectors. Our task is to prove

$$D(\ldots, X_i, \ldots, X_j, \ldots) = -D(\ldots, X_j, \ldots, X_i, \ldots).$$

This is clever. Watch: By the multilinearity (part 4)

$$\begin{aligned} &D(\ldots, X_i + X_j, \ldots, X_i + X_j, \ldots) \\ &\quad = D(\ldots, X_i, \ldots, X_i, \ldots,) + \cdots + D(\ldots, X_i, \ldots, X_j, \ldots) \\ &\qquad + D(\ldots, X_j, \ldots, X_i, \ldots) + \cdots + D(\ldots, X_j, \ldots, X_j, \ldots). \end{aligned}$$

However part 2 states that the left side as well as the first and last terms on the right are zero. Thus

$$0 = D(\ldots, X_i, \ldots, X_j, \ldots) + D(\ldots, X_j, \ldots, X_i, \ldots).$$

Transposition of one of the terms to the other side of the equality sign completes the proof. You should also be able to fashion an easy proof of this part which uses only the axioms directly (and uses none of the other parts of this theorem).

Instead of moving on immediately, it is instructive to compute $D[X_1, X_2]$ where X_1 and X_2 are vectors in $\mathbb{R}^2$, $X_1 = (a, b)$, $X_2 = (c, d)$. Then we are computing

$$D\left[\begin{pmatrix} a \\ b \end{pmatrix}, \begin{pmatrix} c \\ d \end{pmatrix}\right],$$

which is, equivalently, the determinant of the matrix $\begin{pmatrix} a & c \\ b & d \end{pmatrix}$.

$$\begin{aligned}
D\left[\begin{pmatrix} a \\ b \end{pmatrix}, \begin{pmatrix} c \\ d \end{pmatrix}\right] &= a\,D\left[\begin{pmatrix} 1 \\ \frac{b}{a} \end{pmatrix}, \begin{pmatrix} c \\ d \end{pmatrix}\right] && \text{(axiom 2)} \\
&= a\,D\left[\begin{pmatrix} 1 \\ \frac{b}{a} \end{pmatrix}, \begin{pmatrix} c \\ d \end{pmatrix} - c\begin{pmatrix} 1 \\ \frac{b}{a} \end{pmatrix}\right] && \text{(Theorem 21 part 1)} \\
&= a\,D\left[\begin{pmatrix} 1 \\ \frac{b}{a} \end{pmatrix}, \begin{pmatrix} 0 \\ \frac{ad-cb}{a} \end{pmatrix}\right] && \text{(algebra)} \\
&= (ad-bc)\,D\left[\begin{pmatrix} 1 \\ \frac{b}{a} \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right] && \text{(axiom 2)} \\
&= (ad-bc)\,D\left[\begin{pmatrix} 1 \\ \frac{b}{a} \end{pmatrix} - \frac{b}{a}\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right] && \text{(Theorem 21 part 1)} \\
&= (ad-bc)\,D\left[\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right] && \text{(algebra)} \\
&= (ad-bc)\,D\,[e_1, e_2] = ad - bc && \text{(axiom 4).}
\end{aligned}$$

Thus $|\text{Area}| = |(a+c)(b+d) - 2bc - cd - ab| = |ad - bc|$

You can indulge in a bit of analytic geometry (or look at my figure) to show that the area of a parallelogram spanned by X_1 and X_2 is $|ad - bc|$. From our explicit calculation, the existence and uniqueness of the determinant of two vectors in $\mathbb{R}^2$ has been proved.

There are several ways to prove the general existence and uniqueness of a determinant function. Our procedure is to first prove there is at most one determinant function (uniqueness). Then we shall define a function inductively, and verify it satisfies the axioms. By uniqueness, it must be the only function. Two interesting and important preliminary propositions are needed.

The following lemma shows how to evaluate the determinant if all of the elements above the principal diagonal are zeroes (that is, the determinant of a *lower triangular* matrix).

LEMMA: Let $X_1, \cdots, X_n$ be the columns of a lower triangular matrix

$$\begin{pmatrix} a_{11} & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & 0 & & \cdot \\ \cdot & \cdot & & & 0 \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ a_{n1} & a_{n2} & \cdots & \cdots & ann \end{pmatrix}$$

Then

$$D(X_1, \cdots, X_n) = a_{11}a_{22}\cdots a_{nn}D(e_1, \cdots, e_n) = a_{11}a_{22}\cdots a_{nn},$$

that is, the determinant of a triangular matrix is the product of the diagonal elements.

PROOF: If any one of the principal diagonal elements are zero, then the determinant is zero. For example, if $a_{jj} = 0$, then the $n - j + 1$ vectors $X_j, \cdots, X_n$ all have their first j components zero, and hence can span at most an $n - j$ dimensional space. Since $n - j + 1 > n - j$, these vectors must be linearly dependent. Therefore, by Theorem 21, part 3, the determinant is zero, as the theorem asserts. [If you didn't follow this, look at a 3×3 or 4×4 lower triangular matrix and think for a moment].

If none of the diagonal elements are zero, we can carry out the following simple recipe. The recipe gives a procedure for reducing the problem to evaluating a matrix which is zero everywhere except along the diagonal.

First, we get all zeros to the left of a_{22} in the second row by multiplying the second column, X_2, by $-a_{21}/a_{22}$ and adding the resulting vector to X_1. This gives a new first column with $i = 2, \quad j = 1$ element zero. Moreover, the new matrix has the same determinant as the old one (Theorem 21, part 1). It looks like

$$\begin{pmatrix} a_{11} & 0 & 0 & 0 \\ 0 & a_{22} & 0 & \cdot \\ \tilde{a}_{31} & a_{32} & a_{33} & \cdot \\ \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \tilde{a}_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}.$$

Only the first column has changed. Repeat the same process to get all zeros to the left of a_{33}. Thus, multiply the third column by $-\tilde{a}_{31}/a_{33}$ and $-a_{31}/a_{33}$ and add the result to the first and second columns respectively. This gives a new matrix, again with equal determinant, but which looks like

$$\begin{pmatrix} a_{11} & 0 & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & 0 & & 0 \\ 0 & 0 & a_{33} & 0 & & \cdot \\ \hat{a}_{41} & \hat{a}_{42} & a_{43} & a_{44} & & \cdot \\ \cdot & & \cdot & & & 0 \\ \cdot & & \cdot & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \hat{a}_{n1} & \hat{a}_{n2} & a_{n3} & \cdot & \cdot & a_{nn} \end{pmatrix}.$$

Moving on, we gradually eliminate all of the terms to the left of the diagonal but keep the *same* diagonal ones. The final result is

$$\begin{pmatrix} a_{11} & 0 & \cdots 0 & \\ 0 & a_{22} & & \cdot \\ \cdot & \cdot & & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & a_{nn} \end{pmatrix}.$$

It has the same determinant as the original matrix, so

$$\begin{aligned} D(X_1, \dots, X_n) &= D(a_{11}e_1, \dots, a_{nn}e_n) \\ &= a_{11}\cdots a_{nn}D(e_1, \dots, e_n), \end{aligned}$$

where Axiom 2 has been used to pull out the constants. Now Axiom 4, $D(e_1, \dots, e_n) = 1$, can be used to complete the proof. Observe that Axiom 4 is not used until the very last step. Thus, the formula $D =$ (something) $D(e_1, \dots, e_n)$ depends only on Axioms 1-3. We shall need this soon.

The above theorem shows how easy it is to evaluate the determinant of a lower triangular matrix. It becomes particularly valuable when coupled with the next theorem which shows how the determinant of an arbitrary matrix can be reduced to that of a lower triangular matrix. The reduction procedure given here is the *best practical way of evaluating a determinant.* There is a peculiar criss-cross method for evaluating 3×3 determinants which is taught in many high schools. Forget it. The method is not very practical and does not generalize to 4×4 or larger determinants.

Theorem 5.24 . *The evaluation of the determinant $D(X_1, \dots, X_n)$ can be reduced to the evaluation of a lower triangular matrix - and hence has the form*

$$D = (\textit{something})D(e_1, \dots, e_n).$$

The proof gives a way of computing "something" in terms of the original matrix.

Remark: In the above formula, we did not utilize the fact that

$$D(e_1, \cdots, e_n) = 1$$

since this one step in the proof is the only place where Axiom 4 would be used, so we can (and shall) use the fact that this result holds for any function which only satisfies Axioms 1-3.

PROOF: This is just a recipe for carrying out the reduction. It essentially is a repetition of the last part of the preceding lemma. Instead of waving our hands at the procedure, we shall work out a representative

EXAMPLE: Evaluate

$$D = D(X_1, X_2, X_3, X_4) = \begin{vmatrix} 1 & 2 & -1 & 0 \\ -1 & -2 & 3 & 1 \\ 0 & -1 & 4 & -3 \\ 2 & 5 & 0 & 1 \end{vmatrix}$$

by reducing it to a lower triangular determinant.

First we get all zeros to the right of the diagonal in the first row, that is, except in the a_{11} slot, by multiplying X_1 by the constants $-2, 1$ and 0 and adding the resulting vectors to X_2, X_3, and X_4, respectively. We obtain

$$D = \begin{vmatrix} 1 & 2 & -1 & 0 \\ -1 & -2 & 3 & 1 \\ 0 & -1 & 4 & -3 \\ 2 & 5 & 0 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0 & -1 & 0 \\ -1 & 0 & 3 & 1 \\ 0 & -1 & 4 & -3 \\ 2 & 1 & 0 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 2 & 1 \\ 0 & -1 & 4 & -3 \\ 2 & 1 & 2 & 1 \end{vmatrix}$$

Now we get all zeros to the right of the diagonal in the second row. Since the new a_{22} element above is zero, interchange the second and third columns (one could have interchanged the second and fourth). This introduces a factor of -1 (by Theorem 21, part 5). Then multiply the new second column by the constants 0 and $-\frac{1}{2}$, respectively, and add to the last two columns, respectively. This gives

$$D = \begin{vmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 2 & 1 \\ 0 & -1 & 4 & -3 \\ 2 & 1 & 2 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ -1 & 2 & 0 & 1 \\ 0 & 4 & -1 & -3 \\ 2 & 2 & 1 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 \\ 0 & 4 & -1 & -5 \\ 2 & 2 & 1 & 0 \end{vmatrix}$$

And on the third row, where we again want all zeros to the right of the diagonal, so multiply the new third column by -5 and add it to the fourth column:

$$D = -\begin{vmatrix} 1 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 \\ 0 & 4 & -1 & -5 \\ 2 & 2 & 1 & 0 \end{vmatrix} = -\begin{vmatrix} 1 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 \\ 0 & 4 & -1 & 0 \\ 2 & 2 & 1 & -5 \end{vmatrix} = -(1)(2)(-1)(-5) = -10,$$

where we have used the lemma about determinants of lower triangular matrices to evaluate the last determinant.

Uniqueness is now elementary.

Theorem 5.25 . *There is at most one function*

$$D(X_1, \cdots, X_n), \qquad X_k \in \mathbb{R}^n,$$

which satisfies the 4 axioms for a determinant function.

PROOF: Assume there are two such functions,

$$D(X_1, \cdots, X_n)$$

and

$$\tilde{D}(X_1, \cdots, X_n).$$

Let

$$\Delta(X_1, \cdots, X_n) = D(X_1, \cdots, X_n) - \tilde{D}(X_1, \cdots, X_n).$$

We shall show $\Delta(X_1, \cdots, X_n) = 0$ for any choice of $X_1, \cdots, X_n$. Since both D and $\tilde{D}$ satisfy Axioms 1-4, we have

1). $\Delta = D - \tilde{D}$ is real valued.

2). $\Delta(\ldots, \lambda X_j, \ldots) = D(\ldots, \lambda X_j, \ldots) - \tilde{D}(\ldots, \lambda X_j, \ldots)$

$$= \lambda D(\ldots, X_j, \ldots) - \lambda \tilde{D}(\ldots, X_j, \ldots)$$

$$= \lambda \Delta(\ldots, X_j, \ldots).$$

3). $\Delta(\ldots, X_j + X_k, \ldots) = D(\ldots, X_j + X_k, \ldots) - \tilde{D}(\ldots, X_j + X_k, \ldots)$

$$= D(\ldots, X_j, \ldots - \tilde{D}(\ldots, X_j, \ldots)$$

$$= \Delta(\ldots, X_j, \ldots), \quad j \neq k.$$

4). $\Delta(e_1, \ldots, e_n) = D(e_1, \ldots, e_n) - \tilde{D}(e_1, \ldots, e_n) = 1 - 1 = 0$. Thus, Δ satisfies the same first three axioms but $\Delta(e_1, \ldots, e_n) = 0$ in place of Axiom 4. Because the proof of Theorem 22 and its predecessors never used Axiom 4, we know that

$$\Delta(X_1, \ldots, X_n) = (\text{something}) \quad \Delta(e_1, \ldots, e_n) = 0.$$

Thus $\Delta(X_1, \cdots, X_n) = 0$ for any vectors X_j.

If it exists, the determinant function is known to be unique. We intend to *define* the determinant of order n, that is, of n vectors in $\mathbb{R}^n$, in terms of determinants of order $n-1$. The key to such an approach is a relationship between a determinant of order n and determinants of order $n-1$. To motivate our definition, we first examine the case $n = 3$ and utilize the intimate relation between determinant and volume.

Let X_1, X_2 and X_3 be three vectors in $\mathbb{R}^3$. To find the determinant $D(X_1, X_2, X_3)$, we can resolve one of the vectors, say X_1, into its components $X_1 = a_{11}e_1 + a_{21}e_2 + a_{31}e_3$. Since the determinant function is linear (Theorem 21, part 4),

$$D = D(X_1, X_2, X_3) = a_{11}D(e_1, X_2, X_3) + a_{21}D(e_2, X_2, X_3) + a_{31}D(e_3, X_2, X_3).$$

How can we interpret $D(a_{11}e_1, X_2, X_3)$,

$$D(a_{11}e_1, X_2, X_3) = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{vmatrix}?$$

By subtracting suitable multiples of the first column from the other two, we have

$$D(a_{11}e_1, X_2, X_3) = \begin{vmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{vmatrix}.$$

Consider the related volume function. The vectors in the last matrix span a parallelepiped whose base is the parallelogram spanned by $(0, a_{22}, a_{32})$ and $(0, a_{23}, a_{33})$, while the height is a_{11}. Thus, we expect the volume to be a_{11} times the area of the base. Since the area of the base is $\left|\det \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}\right|$, we hope

$$\begin{vmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}.$$

except possibly for a factor of ± 1. This last formula is the connection between determinants of order three and those of order two.

Notice that the determinant on the right in the last equation is obtained from that of $D = D(X_1, X_2, X_3)$ by deleting both the first row and first column. It is called the 1,1 *minor* of D, and written D_{11}. More generally, the i, j minor D_{ij} of D is the determinant obtained by deleting the ith row and jth column of D. If D is of order n, then each D_{ij} is of order $n-1$.

In this notation, we expect from the expansion of $D(X_1, X_2, X_3)$ that

$$D(X_1, X_2, X_3) = \pm ?a_{11}D_{11} \pm ?a_{21}D_{21} \pm ?_{31}D_{31},$$

or

$$\begin{vmatrix} a_{11} & a_{21} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \pm ?a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} \pm ?a_{21} \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} \pm ?a_{31} \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix}.$$

where ? indicates our doubt as to the signs. Explicit evaluation of both sides (using Theorem 22) reveals that the correct sign pattern is $+, -, +$.

Having examined this special case (and the 4×4 case too), we are tentatively led to

Suspicion (Expansion by Minors). If $D(X_1, \ldots, X_n)$ is a determinant function, that is, if it satisfies the axioms, then

$$D(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} D_{ij}, \tag{5-3}$$

where $X_j = (a_{1j}, a_{2j}, \ldots, a_{nj})$.

For the case $n = 3, j = 1$ this is the formula we found above. To verify that the formula is correct, we must verify that the function satisfies our axioms for a determinant. The reasoning goes as follows: we know exactly what determinants of order two are by a previous computation, so the formula gives a candidate for the determinant of order three, which in turn gives a candidate for a determinant function of order four, and so on. Thus, by induction, let us assume that determinants of order $k-1$ are known. We must prove

Theorem 5.26 . *The previous function $D(X_1, \ldots, X_k)$ defined by the above formula is a determinant function, that is, it satisfies the axioms.*

PROOF: 1). $D(X_1, \ldots, X_k)$ is real valued since, by our induction hypothesis, each of the D_{ij}, determinants of order $k-1$, is real valued.

2). $D(\ldots, \lambda X_l, \ldots) = \lambda D(\ldots, X_l, \ldots)$. There are two cases. *If* $l = j$, then λX_j means that $a_{1j}, a_{2j}, \ldots$, is multiplied by λ. Thus

$$D(\ldots, \lambda X_j, \ldots) = \sum_{i=1}^{n} (-1)^{i+j} \lambda a_{ij} D_{ij} = \lambda D(\ldots, X_j, \ldots),$$

so the axiom is satisfied. *If* $l \neq j$, then some vector X other than X_j is multiplied by λ, so

$$D(\ldots, \lambda X_l, \ldots, X_j, \ldots) = \begin{vmatrix} \cdots & \lambda a_{1l} & \cdots & a_{1j} & \cdots \\ \cdots & \lambda a_{2l} & & a_{2j} & \cdots \\ & \cdot & & \cdot & \\ & \cdot & & \cdot & \\ & \cdot & & \cdot & \\ \cdots & \lambda a_{kl} & & a_{kj} & \cdots \end{vmatrix}$$

Since D_{ij} is formed by deleting the ith row and jth column of D, and $l \neq j$, one column in minor Dij will have the factor λ appearing in it. By the induction hypothesis, the factor can be pulled out of each one, and hence from any linear combination of them. Because the expansion formula for D is a linear combination of the minors, the axiom is verified in this case too.

3). Omitted. This one is just plain messy. If you don't care to try the general case for yourself, at least try the case $n = 3$ and verify it there.

4). To prove $D(e_1, \ldots, e_n) = 1$. Of the coefficients $a_{1j}, a_{2j}, \ldots, a_{nj}$, only $a_{jj} \neq 0$, and $a_{jj} = 1$. Thus $D(e_1, \ldots, e_n) = (-1)^{j+j} a_{jj} D_{jj} = D_{jj}$. But by the induction hypothesis, $D_{jj} = 1$ since it has only ones on its main diagonal and zero elsewhere. Therefore $D(e_1, \ldots, e_n) = 1$, as desired.

This theorem completes (except for one segment) the proof that a unique determinant function exists. The uniqueness was proved directly, while the existence was obtained from the known existence of 2×2 determinant functions (the simpler case of 1×1 determinants could also have been used) and proving inductively that a candidate for the $n \times n$ determinant function does satisfy the axioms.

Emerging from the jungle of the existence proof, we are fully equipped with the powerful determinant function and the associated volume function. It will be relatively simple to prove the remaining theorems involving determinants. The trick in most of them is to make clever use of the fact that the determinant function is unique. We shall expose this trick in its bare form.

Theorem 5.27 . *Let $\Delta(X_1, \ldots, X_n)$ be a function of n vectors in $\mathbb{R}^n$ which satisfies axioms 1-3 for the determinant. Then for every set of vectors $X_1, \ldots, X_n$*

$$\Delta(X_1, \ldots, X_n) = \Delta(e_1, \ldots, e_n) D(X_1, \ldots, X_n).$$

Thus, the function Δ differs from D only by a constant multiplicative factor, which is the number Δ assigns to the unit matrix (geometrically, the unit cube) in $\mathbb{R}^n$.

PROOF: If $\Delta(e_1, \ldots, e_n) = 1$, then Δ satisfies Axiom 4 also, so by the uniqueness theorem, it must be D itself. If $\Delta(e_1, \ldots, e_n) \neq 1$, consider

$$\tilde{D}(X_1, \ldots, X_n) := \frac{D(X_1, \ldots, X_n) = \Delta(X_1, \ldots, X_n)}{1 - \Delta(e_1, \ldots, e_n)}.$$

Note that the denominator is a fixed scalar which does not depend on $X_1, \ldots, X_n$. It is a mental calculation to verify that $\tilde{D}$ satisfies all of Axioms 1-4. Therefore $\tilde{D}(X_1, \ldots, X_n) := D(X_1, \ldots, X_n)$ by uniqueness. Solving the last equation for $\Delta(X_1, \ldots, X_n)$ yields the formula.

Consider $D(X_1, \ldots, X_n)$. If $B = ((b_{ij}))$ is a square $n \times n$ matrix representing a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$, how are $D(X_1, \ldots, X_n)$ and $D(BX_1, BX_2, \ldots, BX_n)$ related? The answer to this question is vital if we are to find how volume varies under a linear transformation B. If $A = ((a_{ij}))$ is the matrix whose columns are $X_1, \ldots, X_n$, and $C = ((c_{ij}))$ is the matrix whose columns are $BX_1, BX_2, \ldots, BX_n$, then $C = BA$ [since, for example, c_{11}—the first element in the vector BX_1—is

$$c_{11} = b_{11}a_{11} + b_{12}a_{21} + b_{13}a_{31} + \cdots + b_{1n}a_{n1}.]$$

Because $D(X_1, \ldots, X_n) = \det A$ and $D(BX_1, \ldots, BX_n) = \det C$, our question becomes one of relating $\det C = \det(BA)$ to $\det A$. The result is as simple as one could possibly expect.

Theorem 5.28 . *If A and B are two $n \times n$ matrices, then*

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

or, if $X_1, \ldots, X_n$ are the column vectors of A, then this is equivalent to

$$D(BX_1, \ldots, BX_n) = D(Be_1, Be_2, \ldots, Be_n) D(X_1, \ldots, X_n)$$

(since the matrix whose columns are $Be_1, \ldots, Be_n$ is just B).

PROOF: Let $\Delta(X_1, \ldots, X_n) := D(BX_1, \ldots, BX_n)$. This function clearly satisfies Axiom 1. We shall verify Axioms 2 and 3 at the same time.

$$\Delta(\ldots, \lambda X_j + \mu X_k, \ldots) = D(\ldots, B(\lambda X_n + \mu X_k), \ldots)$$

Because B is a linear transformation, we have

$$= D(\ldots, \lambda BX_j + \mu BX_k, \ldots).$$

By the linearity of D (Theorem 21, part 4)

$$= \lambda D(\ldots, BX_j, \ldots) + \mu D(\ldots, BX_k, \ldots).$$

If $j \neq k$, then the vector BX_k in the second term on the right also appears as another column in the same determinant. Hence the second term vanishes. Thus if $j \neq k$,

$$\Delta(\ldots, \lambda X_j + \mu X_k, \ldots) = \lambda D(\ldots, BX_j, \ldots).$$

The special case $\mu = 0$ shows Axiom 2 holds for Δ, while the case $\lambda = \mu = 1$ verifies Axiom 3. Therefore Δ satisfies Axioms 1-3. Applying the preceding Theorem (25), we have

$$\Delta(X_1, \ldots, X_n) = \Delta(e_1, \ldots, e_n) D(X_1, \ldots, X_n).$$

By definition, $\Delta(e_1, \ldots, e_n) := D(Be_1, \ldots, Be_n)$. Substitution verifies our formula. The commutativity

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

follows from the fact that $\det A$ and $\det B$ are real numbers - which do commute under multiplications.

Corollary 5.29 . *If A is an invertible matrix, then*

$$\det(A^{-1}) = \frac{1}{\det A}.$$

PROOF: Since $AA^{-1} = I$, and $\det I = 1$, we find

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

Ordinary division completes the proof.

Our next theorem is also a corollary, but because of its importance, we call it

Theorem 5.30 . *The vectors $X_1, \ldots, X_n$ in $\mathbb{R}^n$ are linearly independent if and only if $D(X_1, \ldots, X_n) \neq 0$.*

PROOF: $\Leftarrow$ If $D(X_1, \ldots, X_n) \neq 0$, then the vectors $X_1, \ldots, X_n$ are linearly independent, since if they were dependent, then $D = 0$ by part 3 of Theorem 21.

$\Rightarrow$. If $X_1, \ldots, X_n$ are linearly independent vectors in $\mathbb{R}^n$, then the Corollary to Theorem 12 (p. 364) shows that the matrix A whose columns are the X_j is invertible. Let A^{-1} be its inverse. From the computation in the corollary preceding this theorem,

$$(\det A)(\det A^{-1}) = 1.$$

Thus the real number $\det A$ cannot be zero. The equivalent form of our theorem is also a consequence of the Corollary to Theorem 12.

EXAMPLE: (cf. p. 157, Ex. 1b). Are the vectors

$$X_1 = (0,1,1), \quad X_2 = (0,0,-1), \quad X_3 = (0,2,3)$$

linearly dependent? We compute the determinant

$$D(X_1, X_2, X_3) = \begin{vmatrix} 0 & 0 & 0 \\ 1 & 0 & 2 \\ 1 & -1 & 3 \end{vmatrix}.$$

If we knew that "the determinant of a matrix was equal to the determinant of its adjoint" (a true theorem to be proved below), then taking the adjoint we get a matrix with one column zero 0 which gives $D = 0$. Since the quoted theorem is not yet proved, we proceed differently and reduce our 3×3 determinant to 2×2 determinants expanding by minors (p. 411). The simplest column to use is the second.

$$\begin{vmatrix} 0 & 0 & 0 \\ 1 & 0 & 2 \\ 1 & -1 & 3 \end{vmatrix} = (-1)^{1+2}0 \begin{vmatrix} 1 & 2 \\ 1 & 3 \end{vmatrix} + (-1)^{2+2}0 \begin{vmatrix} 0 & 0 \\ 1 & 3 \end{vmatrix} + (-1)^{3+2}(-1) \begin{vmatrix} 0 & 0 \\ 1 & 2 \end{vmatrix}$$
$$= \begin{vmatrix} 0 & 0 \\ 1 & 2 \end{vmatrix} = 0 \cdot 2 - 1 \cdot 0 = 0$$

by the explicit formula for evaluating 2×2 determinants. Thus $D = 0$ so the vectors X_1, X_2, X_3 are linearly dependent.

That nice theorem we could have used in the above example is our next target.

Theorem 5.31 . *If A is an $n \times n$ matrix, then*

$$\det A^* = \det A.$$

PROOF: Let $\mathcal{A}_1, \ldots, \mathcal{A}_n$ be the columns of A and $\mathcal{B}, \ldots, \mathcal{B}_n$ its rows,

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & \cdots & \cdots & a_{2n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ a_{n1} & \cdots & \cdots & a_{nn} \end{pmatrix} \begin{matrix} \mathcal{B}_1 \\ \} \; \mathcal{B}_2 \\ \cdot \\ \cdot \\ \cdot \\ \} \; \mathcal{B}_n \end{matrix}.$$

Consider the function

$$D(\mathcal{B}_1, \ldots, \mathcal{B}_n) = \begin{vmatrix} a_{11} & \cdots & a_{n1} \\ a_{12} & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{1n} & & a_{nn} \end{vmatrix} \begin{matrix} \} \; \mathcal{A}_1 \\ \cdot \\ \cdot \\ \cdot \\ \} \; \mathcal{A}_n \end{matrix} = \det A^*.$$

since the rows of A are the columns of A^*. Let us *define* a new function

$$\hat{D}(\mathcal{A}_1, \ldots, \mathcal{A}_n) := D(\mathcal{B}_1, \ldots, \mathcal{B}_n).$$

Our task is to verify that $\hat{D}(\mathcal{A}_1, \ldots, \mathcal{A}_n)$ satisfies all of Axioms 1-4. Then by uniqueness

$$\det A^* := \hat{D}(\mathcal{A}_1, \ldots, \mathcal{A}_n) = D(\mathcal{A}_1, \ldots, \mathcal{A}_n) = \det A.$$

(1) $\hat{D}(\mathcal{A}_1, \ldots, \mathcal{A}_n)$ is a real number since $\det A^*$, the determinant of the matrix A^* is a real number.

(2) We must show $\hat{D}(\ldots, \lambda\mathcal{A}_j, \ldots) = \lambda\hat{D}(\ldots, \mathcal{A}_j, \ldots)$, that is,

$$\begin{vmatrix} a_{11} & a_{2j} & \cdots & a_{n1} \\ \cdot & & & \\ \cdot & & & \\ \lambda a_{1j} & \lambda a_{2j} & \ldots & \lambda a_{nj} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ a_{1n} & \ldots & \ldots & a_{nn} \end{vmatrix} = \lambda \begin{vmatrix} a_{11} & \cdots & a_{nl} \\ \cdot & & \\ \cdot & & \\ a_{1j} & \cdots & a_{nj} \\ & & \\ & & \\ a_{1n} & \cdots & a_{nn} \end{vmatrix}$$

(a fact we only know so far if a *column* is multiplied by a scalar). Trick: observe that

$$j\text{th row} \ldots \begin{pmatrix} 1 & 0 & 0 & & 0 \\ 0 & 1 & & & 0 \\ 0 & \cdots & \lambda & 1 & 0 \\ 0 & & 0\ldots & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} & \cdots & a_{nl} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{1j} & \cdots & a_{nj} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{1n} & \cdots & a_{nn} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11} & \cdots & a_{nl} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{1j} & \cdots & a_{nj} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{1n} & \cdots & a_{nn} \end{pmatrix}$$

The matrix on the left is the identity matrix I except for a λ in its jth row and jth column. Its determinant is λ (since you can factor λ from the jth column and are left with the identity matrix). By Theorem 26, the determinant of the product on the left is $\lambda\hat{D}(\mathcal{A}_1, \ldots, \mathcal{A}_n)$ while the right is $\hat{D}(\mathcal{A}_1, \ldots, \mathcal{A}_n)$, proving $\hat{D}$ satisfies Axiom 2.

(3) The proof of Axiom 3 involves a similar trick. We have to show $\hat{D}(\ldots, \mathcal{A}_j + \mathcal{A}_k, \ldots) = \hat{D}(\ldots, \mathcal{A}_j, \ldots)$ where $j \neq k$, that is, to show

$$\begin{vmatrix} a_{11} & \cdots & a_{n1} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{1j} + a_{1k} & \cdots & a_{nj} + a_{nk} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{1n} & \cdots & a_{nn} \end{vmatrix} = \begin{vmatrix} a_{11} & \cdots & a_{n1} \\ \cdot & & \\ \cdot & & \\ a_{1j} & \cdots & a_{nj} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{1n} & \cdots & a_{nn} \end{vmatrix}, j \neq k.$$

Observe that

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & \\ \cdots & \cdots & \cdots & \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & & 1 \end{pmatrix} \begin{pmatrix} a_{11} & \cdots & a_{n1} \\ \cdot & & \\ \cdot & & \\ a_{1j} & \cdots & a_{nj} \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ a_{1n} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{n1} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{1j} + a_{1k} & \cdots & a_{nj} + a_{nk} \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ a_{1n} & \cdots & a_{nn} \end{pmatrix},$$

where the matrix on the left is the identity matrix with an extra 1 in the jth row, kth column. Since the determinant of this matrix is one (check by a mental computation), the rule for the determinant of a product of matrices shows that Axiom 3 is satisfied.

(4) Easy, for

$$\hat{D}(e_1, \ldots, e_n) = \begin{vmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & 0 \\ & \cdots & \cdots & 1 & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{vmatrix} = D(e_1, \ldots, e_n) = 1.$$

This verification of the four Axioms coupled with the remarks at the beginning of the proof completes the proof.

Corollary 5.32 . *The column operations of Theorem 21 are also valid as row operations.*

PROOF: Every row operation on a matrix A (like adding two rows) can be split up to : i) take A^* so the rows become columns, ii) carry out the operation on the column of A^* and iii) take the adjoint again. Since the determinant does not change under these operations, we are done.

Corollary 5.33 . *If R is an orthogonal matrix then*

$$\det R = \pm 1.$$

PROOF: If R is orthogonal, then $R^*R = I$ by Theorem 19 (p. 383). Thus,

$$a = \det I = \det(R^*R) = (\det R^*)(\det R) = (\det R)^2,$$

where Theorems 25 and 27 were invoked once each. Now take the square root of both sides.

The orthogonal matrices

$$R_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad R_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

for which $\det R_1 = 1$ and $\det R_2 = -1$ show that both signs are possible. *If* $\det R = -1$, then the orthogonal transformation has not only been a rotation but also a *reflection.* The transformation given by R_2 is

A FIGURE GOES HERE

which can be thought of as the composition (product) of a rotation by $+90^0$ followed by a reflection (mirror image). In fact, R_2 may be factored into $\hat{R}\tilde{R} = r_2$, where

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = R_2.$$

Pictorially

A FIGURE GOES HERE

Our theorems about determinants also imply the following valuable result about volume.

Theorem 5.34 . *Let $X_1, \ldots, X_n$ span a parallelepiped Q in $\mathbb{R}^n$ and the matrix A map $\mathbb{R}^n$ into $\mathbb{R}^n$. Then the volume is magnified by $|\det A|$, that is,*

$$V[AX_1, \ldots, AX_n] = |\det A| \; V[X_1, \ldots, X_n].$$

If we denote the image of Q by $A(Q)$, then this theorem reads

$$Vol[A(Q)] = |\det A| \quad Vol[Q].$$

PROOF: [We should first prove that there is at most one volume function V satisfying its four axioms. Since $V := |D|$ is a volume function, assume there is another volume function V^* and define $\tilde{D}(X_1, \ldots, X_n)$ by

$$\tilde{D}(X_1, \ldots, X_n) := \begin{cases} \frac{V^*(X_1,\ldots,X_n)D(X_1,\ldots,X_n)}{|D(X_1,\ldots,X_n)|} & \text{if} \quad D \neq 0 \\ 0 & \text{if} \quad D = 0. \end{cases}$$

It is simple to check that $\tilde{D}$ satisfies the axioms for a determinant. By uniqueness, $\tilde{D} = D$. Solving the last equation, we find $V^*(X_1, \ldots, X_n) = |D(X_, \ldots, X_n)| \equiv V(X_1, \ldots, X_n)$, so the volume function is also unique.]

The theorem is easily proved. Since $V = |D|$, an application of Theorem 26 tells us that

$$\begin{aligned} V[AX_1, \ldots, AX_n] &= |D(AX_1, \ldots, AX_n)| \\ &= |D(Ae_1, \ldots, Ae_n)| \quad |D(X_1, \ldots, X_n)| \\ &= |\det A| \quad V[X_1, \ldots, X_n]. \end{aligned}$$

Done.

Corollary 5.35 . *Volume is invariant under an orthogonal transformation.*

$$V(RQ) = V(Q)$$

PROOF: If R is an orthogonal transformation, $|\det R| = 1$.

Remark 1. Since we eventually want to define the volume of suitable sets by approximating the sets by parallelepipeds, this theorem will allow us to conclude the same results about how the volume of some set changes under a linear transformation in general and an orthogonal transformation in particular.

REMARK: 2 We *define* the determinant of a linear transformation L which maps $\mathbb{R}^n$ into $\mathbb{R}^n$ as the determinant of a matrix which represents L. This definition makes it mandatory to prove: "the determinant of two different matrices which represent L (different because of a different choice of bases) are equal." However the theorem is an immediate consequence of the following fact we never proved: "if A and B are matrices which represent the *same* linear transformation L with respect to different bases then there is a nonsingular matrix C such that $B = CAC^{-1}$." The matrix C is the matrix expressing one set of bases vectors in terms of the other bases. Using this theorem, we find

$$\det B = \det(CAC^{-1}) = (\det C)(\det A)(\det C^{-1}) = \det A.$$

How does volume change under a translation $T, TX = X + X_0$? A little thought is needed. Imagine a parallelepiped Q spanned by $X_1, \ldots, X_n$. The crux of the matter is to realize that the parallelepiped has the *origin* as one of its vertices and $X_1, \ldots, X_n$ at the others. Under the translation T, not only do the X_j's get translated through X_0, but *so does the origin*, $0 \to X_0, \quad X_1 \to X_1 + X_0, \quad X_2 \to X_2 + X_0$, etc.

A FIGURE GOES HERE

In terms of free vectors, the edge from 0 to X_j becomes the edge from X_0 to $X_j + X_0$ (see figure). Thus the free vector representing this edge is $(X_j + X_0) - X_0$, that is, it is still X_j! This motivates the
DEFINITION: The volume of a parallelepiped is defined to be the volume of the parallelepiped after translating one vertex to the origin.

Theorem 5.36 . *The change in volume of a parallelepiped Q under an affine transformation $AX = LX + X_0$, L linear, is given by:*

$$Vol[A(Q)] = |\det L| \; Vol[Q].$$

In particular, volume is invariant under a rigid body transformation (for then L is an orthogonal transformation).

PROOF: The affine transformation may be factored into $A = TL$, a linear transformation followed by a translation (p. 380). Since L changes volume by $|\det L|$ while translation preserves the volume, the net result is a change by $|\det L|$ as claimed.

a) Application to Linear Equations

What have our geometrically motivated determinants in common with the determinants of high school fame - where they were used to solve systems of linear algebraic equations? Everything, for they are the same. Since determinants are defined only for square matrices, they are applicable to linear algebraic equations only when there are the same number of equations as unknowns. At the end of this section, we shall make some remarks about the case when the number of equations and unknowns are not equal.

Consider the system of equations

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= y_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= y_2 \\ \vdots \qquad & \quad \vdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n &= y_n, \end{aligned}$$

which we can write as

$$x_1\mathcal{A}_1 + \cdots + x_n\mathcal{A}_n = Y,$$

where $\mathcal{A}_j$ is the jth column of the matrix $A = ((a_{ij}))$ and Y is the obvious column vector. The problem is to find numbers $x_1, \ldots, x_n$ such that $x_1\mathcal{A}_1 + \cdots + x_n\mathcal{A}_n = Y$, where Y is given.

Theorem 5.37 . *Let $A = ((a_{ij}))$ be a square $n \times n$ matrix and Y a given vector. The system of linear algebraic equations $AX = Y$ can always be solved for X if and only if* $\det A \neq 0$. *This can be rephrased as, A is invertible if and only if* $\det A \neq 0$.

PROOF: Let $\mathcal{A}_j$ be the jth vector of A. Each $\mathcal{A}_j$ is a vector in $\mathbb{R}^n$. If $\det A \neq 0$, then the $\mathcal{A}_n$'s are linearly independent by Theorem 27, p. 417. But since they are linearly independent and there are n of them, $\mathcal{A}_1, \cdots, \mathcal{A}_n$, they must span $\mathbb{R}^n$. Thus, any $Y \in \mathbb{R}^n$ can be written as a linear combination of the $\mathcal{A}_j$'s. The numbers $x_1, \cdots, x_n$ are just the coefficients in this linear combination.

Conversely, if the equations $AX = Y$ can be solved for *any* $Y \in \mathbb{R}^n$, then the vectors $\mathcal{A}_1, \cdots, \mathcal{A}_n$ span $\mathbb{R}^n$. But if n vectors span $\mathbb{R}^n$, these vectors must be linearly independent, so $\det A \neq 0$, again by Theorem 27, page 417.

Theorem 5.38 . *Let A be a square matrix. The system of homogeneous equations $AX = 0$ has a non-trivial solution if and only if $\det A = 0$.*

PROOF: By Theorem 27, Page 417, $\det A = 0$ if and only if the column vectors $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are linearly dependent. Now if the column vectors $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are linearly dependent, then there are numbers $x_1, \ldots, x_n$, not all zero, such that $x_1\mathcal{A}_1 + \ldots + x_n\mathcal{A}_n = 0$. The vector $X = (x_1, \ldots, x_n)$ is then a non-trivial solution of $AX = 0$. Conversely, if there is a non-trivial solution of $AX = 0$, then $x_1\mathcal{A}_1 + \cdots + x_n\mathcal{A}_n = 0$, so the $\mathcal{A}_j$'s are linearly dependent. Hence $\det A = 0$.

In contrast to the above theorems which give no hint of a procedure for finding the desired vector X, the next theorem gives an explicit formula for the solution of $AX = Y$.

Theorem 5.39 *(Cramer's Rule). Let $A = ((a_{ij}))$ be a square $n \times n$ matrix with columns $\mathcal{A}_1, \ldots, \mathcal{A}_n$. Assume $\det A \neq 0$. Then for any vector Y, the solution of $AX = Y$ is*

$$x_1 = \frac{D(Y, \mathcal{A}_2, \ldots, \mathcal{A}_n)}{D(\mathcal{A}_1, \ldots, \mathcal{A}_n)}, \qquad x_2 = \frac{D(\mathcal{A}_1, Y, \mathcal{A}_3, \ldots, \mathcal{A}_n)}{D(\mathcal{A}_1, \ldots, \mathcal{A}_n)}$$

$$\vdots$$

$$x_n = \frac{D(\mathcal{A}_1, \ldots, \mathcal{A}_{n-1}, Y)}{D(\mathcal{A}_1, \ldots, \mathcal{A}_n)}.$$

For example, in detail, the formula for x_2 is

$$x_2 = \frac{\begin{vmatrix} a_{11} & y_1 & a_{13} & \cdots & a_{1n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & y_n & a_{n3} & \cdots & a_{nn} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{vmatrix}}.$$

PROOF: A snap. Since $\det A \neq 0$, by Theorem 31 we know a solution $X = (x_1, \ldots, x_n)$ exists. Thus $x_1\mathcal{A}_1 + \cdots + x_n\mathcal{A}_n = Y$. Let us obtain the formula for x_2 as a representative case. Observe that

$$D(\mathcal{A}_1, Y, \mathcal{A}_3, \cdots, \mathcal{A}_n) = D(\mathcal{A}_1, x_1\mathcal{A}_1 + \cdots + x_n\mathcal{A}_n, \mathcal{A}_3, \cdots, \mathcal{A}_n).$$

Since D is multilinear, we can expand the above to

$$= x_1D(\mathcal{A}_1, \mathcal{A}_1, \mathcal{A}_3, \cdots, \mathcal{A}_n) + x_nD(\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \cdots, \mathcal{A}_n) + \cdots + x_nD(\mathcal{A}_1, \mathcal{A}_n, \mathcal{A}_3 \cdots \mathcal{A}_n).$$

Now all of these determinants, except the second one, vanishes since each has two identical columns (part 5 of Theorem 21, page 400). Thus

$$D(\mathcal{A}_1, Y, \mathcal{A}_3, \cdot, \mathcal{A}_n) = x_2D(\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_n).$$

Because $\det \mathcal{A} = D(\mathcal{A}_1, \cdots, \mathcal{A}_n) \neq 0$, we can divide to find the desired formula for x_2. Done.

REMARK: This elegant formula is mainly of theoretical use. It is not the most efficient procedure for solving such equations. That honor belongs to the method of reducing to triangular form which was outlined in the proof of Theorem 22. To be more vivid, if Cramer's rule were used to solve a system of 26 equations, approximately $(23+1)! \approx 10^{28}$ multiplications would be required. Reduction to triangular form, on the other hand, would only require about $(1/3)(23)^3 \approx 6000$ multiplications. Think about that.

For non-square matrices, determinants are not applicable. Given a vector Y, one would still like a criterion to determine if one can solve $AX = Y$, that is, one would like a criterion to see if $Y \in \mathcal{R}(A)$.

Theorem 5.40 . *Let $L\colon V_1 \to V_2$ be a linear operator. Then*

$$\mathcal{R}(L)^{\perp} = \mathcal{N}(L^*);$$

or equivalently (for finite dimensional spaces)

$$\mathcal{R}(L) = \mathcal{N}(L^*)^{\perp}.$$

PROOF: If $X \in V$, and $Y \in \mathcal{R}(L)^{\perp}$, then for all X

$$0 = \langle Y,\, LX \rangle = \langle L^*Y,\, X \rangle.$$

This means L^*Y is orthogonal to all X, consequently, $L^*Y = 0$, so $Y \in \mathcal{N}(L^*)$. The converse is proved by observing that our steps are reversible.
Application. For what vectors $Y = (y_1, y_2, y_3)$ can you solve the equations

$$\begin{aligned} 2x_1, +3x_2 &= y_1 \\ x_1 - x_2 &= y_2 \\ x_1 + 2x_2 &= y_3 \qquad ? \end{aligned}$$

If the equations are written as $AX = Y$, then by the above theorem $Y \in \mathcal{R}(A)$ if and only if $Y \perp \mathcal{N}(A^*)$. Let us find a basis for $\mathcal{N}(A^*)$. This means solving the homogeneous equations $A^*Z = 0$,

$$\begin{aligned} 2z_1 + z_2 + z_3 &= 0 \\ 3z_1 - z_2 + 2z_3 &= 0. \end{aligned}$$

If we let $z_1 = \alpha$, and solve the resulting equations for z_2 and z_3, we find that $z_3 = -5\alpha/3$ and $z_2 = -11\alpha/3$. Consequently, all vectors $Z \in (A^*)$ have the form $Z = (3\alpha, -11\alpha, -5\alpha)$. A basis for $\mathcal{N}(A^*)$ is $e = (3, -11, -5)$. Therefore, $Y \perp \mathcal{N}(A^*)$ if and only if $3y_1 - 11y_2 - 5y_3 = 0$. By the above reasoning, the equation $AX = Y$ can be solved for only these Y's.

REMARK: The use of Theorem 34 as a criterion for finding if $Y \in \mathcal{R}(L)$ is much more valuable in infinite dimensional spaces, for it quite often turns out that $\mathcal{N}(L^*)$ is still finite dimensional while $\mathcal{R}(L)$ is infinite dimensional. For more on these ideas, see page 389, Exercise 12 and page 501 Exercises 27- 29.

Exercises

(1) Evaluate the following determinants as you see fit:

a). $\begin{vmatrix} 7 & 3 \\ 2 & -1 \end{vmatrix}$, b). $\begin{vmatrix} \frac{1}{2} & 5 \\ -3 & 4 \end{vmatrix}$.

c). $\begin{vmatrix} -10 & -2 & 3 \\ -3 & 2 & 1 \\ 5 & 0 & -1 \end{vmatrix}$, d). $\begin{vmatrix} 53 & 17 & 29 \\ 36 & 12 & 39 \\ 69 & 23 & 75 \end{vmatrix}$, e). $\begin{vmatrix} 1 & 2 & 0 & 1 \\ 1 & 3 & 4 & 0 \\ 0 & 1 & -5 & 6 \\ 1 & 2 & 3 & 4 \end{vmatrix}$

f). $\begin{vmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{vmatrix}$, g). $\begin{vmatrix} a & 1 & 0 & 0 & 0 \\ b & 1 & 0 & 0 & 0 \\ c & 0 & 0 & 1 & -b \\ c & 0 & 0 & 1 & -a \\ d & e & 1 & f & g \end{vmatrix}$

[Answers: a) -13, b) 17, c) -14, d) 6, e) 5, g) $-(b-a)^2$].

(2) If A and B are the matrices whose respective determinants appear in #1 a) and b), compute $\det(AB)$ by first finding AB. Compare with $(\det A)(\det B)$.

(3) a). Use Cramer's rule (Theorem 33) to solve the equation $AX = Y$, where A is given below. Then observe you have computed A^{-1}, so exhibit it.

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & -3 & -1 \\ 4 & 9 & 1 \end{pmatrix}. \qquad [A^{-1} = \frac{1}{30}\begin{pmatrix} 6 & 8 & 2 \\ -6 & -3 & 3 \\ 30 & -5 & -5 \end{pmatrix}].$$

b). Use the formula for A^{-1} to solve the equations

$$AX = Y \quad \text{where} \quad Y = (1, 2, 0).$$

(4) a). Find the volume of the parallelepiped Q in $\mathbb{R}^3$ which is spanned by the vectors $X_1 = (1,1,1)$, $X_2 = (2,-1,-3)$ and $X_3 = (4,1,9)$. [Answer: Volume = 30].

b). The matrix A,

$$A = \begin{pmatrix} -10 & -2 & 3 \\ -3 & 2 & 1 \\ 5 & 0 & -1 \end{pmatrix} \qquad \text{– (cf. \#1,c)}$$

maps $\mathbb{R}^3$ into itself. Find the volume of the image of Q, that is, the volume of $A(Q)$. [Answer: 420].

(5) Let $B = A - \lambda I$ where A is a square matrix. The values λ for which B is singular are called the *eigenvalues* of A. Find the eigenvalues for

a). $A = \begin{pmatrix} 3 & 2 \\ 2 & -1 \end{pmatrix}$, b). $A = \begin{pmatrix} 3 & 2 \\ 1 & -1 \end{pmatrix}$.

c). $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$.

[Hint: If B is singular, then $0 = \det B = \det(A - \lambda I)$. Now observe that $\det(A - \lambda I)$ is a polynomial in λ. The answer to c) is $\lambda = \frac{1}{2}(a + d \pm \sqrt{(a+d)^2 - 4(ad - bc)})$].

(6) For what value(s) of α are the vectors

$$X_1 = (1,2,3), \quad X_2 = (2,0,1), \quad X_3 = (0,\alpha,-1)$$

linearly dependent?

(7) If X_1, X_2, X_3 and Y_1, Y_2, Y_3 are vectors in $\mathbb{R}^3$, prove that

$$\begin{aligned} &D[X_1, X_2, X_3] - D[Y_1, Y_2, Y_3] \\ &\qquad = D[X_1 - Y_1, X_2, X_3] + D[X_1, X_2 - Y_2, X_3] + D[X_1, X_2, X_3 - Y_3]. \end{aligned}$$

[Hint: First work out the corresponding formula for the 2×2 case.]

(8) Here you shall compute the derivative of a determinant if the coefficients of $A = ((a_{ij}))$ depend on $t, a_{ij}(t)$. Let $X_1(t), \ldots, X_n(t)$ be the vectors which constitute the columns of A. The problem is to compute

$$\frac{dD(t)}{dt} = \frac{d}{dt} D[X_1, \ldots, X_n](t) = \frac{d}{dt}\begin{vmatrix} a_{11}(t), & \cdots, & a_{1n}(t) \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ a_{n1}(t), & \cdots, & a_{nn}(t) \end{vmatrix}$$

a). Use Exercise 7 (generalized to $n \times n$ matrices) to show

$$\begin{aligned} D(t + \Delta t) - D(t) &\equiv D[X_1(t+\Delta t),\, X_2(t+\Delta t), \ldots] - D[X_1(t),\, X_2(t), \ldots] \\ &= \sum_{j=1}^{n} D[X_1(t), \ldots, X_{j-1}(t),\, X_j(t+\Delta t) - X_j(t), X_{j+1}(t+\Delta b), \ldots] \end{aligned}$$

[Hint: Do the cases $n=2$ and $n=3$ first].

b). Use part a to show that

$$\begin{aligned}\frac{dD}{dt} &= \lim_{\Delta t\to 0}\left[\frac{D(t+\Delta t)-D(t)}{\Delta t}\right]\\ &= \sum_{j=1}^{n} D[X_1,\dots,X_{j-1},\frac{dX_j}{dt},X_{j+1},\dots,X_n],\end{aligned}$$

so the derivative of a determinant is found by taking the derivative one column at a time and adding the result.

(9) Let $u_1(t)$, and $u_3(t)$ be solutions of the differential equation

$$u''+a_1(t)u'+a_0(t)u=0.$$

Consider the *Wronski* determinant

$$W(u_1,u_2)(t) := \begin{vmatrix} u_1(t) & u_2(t)\\ u_1'(t) & u_2'(t)\end{vmatrix}$$

(a) Use Exercise 8 to prove

$$\frac{dW}{dt} = -a_1(t)W.$$

(b) Consequently, show

$$W(t)=W(t_0)\exp\left\{-\int_{t_0}^{t} a_1(s)\,ds\right\}.$$

(c) Apply this to show that if the vectors $(u_1(t),u_1'(t))$ and $(u_2(t),u_2'(t))$ are linearly independent at $t=t_0$, then they are always linearly independent.

(d) Let $u_1(t)\dots,u_n(t)$ be solutions of the differential equation

$$u^{(n)}+a_{n-1}(t)u^{(n-1)}+\cdots+a_1(t)u'+a_0(t)u=0.$$

Consider the Wronski determinant of $u_1,\dots,u_n$

$$W(u_1,\dots,u_n)=\begin{vmatrix} u_1 & u_2 & \cdots & u_n\\ u_1' & u_2' & \cdots & u_n'\\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ u_1^{(n-1)} & a_2^{(n-)} & \dots & u_n^{(n-1)}\end{vmatrix}$$

Prove

$$\frac{dW}{dt}=-a_{n-1}(t)W,$$

so again

$$W(t)=W(t_0)\exp\left\{-\int_{t_0}^{t} a_{n-1}(s)\,ds\right\}.$$

(e) Use part d) to conclude that the n vectors

$$(u_1,u_1',\dots,u_1^{(n-1)}),(u_2,u_2',\dots,u_2^{(n-1)},),\cdots(u_n,u_n',\dots,u_n^{(n-1)})$$

(where the u_j are solutions of the O.D.E.) are linearly independent for all t if and only if they are so at $t=t_0$.

(10) A matrix A is *upper* (lower) *triangular* if all the elements below (above) the main diagonal are zero,

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_n\\ 0 & a_{22} & \cdots & \cdot\\ 0 & 0 & & \cdot\\ 0 & 0 & \cdots & a_{nn}\end{pmatrix}.$$

If A is upper (or lower) triangular, prove again that

$$\det A = a_{11}a_{22}\dots a_{nn}.$$

by expanding by minors. What is the relation of this result to the exercise (#4, p. 157) on echelon form?

(11) Let $X_1,\dots,X_n$ be vectors in $\mathbb{R}^n$ and let $\hat{D}(X_1,\dots,X_n)$ be a real valued function which has properties 1 and 4 of Theorem 21. Thus $\hat{D}$ is skew-symmetric, and is linear in each of its columns. Prove $\hat{D}$ necessarily satisfies Axioms 2 and 3 for the determinant, and conclude that

$$\hat{D}(X_1,\dots,X_n)=kD(X_1,\dots,X_n),$$

where the constant $k=D(e_1,\dots,e_n)$.

(12) Let $u_1(t),\dots,u_n(t)$ be sufficiently differentiable functions (C^{n-1} is enough). Define the Wronskian as in Exercise 9 part d. Prove that if the functions $u_1,\dots,u_n$ are linearly dependent, then $W(t)\equiv 0$. Thus, if $W(t_0)\neq 0$, the functions are linearly independent in any interval containing t_0. [Do *not* try to apply the result of Exercise 9 for it is not applicable].

(13) (a) If I is the $n\times n$ identity matrix, evaluate $\det(\lambda I)$ where λ is a constant.

(b) If A is an $n\times n$ matrix, prove

$$\det(\lambda A)=\lambda^n\det A.$$

(c) If A or B are $n\times n$ matrices, is

$$\det(A+B)\stackrel{?}{=}\det A+\det B?$$

Proof or counterexample.

(14) For what value of α does the system of equations

$$\begin{aligned} x + 2y + z &= 0 \\ -2x + \alpha y + 2z &= 0 \\ x + 2y + 3z &= 0 \end{aligned}$$

have more than one solution?

(15) A matrix is *nilpotent* if some power of it is zero, that is, $A^N = 0$ for some positive integer N. Prove that if A is nilpotent, then $\det A = 0$.

(16) (a) Solve the systems of equations
i) $x + y = 1,\ x - .9y = -1$
and
ii) $x + y = 1,\ x - 1.1y = -1$,
and compare your solutions, which should be almost the same.

(b) Solve the systems of equations
i) $x + y = 1,\ x + .9y = -1$,
and
$x + y = 1,\ x + 1.1y = -1$.
and again compare your solutions. Explain the result in terms of the theory in this section.

(c) Consider the solution of the systems of equations

$$\begin{aligned} x + y &= 1 \\ x + \alpha y &= -1 \end{aligned}$$

as the point where the lines $x + y = 1$ and $x + \alpha y = -1$ intersect. Sketch the graph of these lines for α near -1 and then for α near $+1$. Use these observations to again explain the phenomena in parts a) and b).

(17) Let Δ_n be the $n \times n$ determinant of a matrix with a's along the main diagonal and b's on the two "off diagonals" directly above and below the main diagonal. Thus

$$\Delta_5 = \begin{vmatrix} a & b & 0 & 0 & 0 \\ b & a & b & 0 & 0 \\ 0 & b & a & b & 0 \\ 0 & 0 & b & a & b \\ 0 & 0 & 0 & b & a \end{vmatrix}.$$

(a) Prove $\Delta_n = a\Delta_{n-1} - b^2\Delta_{n-2}$.

(b) Compute Δ_1 and Δ_2 by hand. Then use the formula to compute Δ_3 and Δ_4.

(c) If $a^2 \neq 4b^2$, can you show

$$\Delta_n = \frac{1}{\sqrt{a^2 - 4b^2}}\left[\left(\frac{a + \sqrt{a^2 - 4b^2}}{2}\right)^{n+1} - \left(\frac{a - \sqrt{a^2 - 4b^2}}{2}\right)^{n+1}\right]?$$

Later, we shall give a method for obtaining this directly from the equation of part a). [p. 522-523].

(18) Prove Part 5 of Theorem 21 using only the axioms and no other part of Theorem 21.

(19) Apply the result of Exercise 12 on page 389. Try to prove the following. A is a *square* matrix.

a). $\dim \mathcal{N}(A) = \dim \mathcal{N}(A^*)$.

Thus, the homogeneous equation $AX = 0$ has the same number of linearly independent solutions as does the equation $A^*Z = 0$.

b). Let $Z_1, \ldots, Z_k$ span $\mathcal{N}(A^*)$. Then the inhomogeneous equation

$$AX = Y$$

has a solution, that is, $Y \in \mathcal{R}(A)$, if and only if

$$\langle Z_j, Y \rangle = 0, \quad j = 1, 2, \ldots, k.$$

In other words, the equation $AX = Y$ has a solution if and only if Y is orthogonal to the solutions of the homogeneous adjoint equation.

c). Consider the system of linear equations

$$\begin{aligned} 2x - 3y + z &= 1 \\ -3x + 2y - 4z &= \alpha \\ x - 4y - 2z &= \beta. \end{aligned}$$

Let A be the coefficient matrix. Find a basis for $\mathcal{N}(A^*)$. [Answer: $\dim \mathcal{N}(A^*) = 1$ and $Z_1 = (2, 1, -1)$ is a basis]. For what value(s) of the constants α, β can you solve the given system of equations? [Answer: There is a solution if and only if $\beta - \alpha = 2$.] Find a solution if $\alpha = 1$ and $\beta = 3$.

d). Repeat part c) for the system of equations

$$\begin{aligned} x - y &= 1 \\ x - 2y &= -1 \\ x + 3y &= \alpha. \end{aligned}$$

[Answer: $\dim \mathcal{N}(A) = 1$ and $Z_1 = (-5, 4, 1)$ is a basis. There is a solution if and only if $\alpha = -1$].

(20) Use the result of Exercise 12 to prove that each of the following sets of functions are linearly independent everywhere.

a) $u_1(x) = \sin x$, $\quad u_2(x) = \cos x$

b) $u_1(x) = \sin nx$, $\quad u_2(x) = \cos mx$, where $n \neq 0$.

c) $u_1(x) = e^x$, $u_2(x) = e^{2x}$, $u_3(x) = e^{3x}$.

d) $u_1(x) = e^{ax}, u_2(x) = e^{bx}$, $u_3(x) = e^{cx}$, where a, b, and c are distinct numbers.

e) $u_1(x) = 1$, $u_2(x) = x$, $u_3(x) = x^2$, $u_4(x) = x^3$

f) $u_1(x) = e^x$, $u_2(x) = e^{-x}$, $u_3(x) = xe^x$, $u_4(x) = xe^{-x}$.

5.4 An Application to Genetics

A mathematical model is developed and solved. Although this particular model will be motivated by genetics, the resulting mathematical problem also arises in sociometrics and statistical mechanics as well as many other places. In the literature you will find these mathematical ideas listed under the title *Markov chains*.

Part of the value you should glean from our discourse is insight into the process of going from vague qualitative phenomena to setting up a quantitative model. One part of this scientific process we shall not have time to investigate in detail is the very important step of comparing the quantitative results with experimental data. Furthermore, we shall never delve into the fertile realm of generalizing our accumulated knowledge to more complicated - as well as more interesting and realistic - situations.

In bisexual mating, the genes of the resulting offspring occur in *pairs*, one gene in each pair being contributed by each parent. Consider the simplest case of a trait which is determined by a single pair of genes, each of which is one of two types g and G. Thus, the father contributes G or g to the pair, and the mother does likewise. Since experimental results show that the pair Gg is identical to the pair gG, the offspring has one of the three pairs

$$GG \qquad Gg \qquad gg.$$

The gene G *dominates* g if the resulting offspring with genetic types GG and Gg "appear" identical but both are different from gg. In this case, an individual with genetic type GG is called *dominant*, while the types gg and Gg are called *recessive* and *hybrid*, respectively.

An offspring can have the pair GG (resp. gg) if and only if *both* parents contributed a gene of type G (resp. g) while the combination Gg occurs if *either* parent contributed G and the other g. A fundamental *assumption* we shall make is that a parent with genetic type ab can only contribute a gene of type a or of type b. This assumption ignores such things as radioactivity as a genetic force. Thus, a dominant parent, GG can *only* contribute a dominant gene, G, a recessive parent, gg, can *only* contribute g, and a hybrid parent Gg can contribute *either* G or g (with equal probability). Consequently, if two *hybrids* are mated, the offspring has probability $\frac{1}{2}$ of getting G or g from each parent, so the probability of his having genetic type GG of gg is $\frac{1}{4}$ each, while the probability of having genetic type Gg is $\frac{1}{2}$.

We introduce a probability vector $V = (v_1, v_2, v_3)$, with v_1 representing the probability of being genetic type GG, v_2 of being type Gg, and v_3 of being type gg. Thus for an offspring of *two hybrid parents*, $V = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. Observe that, by definition of probability, $0 \leq v_j \leq 1$, $j = 1, 2, 3$, and $v_1 + v_2 + v_3 = 1$ (since with probability one - certainty - the offspring is either GG, Gg, or gg).

Consider the issue of mating an individual whose genetic type is *unknown* with an individual of *known* genetic type (dominant, hybrid or recessive). To be specific, assume the known person is of dominant type. Then the following matrix of *transition probabilities*

$$D = \begin{pmatrix} 1 & 1/2 & 0 \\ 0 & 1/2 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

describes the probability of the offspring's genetic type in the following sense: if the unknown parent had genetic type V_0 (so $V_0 = (1, 0, 0)$ if unknown was dominant, $V_0 = (0, 1, 0)$ if hybrid, and $V_0 = (0, 0, 1)$ if recessive), then

$$V_1 = DV_0,$$

is the probability vector of the offspring. For example, if the unknown parent was hybrid, then $V_1 = DV_0 = (\frac{1}{2}, \frac{1}{2}, 0)$. Thus the offspring can, with equal likelihood, be either dominant or hybrid, but cannot be recessive.

Notice that the matrix D embodies the fact that one of the parents is dominant.

If the individual of unknown genetic type were crossed with an individual of hybrid type, then the corresponding matrix H is

$$H = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{4} & \frac{1}{2} \end{pmatrix},$$

while if the person of unknown type were crossed with the individual of recessive type, then

$$R = \begin{pmatrix} 0 & 0 & 0 \\ 1 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 1 \end{pmatrix}.$$

It is of interest to investigate the question of *genetic stability* under various circumstances. Say we begin with an individual of unknown genetic type and cross it with a *dominant* individual, then cross that offspring with another dominant individual, and so on, always mating the resulting offspring with a dominant individual. Let V_n represent the genetic probability vector for the offspring in the nth generation. Then

$$V_n = DV_{n-1} = D^2 V_{n-2} = \cdots = D^n V_0,$$

where V_0 is the *unknown* vector for the initial parent (of unknown genetic type). Without knowing V_0, can we predict the eventual ($n \to \infty$) genetic types of the offspring? Intuitively, we expect that no matter what the type of the initial parent, the repeated mating with a dominant individual will produce a dominant strain. The question we are asking is, does $\lim_{n\to\infty} V_n$ exist, and if so, what is it?

Assume for the moment that the limit does exist and denote it by V. Then $V = DV$ since

$$V = \lim_{n\to\infty} V_n = \lim_{n\to\infty} V_{n+1} = \lim_{n\to\infty} DV_n = D(\lim_{n\to\infty} V_n) = DV$$

Armed with the equation $DV = V$, we can solve linear equations for the vector $V = (v_1, v_2, v_3)$

$$v_1 + \frac{1}{2}v_2 + 0 = v_1$$

$$0 + \frac{1}{2}v_2 + v_3 = v_2$$

$$0 + 0 + 0 = v_3.$$

Clearly $v_1 = v_2 = v_3 = 0$ is a trivial solution. A non-trivial one can be found by transposing the v_j's to the left side and solving. We find $v_1 = 1, v_2 = 0, v_3 = 0 (v_1 = 1$ since $v_1+v_2+v_3 = 1)$. Thus, *if* the limit V_n exists, the limit must be $V = (1,0,0)$. In genetic terms, this sustains our feeling that the offspring will eventually become genetically dominant.

But does the limit exist? To prove it does, we must show for any probability vector $V_0 = (v_1, v_2, v_3)$, where $v_1 + v_2 + v_3 = 1$, that the limit

$$\lim_{n\to\infty} V_n = \lim_{n\to\infty} D^n V_0,$$

exists and equals $V = (1,0,0)$. By evaluating D, D^2, and D^3 explicitly, we are led to guess

$$D^n = \begin{pmatrix} 1 & 1-\frac{1}{2^n} & 1-\frac{1}{2^{n-1}} \\ 0 & \frac{1}{2^n} & \frac{1}{2^{n-1}} \\ 0 & 0 & 0 \end{pmatrix},$$

which is then easily verified using mathematical induction. Thus

$$V_n = D^n V_0 = \begin{pmatrix} v_1 & + & (1-\frac{1}{2^n}) & + & (1-\frac{1}{2^{n-1}})v_3 \\ 0 & + & \frac{1}{2^n}v_2 & + & \frac{1}{2^{n-1}}v_3 \\ 0 & + & 0 & + & 0 \end{pmatrix}$$

$$= \begin{pmatrix} v_1+v_2+v_3 & - & \frac{1}{2^n}(v_2+2v_3) \\ \frac{1}{2^n}v_2 & + & \frac{1}{2^{n-1}}v_3 \\ & 0 & \end{pmatrix}$$

Since $v_1 + v_2 + v_3 = 1$, we find

$$V_n = D^n V_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \frac{1}{2^n}(v_2 + 2v_3)\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}.$$

It is now clear that the limit as $n \to \infty$ does exist, and is $V = (1,0,0)$. Consequently, if we begin with a random individual (you) and mate that individual and the successive offspring with a dominant gene bearer, then the resulting generations will tend to all dominant individuals. Moreover, the process proceeds exponentially because the "damping factor" is essentially $\frac{1}{2}$ for each generation (see above formula).

Were there enough time, you would see a second application of matrices to the special theory of relativity. Given your knowledge of linear spaces, it is possible to present an elegant exposition of the theory. The Lorentz transformation would appear as an orthogonal transformation - a rotation - in *world space* or *Minkowski's space* as it is often called. This is a four dimensional space three of whose dimension are those of ordinary space, while the fourth dimension is an *imaginary* ($i = \sqrt{-1}$) time dimension. Goldstein's *Classical Mechanics* contains the topic. Regrettably, he does not begin with the Michelson - Morley experiment but rather plunges immediately into mathematical technicalities.

Exercises

1. If you begin with an individual of unknown genetic type and cross it with a *hybrid* individual and then cross the successive offspring with hybrids, does the resulting strain approach equilibrium? If so, what is it?
2. Same as 1 but you mate an individual of unknown type with a *recessive* individual.
3. Beginning with an individual of unknown genetic type, you mate it with a *dominant* individual, mate the offspring with a *hybrid*, mate that offspring with a dominant, and continue mating *alternate* generations with dominants and hybrids respectively. Does the resulting strain approach equilibrium? If so, what is it? (You will need to define equilibrium to cope with this problem. There are several reasonable definitions.)
4. a). The city X has found that each year 5% of the city dwellers move to the suburbs, while only 1% of the suburbanites move to the city. Assuming the total population of the city plus suburb does not change, show that the matrix of transition probabilities is

$$P = \begin{pmatrix} .95 & .01 \\ .05 & .99 \end{pmatrix},$$

where a vector $V = (v_1, v_2) =$ (proportion of people in city, proportion of people in suburb).

b). Given any initial population distribution V, does the population approach an equilibrium distribution? If so, find it.

5. A long queue in front of a Moscow market in the Stalin era sees the butcher whisper to the first in line. He tells her "Yes, there is steak today." She tells the one behind her and so on down the line. However, Moscow housewives are not reliable transmitters. If one is told "yes", there is only an 80% chance she'll report "yes" to the person behind her. On the other hand, being optimistic, if one hears "no", she will report "yes" 40% of the time. If the queue is very long, what fraction of them will hear "there is no steak"? [This problem can be solved without finding a formula for P^n, although you might find it a challenge to find the formula].

5.5 A pause to find out where we are

.

We all know the homily about the forest and the trees. The next few pages are about the forest.

In the beginning we introduced dead linear spaces with their algebraic structure (Chapter II). Then we investigated the geometry induced by defining an inner product on a linear space and saw how easily many of the results in Euclidean geometry generalize (Chapter III).

Our next step was to consider mappings, linear mappings, between linear spaces (Chapter IV). Not much could be said in general, so we began investigating a particular case, linear maps between *finite* dimensional spaces. Two important special cases of this

$$L\colon \mathbb{R}^1 \to \mathbb{R}^n,$$

and

$$L\colon \mathbb{R}^n \to \mathbb{R}^1,$$

were treated before the general case,

$$L\colon \mathbb{R}^n \to \mathbb{R}^m.$$

A key theorem which facilitates the theory of linear mappings between finite dimensional spaces is the *representation theorem* (page 374): every such map can be represented as a matrix.

What next? There are two equally reasonable alternatives:

(A) We can continue with *linear maps*,

$$L\colon V_1 \to V_2,$$

and consider the case where V_1 or V_2, or both are *infinite* dimensional. The general theory here is in its youth and still undeveloped. Only one of the sources of difficulty is that a generalization of the representation theorem (page 374) remains unknown - except for some special cases. Thus, many special types of mappings have to be investigated individually. We shall consider only one type of linear mapping between infinite dimensional spaces, those defined by linear differential operators (Chapter VI and Chapter VII, Section 3).

(B) The second alternative is to continue our study of mappings between *finite* dimensional spaces, only now switch to *non linear* mappings. This theory should parallel the transition in elementary calculus from the analytic geometry of straight lines,

$$f(x) = a + bx,$$

that is, affine mappings, to genuine non linear mappings, as

$$f(x) = x^2 - 7\sqrt{x}$$

or

$$f(x) = x^3 - e^{\sin x}.$$

You recall, one important idea was to approximate the graph of a function $y = f(x)$ at a point x_0 by its tangent line at x_0, since for x near x_0, the curve and the tangent line there approximately agree. For example, one easily proves that at a maximum or minimum, the tangent line must be horizontal, $f' = 0$.

In generalizing this to functions of several variables,

$$Y = F(X) = F(x_1, \cdots, x_n),$$

the role of the derivative at X_0 is assumed by the affine map,

$$A(X) = Y_0 + LX,$$

which is tangent to F at X_0. Thus, linear algebra appears as the natural extension of analytic geometry to higher dimensional spaces. See Chapters VII - IX for this.

Chapter 6

Linear Ordinary Differential Equations

6.1 Introduction

.

A *differential equation* is an equation relating the values of a function $u(t)$ with the values of its derivatives at a point,

$$F(t, u(t), \frac{du}{dt}, \ldots, \frac{d^n u}{dt^n}) = 0 \tag{6-1}$$

The *order* of the equation is the order, n, of the highest derivative which appears. For example, the equations

$$\left(\frac{d^2u}{dt^2}\right)^3 - 7\frac{du}{dt} + t^2u^2 - \sin\, t = 0$$

$$\frac{du}{dt} - t\,\sin u^2 = 0$$

are of order two and one respectively. A function $u(t)$ is a *solution* of the differential equation if it has at least as many derivatives as the order of the equation, and if substitution of it into the equation yields an identity. Thus, the equation

$$\left(\frac{du}{dt}\right)^2 + u^2 = 1$$

has the function $u(t) = \sin t$ as a solution, since for all t

$$\left(\frac{d}{dt}\sin\, t\right)^2 + (\sin\, t)^2 = 1.$$

A differential equation (1) for the unknown function $u(t)$ is *linear* if it has the form

$$Lu := a_n(t)\frac{d^n u}{dt^n} + a_{n-1}(t)\frac{d^{n-1}}{dt^{n-1}} + \cdots + a_0(t)u = 0 \tag{6-2}$$

You should verify that this coincides with the notion of a linear operator used earlier. Equation (2) is sometimes called *linear homogeneous* to distinguish it from the *inhomogeneous* equation

$$Lu = f(t), \tag{6-3}$$

that is

$$a_n(t)\frac{d^n u}{dt^n} + \cdots + a_0(t)u = f(t). \tag{6-4}$$

The subject of this chapter is linear ordinary differential equations with *variable* coefficients (to distinguish them from the special case where the a_j's are constants). This operator L defined by (2) has as its domain the set of all sufficiently differentiable functions—n derivatives is enough. These functions constitute an infinite dimensional linear space. Thus, the differential operator L acts on an infinite dimensional space, as opposed to a matrix which acts on a finite dimensional space.

Differential equations abound throughout applications of mathematics. This is because most phenomena are described by laws which relate the rate of change of a function - the derivative - at a given time (or point) to the values of the function at that same time. For example, we have seen that at any time the acceleration of a harmonic oscillator is determined by its position and velocity at the same time,

$$\ddot{u} = -\mu\dot{u} - ku.$$

When confronted by a differential equation, your first reaction should be to attempt to find the solution explicitly. We were able to do this for linear constant coefficient equations (Chapter 4, Section 2). One of the main goals of this chapter is to show you how to solve as many *linear* ordinary differential equations as possible. However, it is naive to expect to solve an arbitrary equation which crops up in terms of the few functions we know: $x^\alpha, e^x, \log x, \sin x$, and $\cos x$. In fact, to even solve the elementary equation

$$\frac{du}{dx} = \frac{1}{x},$$

appearing in elementary calculus, we were forced to *define* a new function as the solution of this equation

$$u(x) = \log x + c$$

and obtain the properties of this function and its inverse e^x directly from the differential equation. Many many functions arise which cannot be expressed in terms of the few elementary functions we know and love. Most of these functions - like Bessel's functions, elliptic functions, and hypergeometric functions, arise directly because they are the solutions of differential equations nature has forced us to consider.

How do we know these strange sounding functions are solutions of the differential equations? Well, we somehow prove a solution exists and then simply give a name to the solution - much as babies are given names at birth. Furthermore, as is the case with babies, their actual "names" are the least important aspect.

To summarize briefly, we shall solve as many equations as we can. For the remaining ones (which include most equations), we shall attempt to describe a few of the main properties so that if one arises in your work, you will have a place to begin the attack. Later on, we shall again return to the more complicated situation of *nonlinear* equations. Much less can be said there. Only very few general results are known.

Lest you get the wrong idea, we shall cover but a fraction of the known theory for just linear ordinary differential equations. In the next chapter, we shall only look at one *partial differential equation* (the wave equation for a vibrating violin string). The general theory there is too complicated to allow discussion for more than one particular equation.

Exercises

1. Assume there *exists* a *unique* function $E(x)$ which satisfies the following differential equation for all x and satisfies the initial condition

$$\frac{du}{dx} = u, \qquad u(0) = 1.$$

(a) Use the "chain rule" and uniqueness to prove for any $a \in \mathbb{R}$

$$E(x+a) = E(a)E(x)$$

[Hint: Prove $\tilde{E}(x) := E(x+a)$ is also a solution of the equation. Then apply the uniqueness to the function $\tilde{E}(x)/E(a)$].

(b) Prove

$$E(-x) = \frac{1}{E(x)}.$$

(c) Prove for any x

$$E(nx) = [E(x)]^n, \quad n \in \mathbb{Z}.$$

In particular, show

$$E(n) = [E(1)]^n, \quad n \in \mathbb{Z}$$

and

$$E(\frac{1}{m}) = [E(1)]^{1/m}, \quad m \in \mathbb{Z}_+$$

(d) Prove

$$E(\frac{n}{m}) = [E(1)]^{n/m}, \quad n \in \mathbb{Z}, \quad m \in \mathbb{Z}_+$$

[Thus, the function $E(x)$ is defined for all rational $x = \frac{n}{m}$ as the number $E(1)$ to the power n/m. Since $E(x)$ is continuous (even differentiable by definition, we can extend the last formula to irrational x by continuity: if r_j is a sequence of rational numbers converging to the real number x (which may or may not be rational) then by continuity

$$E(x) = \lim_{j\to\infty} E(r_j) = \lim_{j\to\infty} [E(1)]^r j = E(1)^x.$$

Consequently, $E(x)$ is the familiar exponential function e^x].

2. Find the general solutions of the following equations by any method you can.

(a) $\frac{du}{dx} - 2u = 0$

(b) $\frac{du}{dx} = x^2 + \sin x$

(c) $\left(\frac{du}{dx}\right)^2 + 4u^2 = 1$

(d) $\frac{du}{dx} = \frac{x}{u+1}$

(e) $\frac{du}{dx} = x^2 e^u$

(f) $\frac{d^2u}{dx^2} + 3\frac{du}{dx} - 4u = 4$

6.2 First Order Linear

.

Except for those differential equations which can be solved by inspection, the next most simple equation is one which is linear and first order, the *homogeneous equation*

$$\frac{du}{dx} + a(x)u = 0, \tag{6-5}$$

and the *inhomogeneous equation*

$$\frac{du}{dx} + a(x)u = f(x). \tag{6-6}$$

The homogeneous equation can be solved by first writing it in the form

$$\frac{1}{u}\frac{du}{dx} = -a(x)$$

and then integrating both sides

$$\log u(x) = -\int^x a(s)\,ds + C_1.$$

Thus

$$u(x) = Ce - \int^x a(s)\,ds \tag{6-7}$$

is the solution of equation (4) for any constant C. In the very special case $a(s) \equiv$ constant, the solution does have the form found earlier (Chapter 4, Section 2) for a linear equation with constant coefficients.

How can we integrate the inhomogeneous equation (5)? A useful device is needed. Multiply both sides of this equation by an unknown function $q(x)$

$$q(x)\frac{du}{dx} + q(x)a(x)u = q(x)f(x),$$

If we can find $q(x)$ so that the left side is a derivative,

$$q(x)\frac{du}{dx} + q(x)a(x)u = \frac{d}{dx}(q(x)u), \tag{6-8}$$

then the equation reads

$$\frac{d}{dx}(q(x)u) = q(x)f(x),$$

which can be integrated immediately,

$$q(x)u(x) = \int^x q(s)f(s)\,ds + c, \tag{6-9}$$

and then solved for $u(x)$ by dividing by $q(x)$.

Thus, the problem is reduced to finding a $q(x)$ which satisfies (7). Evaluating the right side of (7), we find

$$q\frac{du}{dx} + qa\,u = u\frac{dq}{dx} + q\frac{du}{dx},$$

so $q(x)$ must satisfy

$$\frac{dq}{dx} = q(x)a(x).$$

It is easy to find a function $q(x)$ which satisfies this - for it is a homogeneous equation of the form (4). Therefore

$$q(x) = e^{\int^x a(t)\,dt},$$

the reciprocal of the solution (6) to the homogeneous equation, does satisfy (7). Notice we have ignored the arbitrary constant factor in the solution since all we want is any one function $q(x)$ for (7).

Now we can substitute into (8) to find the solution of the inhomogeneous equation

$$u(x) = \frac{1}{q(x)}\int^x q(s)f(s)\,ds + \frac{c}{q(x)}, \tag{6-10}$$

where $q(x)$ is given by the formula at the top of the page. If it makes you happier, substitute the expression for $q(x)$ into (9) to obtain the messy formula. We have left some room.

A FIGURE GOES HERE

EXAMPLES: 1. $\frac{du}{dx} + \frac{2}{x}u = (1+x^3)^{17}, \quad x \neq 0.$
First,
$$q(x) = \exp(\int^x \frac{2}{s}\, ds) = \exp(2\ln\, x) = \exp(\ln x^2) = x^2.$$
Thus
$$\frac{d}{dx}(x^2 u) = x^2(1+x^3)^{17}.$$
Integrating both sides we find
$$x^2 u(x) = \frac{(1+x^3)^{18}}{54} + C.$$
Therefore
$$u(x) = \frac{1}{54}\frac{(1+x^3)^{18}}{x^2} + \frac{C}{x^2}, \quad x \neq 0.$$

2. $\frac{du}{dx} + 2xu = x$
First,
$$q(x) = \exp(\int^x 2s\, ds) = \exp x^2.$$
Thus,
$$\frac{d}{dx}(e^{x^2}u) = e^{x^2}x.$$
Integrating both sides, we find
$$e^{x^2}u(x) = \frac{1}{2}e^{x^2} + C,$$
so
$$u(x) = \frac{1}{2} + Ce^{-x^2}.$$
This formula could have been guessed much earlier since we know the general solution of the inhomogeneous equation can be expressed as the sum of a particular solution to that equation plus the general solution of the homogeneous equation. The particular solution $u_0(x) = \frac{1}{2}$ can be obtained by inspection of the D.E.

Let us summarize our results.

Theorem 6.1 . *Consider the first order linear inhomogeneous equation*
$$Lu := \frac{du}{dx} + a(x)u = f(x).$$
If $a(x)$ and $f(x)$ are continuous functions, the equation has the solutions
$$u(x) = \tilde{u}(x)\int^x \frac{f(s)}{\tilde{u}(s)}\, ds + C\tilde{u}(x) \tag{6-11}$$
where
$$\tilde{u}(x) = \exp(-\int^x a(s)\, ds)$$
is a non-trivial solution of the homogeneous equation. Moreover, if we specify the initial condition $u(x_0) = \alpha$, then the solution which satisfies this initial condition is unique.

PROOF: The *existence* follows from the explicit formula (9) or (10) and from the fact that a continuous function is always integrable.

Uniqueness. This will be quite similar to the proof carried out in Chapter 4. If $u_1(x)$ and $u_2(x)$ are two solutions of the inhomogeneous equation $Lu = f$, with the *same* initial conditions, then the function
$$w(x) := u_1(x) - u_2(x)$$
satisfies the homogeneous equation
$$Lw := w' + a(x)w = 0,$$
and is zero at x_0,
$$w(x_0) = u_1(x_0) - u_2(x_0) = 0.$$
Our task is to prove $w(x) \equiv 0$. Multiply the equation (20) by $w(x)$. Then
$$ww' = -a(x)w^2,$$
or
$$\frac{1}{2}\frac{d}{dx}w^2 = -a(x)w^2.$$
Since $a(x)$ is continuous, for any closed and bounded interval $[A, B]$ there is a constant k (depending on the interval) such that $-a(x) \leq k$ for all $x \in [A, B]$. Consequently,
$$\frac{1}{2}\frac{d}{dx}w^2 \leq kw^2,$$
or
$$\frac{d}{dx}w^2 - 2kw^2 \leq 0.$$
Now we need an important identity which can be verified by direct computation: for any smooth function g, and any constant α, $g' + \alpha g = e^{-\alpha x}(e^{\alpha x}g)'$. We apply this to the above inequality with $g = w^2$ and $\alpha = -2k$ to conclude that
$$e^{2kx}\frac{d}{dx}[e^{-2kx}w^2] \leq 0.$$
Because e^{2kx} is always positive, by the mean value theorem this inequality states that $e^{-2kx}w^2$ is a decreasing function of x. Thus
$$e^{-2kx}w^2(x) \leq e^{-2kx_0}w^2(x_0), \quad x \geq x_0,$$
or
$$w^2(x) \leq e^{2k(x-x_0)}w^2(x_0), \quad x \geq x_0.$$
But since $w(x_0) = 0$ and $w^2(x) \geq 0$ this means that
$$0 \leq w^2(x) \leq 0.$$

Therefore $w(x) \equiv 0 \quad x \geq x_0$.

To prove $w(x) \equiv 0$ for $x \leq x_0$, merely observe that the equation (11) has the same form if x is replaced by $-x$. Thus the above proof applies and shows $w(x) \equiv 0$ for $x \leq x_0$ too.

Remark: Although a formula has been exhibited for the solution, this does not mean that the integrals which occur can be evaluated in terms of elementary functions. These integrals however can be at least evaluated approximately using a computer if a numerical result is needed.

Exercises

(1) . Find the solution of the following equations with given initial values

(a) $u' + 7u = 3, \quad u(1) = 2$
(b) $5u' - 2u = e^{3x}, \quad u(0) = 1$.
(c) $3u' + u = x - 2x^2, \quad u(-1) = 0$.
(d) $xu' + u = 4x^3 + 2, \quad u(1) = -1$.
(e) $u' + (\cot x)u = e^{\cos x} + 1, \quad u(\frac{\pi}{2}) = 0$. $[\int \cot x\, dx = \ln(\sin x)]$.

(2) . The differential equation

$$L\frac{du}{dt} + Ru = E \sin \omega t,\ L, R, E \text{ constants}$$

arises in circuit theory. Find the solution satisfying $u(0) = 0$ and show that it can be written in the form

$$u(t) = \frac{\omega EL}{R^2 + \omega^2 L^2} e^{-Rt/L} + \frac{E}{\sqrt{R^2 + \omega^2 L^2}} \sin(\omega t - \alpha)$$

where

$$\tan \alpha = \frac{\omega L}{R}.$$

(3) Bernoulli's equation is

$$u' + a(x)u = b(x)u^k, \quad k \quad \text{a constant.}$$

(a) Use the substitution $v(x) = u(x)^{1-k}$ to transform this nonlinear equation to the linear equation

$$v' + (1-k)a(x)v = (1-k)b(x).$$

(b) Apply the above procedure to find the general solution of

$$u' - 2e^x u = e^x u^{3/2}.$$

(4) . Consider the equation

$$u' + au = f(x),$$

where a is a constant, f is continuous in the interval $[0, \infty]$, and $|f(x)| < M$ for all x.

(a) Show that the solution of this equation is

$$u(x) = e^{-ax}u(0) + e^{-ax}\int_0^x e^{at} f(t)\, dt$$

(b) Prove (if $a \neq 0$)

$$\left|u(x) - e^{-ax}u(0)\right| \leq \frac{M}{a}[1 - e^{-ax}].$$

(5) (a) Show the uniqueness proof yields the following stronger fact. If $u_1(x)$ and $u_2(x)$ are both solutions of the same equation

$$u' + a(x)u = f(x)$$

but satisfy *different* initial conditions

$$u_1(x_0) = \alpha, \quad u_2(x_0) = \beta,$$

then

$$|u_1(x) - u_2(x)| \leq e^{k(x-x_0)}\,|\alpha - \beta|, \quad x \geq x_0$$

for all $x \in [A, B]$, where $-a(x) < k$ in the interval. Thus, if the initial values are close, then the solutions cannot get too far apart.

(b) Show that if $a(x) \leq A < 0$, where A is a constant, then as $x \to 0$ any two solutions of the same equation - but with possibly different initial values - tend to the same function.

(6) . Show that the differential equation

$$y' = a(x)F(y) + b(x)G(y)$$

can be reduced to a linear equation by the substitution

$$u = F(y)/G(y) \quad \text{or} \quad u = G(y)/F(y)$$

if $(FG' - GF')/G$ or $(FG' - GF')/F$, respectively, is a constant. Use this substitution to again solve Bernoulli's equation.

(7) . Let $S = \{\, u \in C' \colon u(0) = 0 \,\}$, and define the operator L from S to C by

$$Lu = u' + u.$$

Prove L is injective and $\mathcal{R}(L) = C$.

(8) . Set up the differential equation and solve. The rate of growth of a bacteria culture at any time t is proportional to the amount of material present at that time. If there was one ounce of culture in 1940 and 3 ounces in 1950, find the amount present in the year 2000. The *doubling time* is the interval it takes for a given amount to double. Find the doubling time for this example.

(9) . Find the general solution of $x^2u' + 3xu = \sin x$.

(10) . Assume that a body decreases its temperature $u(t)$ at a rate proportional to the difference between the temperature of the body and the temperature T of the surrounding air. A body originally at a temperature of 100^0 is placed in air which is kept at a temperature of 50^0. If at the end of one hour the temperature of the body has fallen 20^0, how long will it take for the body to reach 60^0?

(11) . Here is one simple mathematical model governing economic behavior. Think of yourself as a widget manufacturer for now. Let

i) $S(t)$ be the supply of widgets available at time t. This is the only function you can control directly.

ii) $P(t)$ be the market price of a widget at time t.

iii) $D(t)$ is the demand for widgets at time t—the number of widgets people want to buy at time t. You cannot control this given function.

It has been found that the market price $P(t)$ changes at a rate proportional to the difference between demand and supply,

$$\frac{dP}{dt} = k(D(t) - S(t)),$$

where $k > 0$ is a fixed constant.

You decide to vary the supply so that it is a fixed constant S_0 plus an amount proportional to the market price,

$$S(t) = S_0 + \alpha P(t), \quad \alpha > 0.$$

(a) Set up the differential equation for $S(t)$ in terms of the given function $D(t)$ and solve it.

(b) Analyze the solution and give an argument making it plausible that the market for widgets behaves roughly in this way. What criticisms can you make of the model?

(c) How does the market behave if the demand increases for a long time and then levels off at some constant value, $D(t) = D(t_1)$ for $t \geq t_1$? A qualitative description of $S(t)$ and $P(t)$ is called for here. In particular, say whether price increases without bound (bringing the evils of inflation) or whether it, too, levels off.

(12) It is found that a juicy rumor spreads at a rate proportional to the number of people who “know”. If one person knows initially, $t = 0$, and tells one other person by the next day, $t = 1$, approximately how long does it take before 4000 people know? Analyze the mathematical model as $t \to \infty$ and state why it is, in fact, the wrong model. (The question to ask yourself is, “how long will it take before everyone even remotely concerned knows?”). The same mathematical model applies to the spreading of contagious diseases - and many other similar phenomena.

6.3 Linear Equations of Second Order

In this section we will consider a portion of the general theory of second order linear O.D.E.'s, with variable coefficients,

$$Lu := a_2(x)\frac{d^2u}{dx^2} + a_1(x)\frac{du}{dx} + a_0(x)u = f(x).$$

Although *all* of the results obtained generalize immediately to linear equation of order n, only the special case $n = 2$ will be treated. This special case has the advantage of clearly illustrating the general situation and supplying proofs which generalize immediately - while avoiding the inevitable computation complexities inherent in the general case.

There are three parts:
A). a review of the constant coefficient case,
B). power series solutions, and
C). the general theory.

Whereas the first two parts are concerned with obtaining explicit formulas for the solutions, the last resigns itself to some statements which can be made without finding the solution explicitly.

a) A Review of the Constant Coefficient Case.

Here we have the operator

$$Lu := a_2u'' + a_1u' + a_0u, \tag{6-12}$$

where a_0, a_1, and a_2 are constants. In order to solve the homogeneous equation

$$Lu = 0,$$

the function $e^{\lambda x}$ is tried. Substitution yields

$$L(e^{\lambda x}) = (a_2\lambda^2 + a_1\lambda + a_0)e^{\lambda x} = p(\lambda)e^{\lambda x}. \tag{6-13}$$

labeleq:13 The polynomial $p(\lambda)$ is called the *characteristic polynomial* for L. If λ_1 is a root of this polynomial, $p(\lambda_1) = 0$, then $u_1(x) = e^{\lambda_1 x}$ is a solution of the homogeneous equation $Lu = 0$. If λ_2 is another root of this polynomial $\lambda_1 \neq \lambda_2$, $u_2(x) = e^{\lambda_2 x}$ is another solution. Then every function of the form

$$u(x) = Au_1(x) + Bu_2(x) = Ae^{\lambda_1 x} + Be^{\lambda_2 x}, \tag{6-14}$$

where A and B are constants, is a solution of the homogeneous equation. The uniqueness theorem showed that *every* solution of $Lu = 0$ is of the form (14).

If the two *roots* of $p(\lambda)$ *coincide*, then a second solution is $u_2(x) = xe^{\lambda_1 x}$, and every function of the form

$$u(x) = Au_1(x) + Bu_2(x) = Ae^{\lambda_1 x} + Bxe^{\lambda_1 x} \tag{6-15}$$

where A and B are constants, is a solution of the homogeneous equation. Again the uniqueness theorem showed that every solution of $Lu = 0$ is of the form (15).

In both (14) and (15), the constants A and B can be chosen to find a unique function $u(x)$ which satisfies the homogeneous equation

$$Lu = 0$$

as well as the *initial conditions*

$$u(x_0) = \alpha, \qquad u'(x_0) = \beta,$$

where α and β are specified constants.

It turns out that the inhomogeneous O.D.E.

$$Lu = f,$$

where f is a given continuous function, can *always* be solved once two linearly independent solutions u_1 and u_2 of the homogeneous equation $Lu_j = 0$ are known. Since the procedure for solving the inhomogeneous equation also works if the coefficients in the differential operator L are not constant, it is described later in this section in the more general situation (p. 487-8, Theorem 8). Somewhat simpler techniques can be used for the constant coefficient equation if the function f is a linear combination of functions of the form $x^k e^{rx}$, where k is a nonnegative integer and r is some real or complex constant (cf. Exercise 6, p. 300). Because both $\sin nx$ and $\cos nx$ are of this form, Fourier series can be used to supply a solution for any function f which has a convergent Fourier series (cf. Exercise 13, p. 303).

Section 5 of this chapter contains an interesting generalization of the theory for constant coefficient ordinary differential operators to operators which are "translation invariant".

b) Power Series Solutions

.

Many ordinary differential equations (linear and nonlinear) can be solved by merely assuming the solution can be expanded in a power series $u(x) = \sum c_n x^n$, and plugging into the differential equation to find the coefficients c_n. A simple example illustrates this.

EXAMPLE: Solve $u'' - 2xu' = 0$ with the initial conditions $u(0) = 1, u'(0) = 0$.

Solution: We try

$$u(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n + \cdots .$$

Then

$$u'(x) = c_1 + 2c_2 x + 3c_3 x^2 + \cdots + nc_n x^{n-1} + \cdots$$

so

$$2xu'(x) = 2c_1 x + 4c_2 x^2 + \cdots + 2nc_n x^n + \cdots$$

Also

$$u''(x) = 2c_2 + 2 \cdot 3c_3 x + 3 \cdot 4c_4 x^2 + \cdot + (n-1)nc_n x^n + \cdots$$

Adding $u'' - 2xu' - u$ and collecting like powers of x we find that $0 = u'' - 2xu' - u = [2c_2 - c_0] + [2 \cdot 3c_3 - 2c_1 - c_1]x + [3 \cdot 4c_4 - 4c_2 - c_2]x^2$

$$+ \cdots + [(k+1)(k+2)c_{k+2} - 2kc_k - c_k]x^k + \cdots$$

If the right side, a Taylor series, is to be zero (= the left side), then the coefficient of each power of x must vanish because the only convergent Taylor series for zero is zero itself.

The coefficient of

x^0 is	$2c_2 - c_0$
x^1 is	$6c_3 - 3c_1$
x^2 is	$12c_4 - 5c_2$
x^k is	$(k+1)(k+2)c_{k+2} - (2k+1)c_k$

Equating these to zero we find that

$$c_2 = \frac{c_0}{2}, \quad c_3 = \frac{c_1}{2}, \quad c_4 = \frac{5c_2}{12} = \frac{5}{24}c_0$$

and, more generally,

$$c_{k+2} = \frac{2k+1}{(k+2)(k+1)}c_k. \tag{6-16}$$

Thus, for this example e_{even} is some multiple of c_0 while c_{odd} is some multiple of c_1. Since $u(0) = c_0$ and $u'(0) = c_1$, the constants c_0 and c_1 are determined by the initial conditions.

$$c_0 = 1, \quad c_1 = 0.$$

Consequently, all of the odd coefficients $c_3, c_5, \ldots$ vanish, while

$$c_2 = \frac{1}{2}, \quad c_4 = \frac{5}{24}, \quad c_6 = \frac{3}{10}c_4 = \frac{1}{16}, \quad c_8 = \ldots,$$

so the first few terms in the series for $u(x)$ are

$$u(x) = 1 + \frac{1}{2}x^2 + \frac{5}{24}x^4 + \frac{1}{16}x^6 + \cdots \tag{6-17}$$

We should investigate if this formal power series expansion converges. Using (16), the ratio of successive terms in the series for $u(x)$ is

$$\left|\frac{c_{k+2}x^{k+2}}{c_k x^k}\right| = \left|\frac{(2k+1)}{(k+2)(k+1)}x^2\right|$$

Therefore the ratio test shows the *formal* power series actually converges for all x. By Theorem 16, p. 82, the series can be differentiated term by term and does satisfy the equation.

Although the computation is lengthy, the series (17) is a solution. Since there is no way of finding the solution in terms of elementary functions, we must be contented with the power series solution. You have seen (Chapter 1, Section 7) how properties of a function can be extracted from a power series definition.

This example is typical.

Theorem 6.2 . *If the differential equation*

$$a_2(x)u'' + a_1(x)u' + a_0(x)u = 0$$

has analytic coefficients about $x = 0$, that is, if the coefficients all have convergent Taylor series expansions about $x = 0$, and if $a_2(0) \neq 0$, then given any initial values

$$u(0) = \alpha, \qquad u'(0) = \beta,$$

there is a unique solution $u(x)$ which satisfies the equation and initial conditions. Moreover, the solution is analytic about $x = 0$ and converges in the largest interval $[-r, r]$ in which the series for a_1/a_2 and a_0/a_2 both converge.

Outline of Proof. There are two parts: i) find a formal power series $u(x) = \sum c_n x^n$, and ii) prove the formal power series converges. Since explicit formulas can be found for the c_n's (cf. Exercise 30a) the first part is true. Proof of the second part is sketched in the exercises too (Exercise 30b).

From the explicit formulas mentioned above for the c_n's, it is clear there is at most one *analytic* solution. But because the general uniqueness proof (p. 510, Theorem 9) states there is at most one solution which is twice differentiable and since $u(x)$ is certainly such a function - the uniqueness of $u(x)$ among all twice differentiable functions follows as soon as Theorem 9 is proved.

The restriction $a_2(0) \neq 0$ which was made in Theorem 3 is very important. If $a_2(0) = 0$ then the differential equation

$$a_2(x)u'' + a_1(x)u' + a_0(x)u = 0$$

is degenerate at $x = 0$ because the coefficient of the highest order derivative vanishes there. Then the point $x = 0$ is called a *singularity* of the differential equation. A simple example illustrates the situation. The function $u(x) = x^{5/2}$ satisfies the differential equation

$$4x^2u'' - 15u = 0$$

and the initial conditions $u(0) = 0, u'(0) = 0$. However $u(x) \equiv 0$ is also a solution. Thus it will be impossible to prove any uniqueness theorem at $x = 0$ for this equation. Perhaps the singular nature of this equation at $x = 0$ is more vivid if the equation is written as

$$u'' - \frac{15}{4x^2}u = 0.$$

Although the possibility of a uniqueness result is ruled out for equations with singularities, it is important to be able to find the non-zero solutions of these equations, important because many of the equations which arise in practice do happen to have singularities (Bessel's equation, Legendre's equation, the hypergeometric equation, ...). In all of the commonly occurring cases, the coefficients $a_0(x)$, $a_1(x)$, and $a_2(x)$,

$$a_2u'' + a_1u' + a_0u = 0,$$

are analytic functions. Thus the only obstacle to applying Theorem 3 is the condition $a_2 \neq 0$. We persist, however, in the belief that a power series, or some modification of it, should work. The modification must allow for such solutions as $u(x) = x^{3/2}$ which do not have Taylor expansions about $x = 0$. Undoubtedly the most naive candidate for a solution is to try

$$u(x) = x^\rho \sum_{n=0}^{\infty} c_n x^n, \tag{6-18}$$

where ρ may be any real number. The particular choice $\rho = 3/2$, $c_0 = 1, c_1 = c_2 = c_3 = \ldots = 0$ does yield the function $u(x) = x^{3/2}$. It turns out that (18) is *usually* the correct guess.

Again, we turn to an example. *Bessel's equation* of order n,

$$x^2u'' + xu' + (x^2 - n^2)u = 0,$$

which arises in the study of waves in a two dimensional circular domain, like those on tympani, in a tea cup, or on your ear drum. Let us find a solution to Bessel's equation of order one,

$$x^2u'' + xu' + (x^2 - 1)u = 0 \tag{6-19}$$

This equation does have a singularity at the origin, $x = 0$. If u has the form (18), then

$$u(x) = \sum_{n=0}^{\infty} c_n x^{n+\rho},$$

$$u'(x) = \sum_{n=0}^{\infty} (n+\rho)c_n x^{n+\rho-1},$$

and

$$u''(x) = \sum_{n=0}^{\infty} (n+\rho)(n+\rho-1)c^n x^{n+\rho-2}.$$

Substituting this into the differential equation (19), we find

$$\sum_{n=0}^{\infty}(n+\rho)(n+\rho-1)c_n x^{n+p} + \sum_{n=0}^{\infty}(n+\rho)c_n x^{n+\rho}$$
$$+\sum_{n=0}^{\infty} c_n x^{n+\rho+2} - \sum_{n=0}^{\infty} c_n x^{n+\rho} = 0. \tag{6-20}$$

We must equate the coefficients of successive powers of x to zero. The lowest power of x which appears is x^{ρ}, the next $x^{\rho+1}$, and so on.

$$\begin{aligned}
x^{\rho}:&\quad \rho(\rho-1)c_0 + \rho c_0 - c_0 = 0\\
x^{\rho+1}:&\quad (\rho+1)\rho c_1 + (\rho+1)c_1 - c_1 = 0\\
x^{\rho+2}:&\quad (\rho+2)(\rho+1)c_2 + (\rho+2)c_2 + c_0 - c_2 = 0\\
x^{\rho+3}:&\quad (\rho+3)(\rho+2)c_3 + (\rho+3)c_3 + c_1 - c_3 = 0\\
&\ \ \vdots\\
x^{\rho+n}:&\quad (\rho+n)(\rho+n-1)c_n + (\rho+n)c_n + c_{n-2} - c_n = 0.
\end{aligned}$$

From the equation for the power x^{ρ}, we find

$$(\rho^2-1)c_0 = 0$$

The polynomial $q(\rho) = \rho^2 - 1$ which appears in the coefficient of the lowest power of x in (20) is called the *indicial polynomial* since it will be used to determine the *index* ρ. If $c_0 \neq 0$, the equation $(\rho_2 - 1)c_0 = 0$ can be satisfied only if ρ is a root of the indicial polynomial. Thus $\rho_1 = 1$, $\rho_2 = -1$.

Consider the largest root $\rho_1 = 1$. Then the equation for the coefficients of $x^{\rho+1}$ in (20) is

$$x^{\rho+1} = x^2 : 3c_1 = 0 \Rightarrow c_1 = 0,$$

while the equation for the coefficient of $x^{\rho+n}$ in (20) is

$$x^{\rho+n} = x^{1+n} : (n+1)nc_n + (n+1)c_n + c_{n-2} - c_n = 0,$$

or

$$c_n = -\frac{c_{n-2}}{n(n+2)}, \qquad n = 2, 3, \dots$$

Since $c_1 = 0$, this equation implies $c_{\text{odd}} = 0$ and determines the c_{even} in terms of c_0,

$$c_2 = -\frac{c_0}{2\cdot 4}, \quad c_4 = -\frac{c_2}{4\cdot 6} = \frac{c_0}{2\cdot 4^2\cdot 6}, \quad c_6 = -\frac{c_4}{6\cdot 8} = \frac{c_0}{2\cdot 4^2\cdot 6^2\cdot 8}$$

$$c_{2k} = \frac{(-1)^k c_0}{2\cdot 4^2\cdot 6^2\cdot 8^2\cdots(2k)^2(2k+2)} = \frac{(-1)^k c_0}{2^{2k}k!(k+1)!}.$$

Thus, the formal series we find for the solution, $J_1(x)$, of the Bessel equation of first order corresponding to the largest indicial root, $\rho_1 = 1$ is

$$J_1(x) = \frac{1}{2}x^1(1 - \frac{x^2}{2\cdot 4} + \frac{x^4}{2\cdot 4^2\cdot 6} - \cdots)$$

or

$$J_1(x) = \frac{1}{2}x\sum_{k=0}^{\infty}\frac{(-1)^k x^{2k}}{2^{2k}k!(k+1)!} \tag{6-21}$$

since it is customary to choose the constant c_0 for $J_1(x)$ as $c_0 = \frac{1}{2}$ (and the constant c_0 for $J_n(x)$ as $1/2^n n!$ when n is a positive integer).

The other (smaller) root, $\rho_2 = -1$, is much more difficult to treat. If the above steps are imitated (which you should try), division by zero needed to solve for c_2 from c_0. It turns out that the solution corresponding to the smaller root $\rho_2 = -1$ is not of the form (18). We shall not enter into this matter further except to note that the difficulty occurs because the *two roots* ρ_1 and ρ_2 *differ by an integer.* If the two roots ρ_1 and ρ_2 do *not* differ by an integer, the above method yields two different solutions of the form (18) for the equation. In any case, this method always gives a solution of the form (18) for the *largest* root of the indicial equation.

It is easy to check that the power series (21) does converge for all x and is therefore a solution to Bessel's equation of the first order. From the power series, with considerable effort one can obtain a series of identities for Bessel functions which exactly parallels those for the trigonometric functions. The functions $J_n(x)$ behaving in many ways similar to $\sin nx$ or $\cos nx$. Here is a graph of $J_1(x)$:

A FIGURE GOES HERE

For x very large, $J_1(x)$ is asymptotically

$$J_1(x) \sim \sqrt{\frac{2}{\pi}}\frac{\cos(x - 3\pi/4)}{\sqrt{x}},$$

which is a cosine curve whose amplitude decreases like $1/\sqrt{x}$. For good reason this curve resembles the height of surface waves on a lake after a pebble has been dropped into the water, or those on the surface of a cup of tea.

Having worked out this example in detail, we shall state a definition in preparation for our theorem.

Definition: The differential equation

$$a_2(x)u'' + a_1(x)u' + a_0(x)u = 0,$$

where the $a_j(x)$ are analytic about $x = 0$, it has a *regular singularity* at $x = 0$ if it can be written in the form

$$x^2u'' + A(x)xu' + B(x)u = 0,$$

where the functions $A(x)$ and $B(x)$ are analytic about $x=0$. Otherwise the singularity is *irregular.*

EXAMPLES:

(1) . $x^2(1+x)u''+2(\sin x)u'-e^x u=-$ has a regular singularity at $x=0$ since the equation may be written as

$$x^2u''+\frac{2(\sin x)}{1+x}u'-\frac{-e^x}{1+x}u=0,$$

where the coefficients $2\sin x/(1+x)x$ and $e^x/1+x$ do have convergent Taylor series about $x=0$. (Here we observed that $\frac{\sin x}{x}=1-\frac{x^2}{3!}+\cdots$).

(2) . $xu''-7u'+\frac{3}{\cos x}u=0$ has a regular singularity at $x=0$ since it can be written in the form

$$x^2u''-7xu'+\frac{3x}{\cos x}u=0,$$

where the coefficients -7 and $3x/\cos x$ are analytic about $x=0$.

(3) . $x^2u''-2u'+xu=0$ has an irregular singularity at $x=0$ since it cannot be written in the desired form.

(4) . $x^3u''-2xu'+u=0$ has an irregular singularity at $x=0$.

Theorem 6.3 . *(Frobenius) Consider the equation with a regular singularity at* $x=0$

$$a_2(x)u''+a_1(x)u'+a_0(x)u=0,$$

so it can be written in the form

$$x^2u''+A(x)xu'+B(x)u=0,$$

where the analytic function $A(x)$ *and* $B(x)$ *have convergent power series for* $|x|<r$. *Let* ρ_1 *and* ρ_2 *be the roots of the indicial polynomial*

$$q(\rho)=\rho(\rho-1)+A(0)\rho+B(0),$$

where $\rho_1\ge\rho_2$ *(or* $Re\rho_1\ge Re\rho_2$ *if roots are complex). Then the differential equation has one solution* $u_1(x)$ *of the form*

$$u_1(x)=x^{\rho_1}\sum_{n=0}^{\infty}c_nx^n \qquad (c_0\ne 0),$$

the series converging for all $|x|<r$. *Moreover, if* $\rho_1-\rho_2$ *is not an integer (or zero), there is a second solution* $u_2(x)$ *of the form*

$$u_2(x)=x^{\rho_2}\sum_{n=0}^{\infty}\tilde{c}_nx^n \qquad (\tilde{c}_0\ne 0),$$

where this series also converges in the interval $|x|<r$. *In the special case* $\rho_1-\rho_2=$ *integer, there may not be a solution of the form (18) - see Exercise 19c. Notice: although the power series do converge at* $x=0$, *the functions* $u_1(x)$ *and* $u_2(x)$ *may not be solutions at that point because the functions* x^ρ *may not be twice differentiable (for example, if* $\rho=\frac{1}{2}$ *then* $\sqrt{x}$ *has no derivatives at* $x=0$*).*

Outline of Proof. Like Theorem 2, this proof also has two parts; i) finding the coefficients c_n for the formal power series, and ii) proving the formal power series converges. As in Theorem 3, part i) is proved by exhibiting formulas for the c_n's, while part ii) is proved by comparing the series $\sum c_nx^n$ with another convergent series $\sum C_nx^n$ whose coefficients are larger, $|c_n|\le C_n$.

To illustrate the procedure of part i), we will obtain the stated formula for the indicial polynomial $q(\rho)$. Let $A(x)=\sum_{n=0}^{\infty}\alpha_nx^n$ and $B(x)=\sum_{n=0}^{\infty}\beta_nx^n$ be the power series expansions of $A(x)$ and $B(x)$. Then assuming $u(x)$ has a solution in the form (18), we find by substituting these formulas into the differential equation that

$$\sum_{n=0}^{\infty}(\rho+n)(\rho+n-1)c_nx^{\rho+n}+(\sum_{n=0}^{\infty}\alpha_nx^n)(\sum_{n=0}^{\infty}(\rho+n)c_nx^{\rho+n})$$
$$+(\sum_{n=0}^{\infty}\beta_nx^n)(\sum_{n=0}^{\infty}c_nx^{\rho+n})=0.$$

The lowest power of x appearing is x^ρ, then comes $x^{\rho+1},\ldots$.

$$\begin{aligned}
x^\rho: &\quad \rho(\rho-1)c_0+\alpha_0\rho c_0+\beta_0c_0=0\\
x^{\rho+1}: &\quad (\rho+1)\rho c_1+[\alpha_1\rho c_0+\alpha_0(\rho+1)c_1]+[\beta_1c_0+\beta_0c_1]=0\\
&\cdot\\
&\cdot\\
&\cdot\\
x^{\rho+n}: &\quad (\rho+n)(\rho+n-1)c_n+\sum_{k=0}^{n}\alpha_{n-k}[(\rho+k)c_k]+\sum_{k=0}^{n}\beta_{n-k}c_k=0,
\end{aligned}$$

the last formula arising from the formula for the coefficients in the product of two power series (p. 76). If $c_0\ne 0$, the first equation states

$$q(\rho):=\rho(\rho-1)+\alpha_0\rho+\beta_0=0,$$

where $q(\rho)$ is the indicial polynomial. Since $\alpha_0=A(0)$ and $\beta_0=B(0)$, this is precisely the formula given in the theorem.

c) General Theory

We begin immediately by stating

Theorem 6.4 *(Existence and Uniqueness). Consider the second order linear O.D.E.*

$$Lu := a_2(x)u'' + a_1(x)u' + a_0(x)u = f(x),$$

where the coefficients a_0, a_1, and a_2 as well as f are continuous functions, and $a_2(x) \neq 0$. There exists a unique twice differentiable function $u(x)$ which satisfies the equation and the initial conditions

$$u(x_0) = \alpha, \qquad u'(x_0) = \beta,$$

where α and β are arbitrary constants.

If time permits, the existence proof will be carried out in the last chapter as a special case of a more general result. The uniqueness will be proved later too, as a special case of Theorem 9, page 510 - in the next section. We will not be guilty of circular reasoning.

Now what? Although this theorem appears to make further study unnecessary, there are several general statements which can be made because the equation is *linear.* Two other theorems are particularly nice; the first is $\dim \mathcal{N}(L) = 2$, while the second gives a procedure for solving the inhomogeneous equation once two linearly independent solutions of the homogeneous equation are known.

A preliminary result on linear dependence and independence of functions is needed. If the differentiable functions $u_1(x)$ and $u_2(x)$ are linearly dependent, there are constants c_1 and c_2 not both zero such that

$$c_1u_1(x) + c_2u_2(x) \equiv 0.$$

Differentiating this equation, we find

$$c_1u_1'(x) + c_2u_2'(x) \equiv 0.$$

Since the two homogeneous algebraic equations for c_1 and c_2 have a non-trivial solution, by Theorem 32 (page 428), the determinant

$$W(x) := W(u_1, u_2)(x) := \begin{vmatrix} u_1(x) & u_2(x) \\ u_1'(x) & u_2'(x) \end{vmatrix} = 0$$

must vanish. This determinant is called the *Wronskian* of u_1 and u_2. We have proved

Theorem 6.5 . *If the differentiable functions $u_1(x), u_2(x)$ are linearly dependent in the interval $[\alpha, \beta]$, then necessarily $W(x) \equiv 0$ throughout $[\alpha, \beta]$. Thus, if $W \neq 0$, the u_j 's are independent.*

Remark: The condition $W = 0$ is necessary for linear dependence but *not sufficient* in general, as can be seen from the example

$$u_1(x) = \begin{cases} x^2 & , \ x \geq 0, \\ 0 & , \ x < 0 \end{cases} \qquad u_2(x) = \begin{cases} 0 & , \ x \geq 0 \\ x^2 & , \ x < 0, \end{cases}$$

for which $W(u_1, u_2) \equiv 0$ for all x but u_1 and u_2 are linearly independent. However it *is* sufficient if u_1 and u_2 are solutions of a second order linear O.D.E., $Lu_j = 0$. An even stronger statement is true in this case. All we need require is that W vanish at one point x_0.

Theorem 6.6 . *Let u_1 and u_2 both be solutions of*

$$Lu := a_2u'' + a_1u' + a_0u = 0,$$

where $a_2 \neq 0$. If $W(x_0) = 0$ at some point x_0, then u_1 and u_2 are linearly dependent - which implies by Theorem 6 that $W(x) \equiv 0$ for all x. In other words, if $W(x_0) \neq 0$, then u_1 and u_2 are linearly independent.

Proof: Since $W(x_0) = 0$, the homogeneous algebraic equations

$$c_1u_1(x_0) + c_2u_2(x_0) = 0$$

$$c_1u_1'(x_0) + c_2u_2'(x_0) = 0$$

have a non-trivial solution c_1, c_2. Let

$$v(x) = c_1u_1(x) + c_2u_2(x).$$

We went to prove $v(x) \equiv 0$. Observe $Lv = 0$. Moreover $v(x_0) = 0$ and $v'(x_0) = 0$. Thus by uniqueness, $v(x) \equiv 0$, establishing the linear dependence of u_1 and u_2.

The same type of reasoning proves

Theorem 6.7 . *Let $Lu := a_2u'' + a_1u' + a_0u$, where $a_2(x) \neq 0$. Then*

$$\dim \mathcal{N}(L) = 2.$$

Proof: We exhibit two special solutions ϕ_1 and ϕ_2 of $Lu = 0$ and prove they constitute a basis for $\mathcal{N}(L)$. Let

$$\phi_1(x) \quad \text{satisfy} \quad L\phi_1 = 0 \quad \text{with} \quad \phi_1(x_0) = 1,\ \phi_1'(x_0) = 0$$

$$\phi_2(x) \quad \text{satisfy} \quad L\phi_2 = 0 \quad \text{with} \quad \phi_2(x_0) = 0,\ \phi_2'(x_0) = 1.$$

There are such functions by the existence theorem.

i) They are linearly independent.

$$W(x_0) = W(\phi_1, \phi_2)(x_0) = \begin{vmatrix} \phi_1(x_0) & \phi_2(x_0) \\ \phi_1'(x_0) & \phi_2'(x_0) \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1 \neq 0.$$

Thus by Theorem 7, ϕ_1 and ϕ_2 are linearly independent.

ii) They span $\mathcal{N}(L)$. Let $u(x)$ be any element in $\mathcal{N}(L)$ and consider the function

$$v(x) = u(x) - [u(x_0)\phi_1(x) + u'(x_0)\phi_2(x)].$$

Then $Lv = 0$ and $v(x_0) = 0, v'(x_0) = 0$. By uniqueness, $v(x) \equiv 0$. Thus every $u \in \mathcal{N}(L)$ can be written as

$$u(x) = A\phi_1(x) + B\phi_2(x),$$

where the constants A and B are $A = u(x_0), B = u'(x_0)$.

All of our attention has been on the homogeneous equation $Lu = 0$. Let us solve the inhomogeneous equation. This is particularly simple for a linear differential equation once we have a basis for $\mathcal{N}(L)$.

Theorem 6.8 *(Lagrange). Let $u_1(x)$ and $u_2(x)$ be a basis for $\mathcal{N}(L)$, where $Lu := a_2(x)u'' + a_1(x)u' + a_0(x)u$, with $a_2 \neq 0$. Then the inhomogeneous equation $Lu = f$ has the particular solution*

$$u_p(x) = u_1(x)\int^x \frac{W_1(s)}{W(s)} f(s)\,ds + u_2(x)\int^x \frac{W_2(s)}{W(s)} f(s)\,ds,$$

where $W(s) := W(u_1, u_2)(s)$ and $W_j(s)$ is obtained from $W(s)$ by replacing the j th column (u_j, u_j') of W by the vector $(0, 1/a_2)$.

Remark: If we let

$$G(x; s) = \frac{u_1(x)W_1(s) + u_2(x)W_2(s)}{W(s)}$$

then the above formula assumes the elegant form

$$u_p(x) = \int^x G(x; s)f(s)\,ds.$$

Proof: A device (due to Lagrange) called *variation of parameters* is needed. We already used a form of this device to solve the inhomogeneous first order linear equation (5, p. 457). The trick is to let

$$u_p(x) = v_1(x)u_1(x) + v_2(x)u_2(x)$$

where the functions $v_1(x)$ and $v_2(x)$ are to be found. This attempt to find u_p is reminiscent of writing the general solution of the homogeneous equation as $Au_1 + Bu_2$. Differentiate:

$$u_p'(x) = v_1u_1' + v_2u_2' + [v_1'u_1 + v_2'u_2].$$

The functions v_1 and v_2 will be chosen to make

$$v_1'u_1 + v_2'u_2 = 0.$$

Using this, we differentiate again

$$u_p''(x) = v_1u_1'' = v_2u_2'' + [v_1'u_1' + v_2'u_2']$$

Now multiply u_p'' by a_2, u_p' by a_1, u_p by a_0, and add to find

$$Lu_p = v_1Lu_1 + v_2Lu_2 + a_2[v_1'u_1' + v_2'u_2']$$

$$= a_2[v_1'u_1' + v_2'u_2'].$$

If we can choose v_1 and v_2 so that $a_2[\qquad] = f$, then indeed $Lu_p = f$, so $u_0 = v_1u_1 + v_2u_2$ is a particular solution. It remains to see if v_1 and v_2 can be found which satisfy the two needed conditions

$$v_1'u_1 + v_2'u_2 = 0$$

$$v_1'u_1' + v_2'u_2' = \frac{f}{a_2}.$$

These two linear equations for v_1' and v_2' may be solved by Cramer's rule (Theorem 33, page 429),

$$v_1' = \frac{\begin{vmatrix} 0 & u_2 \\ f/a_2 & u_2' \end{vmatrix}}{W} = \frac{f\begin{vmatrix} 0 & u_2 \\ 1/a_2 & u_2' \end{vmatrix}}{W} = \frac{W_1}{W}f$$

$$v_2' = \frac{\begin{vmatrix} u_1 & 0 \\ u_1' & f/a_2 \end{vmatrix}}{W} = \frac{f\begin{vmatrix} u_1 & 0 \\ u_1' & 1/a_2 \end{vmatrix}}{W} = \frac{W_2}{W}f$$

Integration of these equations yields v_1 and v_2, which, when substituted into $u_p = u_1v_1 + u_2v_2$, do give the stated result

With this theorem, knowing the general solution of the homogeneous equation $L\tilde{u} = 0$ allows us to find a particular solution of the homogeneous equation $Lu_p = f$. The general solution u of the inhomogeneous equation $Lu = f$ is then the u_p coset of $\mathcal{N}(L)$, that is, all functions of the form

$$u = u_p + \tilde{u}.$$

This puts the burden on finding the general solution of the homogeneous equation.

Examples:

(1) . The homogeneous equation $x^2u'' - 3xu' + 3u = 0,\ x \neq 0$, has the two linearly independent solutions $u_1(x) = x$, $u_2(x) = x^3$—which might have been found by the power series method. Therefore a particular solution of the inhomogeneous equation

$$x^2u'' - 3xu' + 3u = 2x^4$$

can be found by the variation of parameters. We try

$$u_p = v_1x^3 + v_2x$$

and are led to the equations

$$v_1' = \frac{\frac{-2x^4}{x^2}x^3}{2x^3}, \quad v_2' = \frac{\frac{2x^4}{x^2}x}{2x^3}$$

or

$$v_1' = -x^2, \qquad v_2' = 1.$$

Thus

$$v_1(x) = -\frac{x^3}{3}, \qquad v_2(x) = x.$$

Therefore

$$u_p(x) = x(-\frac{x^3}{3}) + x^3(x) = \frac{2}{3}x^4$$

The general solution to the inhomogeneous equation is found by adding the general solution of the homogeneous equation to this particular solution,

$$u(x) = Ax + Bx^3 + \frac{2}{3}x^4.$$

(2) The homogeneous equation $u''+u=0$ has the linearly independent solutions $u_1(x) = \cos x$, $u_2(x) = \sin x$. Let us solve

$$u'' + u = f(x),$$

where f is an arbitrary continuous function. Trying

$$u_p(x) = v_1 \cos x + v_2 \sin x,$$

we are led to

$$v_1' = \frac{-f\sin x}{1}, \qquad v_2' = \frac{f\cos x}{1}.$$

Thus

$$v_1(x) = -\int^x f(s)\sin\, s\, ds, \qquad v_2(x) = \int^x f(s)\cos\, s\, ds.$$

Therefore

$$\begin{aligned} u_p(x) &= -\cos x \int^x f(s)\sin s\, ds + \sin x \int^x f(s)\cos s\, ds \\ &= \int^x f(s)[-\sin s\cos x + \cos s \sin x]\, ds \\ &= \int^x f(s)\sin(x-s)\, ds. \end{aligned}$$

Consequently, the handsome formula

$$u(x) = A\sin x + B\cos x + \int^x f(s)\sin(x-s)\, ds$$

is the general solution of the inhomogeneous equation $u''+u=f$.

Exercises

(1) Solve the following initial value problems any way you can. Check your answers by substituting back into the differential equation.

(a) $u' + 2u = 0, \quad u(1) = 2$
(b) $u'' + 3u' + 2u = 7, \quad u(0) = 0,\ u'(0) = 0$
(c) $u'' + 3u' + 2u = 2e^x, \quad u(0) = 0,\ u'(0) = 1$
(d) $u'' + 3u' + 2u = e^{-2x}, \quad u(0) = 1,\ u'(0) = 0$
(e) $(\tan x)\frac{du}{dx} + u - \sin^2 x = 0, \quad u(\frac{\pi}{6}) = 1$
(f) $u'' + u = \tan x, \quad u(0) = u'(0) = 1,\ x\epsilon(-\frac{\pi}{2}, \frac{\pi}{2})$.
(g) $u''' - 8u = 0, \quad u(0) = 1,\ u'(0) = 2,\ u''(0) = 3$
(h) $u'''' - k^4 u = 0$. General solution.
(i) $u'' - 6u' + 10u = x^2 + \sin x, \quad u(0) = u'(0) = 0.$
(j) $u'''' - 7u''' - 8u'' = 0, \quad u(0) = 3,\ u'(0) = 8,\ u''(0) = 65,\ u'''(0) = 511.$
(k) $xu' + u = x^3, \quad u(1) = 1.$
(l) $u'' + 4u = 4x^2 + \cos 2x, \quad u(0) = 0,\ u'(0) = 1$
(m) $u''' - u' = e^x$. General solution.
(n) $u''' = 3u'' + 3u' - u = 0, \quad u(0) = 1,\ u'(0) = 2,\ u''(0) = 3$
(o) $u^{(5)} - u^{(4)} + 3u^{(3)} - 3u^{(2)} - 4u^{(1)} + 4u = 0$. General solution.

[Hint: $\lambda^5 - \lambda^4 + 3\lambda^3 - 3\lambda^2 - 4\lambda + 4 = (\lambda^2-1)(\lambda^2+4)(\lambda-1)$].

(2) Find the first four non-zero terms (if there are that many) in the power series solutions about $x=0$ for the following equations.

(a) $u'' - xu' - u = 0, \quad u(0) = u'(0) = 1$
(b) $u'' - 2xu' + 2u = 0, \quad u(0) = 0, u'(0) = 1.$
(c) $u'' - 2xu' - 2u = 0, \quad u(0) = 1,\ u'(0) = 0.$
(d) $u'' + xu = 0, \quad u(0) = 1,\ u'(0) = -1.$
(e) $u''' - xu = 0, \quad u(0) = 1, u'(0) = u''(0) = 0.$
(f) $u'' - x^2 u = \frac{1}{1-x^2}, \quad u(0) = 0,\ u'(0) = 0.$ [Hint: $\frac{1}{1-x^2} = 1 + x^2 + x^4 + \cdots$]
(g) $u'' - \frac{1}{1-x}u = 0, \quad u(0) = 0,\ u'(0) = 1.$ [Hint: $\frac{1}{1-x} = ?$]

(3) a) - e) Find where the power series in Ex. 2 a-e converge.

(4) Find the first four non-zero terms (if there are that many) in the power series solutions corresponding to the larger root of the indicial polynomial.

(a) $2x^2u'' - 3xu' + 2u = 0$

(b) $xu'' + 2u' - xu = 0$. [Answer: $u(x) = c_0 \sum_0^\infty \frac{x^{2n}}{(2n+1)!}$].

(c) $4xu'' + 2u' + u = 0$.

(d) $xu'' + (\sin x)u' + x^2 u = 0, \quad u(0) = 0,\ u'(0) = 1$.

(e) $xu'' + u' = x^2$.

(5) (a-e). Investigate the convergence of the series solutions found in Exercise 4 above.

(6) Find the power series solution about $x = 0$ for the nth order Bessel equation corresponding to the highest root of the indicial polynomial. The answer is:

$$J_n(x) = (\frac{x}{2})^n \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(k+n)!} (\frac{x}{2})^{2k},$$

where we have chosen $c_0 = 1/2^n n!$.

(7) Find two linearly independent power series solutions of

$$u'' + xu' + u = 0$$

and prove they are linearly independent. Find all solutions.

(8) The *Hermite equation* is

$$u'' - 2xu' + 2\alpha u = 0.$$

For which value(s) of the constant α are the solutions polynomials - that is, a solution with a *finite* Taylor series. These are the *Hermite polynomials*.

(9) Find the first three non-zero terms in the power series about $x = 0$ for two linearly independent solutions of

$$2x^2u'' + xu' + (x-1)u = 0.$$

(10) The homogeneous equation $Lu := 2x^2u'' - 3xu' - 2u = 0$ has the two linearly independent solutions $u_1(x) = x^2$, $u_2(x) = \sqrt{x}$ (see Ex. 20c below). Find the general solution of the inhomogeneous equation $Lu = \log(x^3)$.

(11) Let $Lu = (1-x^2)u'' - 2xu' + n(n+1)u$ where n is an integer. Show that $Lu = 0$ has a polynomial solution - the *Legendre polynomial*. Compute this for $n = 3$. (cf. page 104l Ex. 10).

(12) Let $J_0(x)$ be a solution of the zero$^{\text{th}}$ order Bessel equation. Prove $\frac{dJ_0}{dx}$ is a solution of the first order Bessel equation. [Hint: Work directly with the equation itself, *not* with power series].

(13) Consider the equation

$$a_2(x)u'' + a_1(x)u' + a_0(x)u = 0.$$

(a) Let $u(x) := u_1(x)v(x)$. Show that the result arranged as an equation for $v(x)$ is

$$a_2u_1v'' + (2a_2u_1' + a_1u_1)v' + (a_2u_1'' + a_1u_1' + a_0u_1)v = 0$$

(b) If u_1 is known to be one solution of the equation, show that the second solution is $u_2(x)$

$$u_2(x) = u_1(x) \int w(x)\, dx$$

where $w(x)$ is a solution of the *first* order equation

$$a_2u_1w' + (2a_2u_1' + a_1u_1)w = 0.$$

Thus, if one solution of a second order linear O.D.E. is known, the problem of finding a second solution is reduced to the problem of solving a first order linear O.D.E. - which can always be solved by separation of variables.

(14) Apply Exercise 13 to the following:

(a) One solution of $2x^2u'' - 3xu' + 2u = 0$ is $u_1(x) = x^2$. Find another.

(b) One solution of $x^2u'' - xu' + u = 0$ is $u_1(x) = x$. Find another.

(c) One solution of $(1+x)xu'' - xu' + u = 0$ is $u_1(x) = x$. Find another, and then write down the general solution.

(d) One solution of the equation $x^2u'' + 2xu' = 0$ is clearly $u_1(x) = 1$. Find another. Prove the solutions are linearly independent for $x > 0$. Find the general solution of $x^2u'' + 2xu' = 1$.

(15) Consider the O.D.E. $u'' + a(x)u' + b(x)u = 0$, where a and b are continuous about x_0. If the graphs of two solutions are tangent at $x = x_0$, are these two solutions linearly dependent? Explain: Can you make an even stronger deduction?

(16) (a) Let L be a constant coefficient differential operator with characteristic polynomial $p(\lambda)$. If $p(\lambda) = p(-\lambda)$, prove

$$L(\sin kx) = p(ik) \sin kx$$

(b) Apply this to find a particular solution of $u'''' - u = \sin 2x$

(17) Find a particular solution of the equation

$$u'' - n^2u = f, \quad n \neq 0.$$

[You will need: $\sin h(\alpha - \beta) = \sin h\, \alpha \cos h\, \beta - \sin h\, \beta \cos h\, \alpha$].

[Answer: $u(x) = \frac{1}{n} \int_0^x f(s) \sin h\, n(x-s)\, ds$.]

(18) Use the method of variation of parameters to find a particular solution to $u'' = f$. Compare with Exercise 5, p. 282.

(19) Consider the differential operator

$$Lu := x^2u'' + axu' + bu,$$

where a and b are constants. This is called *Euler's equation*. It is the simplest equation with a regular singularity at $x = 0$.

(a) Show that $Lx^\rho = q(\rho)x^\rho$, where $q(\rho)$ is the indicial polynomial for L.

(b) If the roots of $q(\rho) = 0$ are distinct, find two solutions of $Lu = 0$, $x > 0$, and prove the solutions are linearly independent for $x > 0$.

(c) If the roots ρ_1 and ρ_2 of $q(\rho) = 0$ coincide, take the derivative with respect to ρ of the equation in a) - holding x fixed - to obtain the candidate $u_2(x) = x^{\rho_1}\ln x$ for a second solution. Verify by substitution that u_2 is a solution in this case and prove the two solutions

$$u_1(x) = x^{\rho_1},\ u_2(x) = x^{\rho_1}\ln x, \quad x > 0$$

are linearly independent for $x \neq 0$.

(20) Apply the method of Exercise 19 to find two linearly independent solutions for each of the following Euler equations

a). $x^2u'' + xu' = 0$.

b). $2x^2u'' - 3xu' + 2u = 0$.

c). $2x^2u'' - 3xu' - 2u = 0$.

d). $x^2u'' - xu' + u = 0$.

(21) (a) Use the result of Ex. 19 a) to find a particular solution of the equation $Lu = x^\alpha$, where

$$Lu := x^2u'' + axu' + bu,$$

with a and b constant, and where α is *not* a root of the indicial polynomial $q(\rho)$ (cf. Ex. 6, p. 300).

(b) If neither α not β are roots of $q(\rho)$, find a particular solution to the inhomogeneous equation

$$Lu = Ax^\alpha + Bx^\beta.$$

(c) Apply this procedure to find the general solution of

$$2x^2u'' - 3xu' - 2u = 3x - 4x^{1/3}.$$

(d) How can you solve $Lu = x^\alpha$ if α is a root of the indicial polynomial?

(22) (a) If u has n derivatives and λ is a constant, prove

$$D^n[e^{\lambda x}u] = e^{\lambda x}(D + \lambda I)^n u.$$

Thus $(D + \lambda I)^n u = e^{-\lambda x}D^n[e^{\lambda x}u]$.

(b) Let $L = (D-a)^n$ be a constant coefficient differential operator with characteristic polynomial $p(\lambda) = (\lambda - a)^n$. Show $u(x)$ is a solution of the equation $Lu = 0$ if and only if $u(x)$ has the form

$$u(x) = e^{ax}Q(x),$$

where $Q(x)$ is a polynomial of degree $\leq n - 1$.

(23) Consider the O.D.E. $Lu = f$, where L is a second order *constant* coefficient operator, and let λ_1 and λ_2 be the characteristic roots of L_1. Assume i) $Re\lambda_1 < 0$ and $Re\lambda_2 < 0$, and ii)there is some constant M such that $|f(x)| \leq M$ for all $x \in [0, \infty]$.

(a) Prove every solution of $Lu = f$ is bounded for $x \in [0, \infty]$.

(b) If $\lim_{x\to\infty} f(x) = 0$, prove that as $x \to \infty$, every solution of $Lu = f$ tends to zero.

(24) Consider the operator $Lu := a_2(x)u'' + a_1(x)u' + a_0(x)u$, where the a_j's are continuous for $x \in [\alpha, \beta]$. Let u_1, u_2 and ϕ_1, ϕ_2 both be bases for $\mathcal{N}(L)$. Prove there is a constant $k \neq 0$ such that

$$W(u_1, u_2)(x) = kW(\phi_1, \phi_2)(x) \quad \text{for all} \quad x \in [\alpha, \beta].$$

(25) (a) Generalize the procedure of Ex. 21b and show how the inhomogeneous Euler equation $Lu = f$ can be solved if f has a power series expansion. You will have to assume that no root of the indicial polynomial is a positive integer.

(b) Apply a) to find a particular solution (as a power series) of

$$2x^2u'' + 3xu' - u = \frac{1}{1-x}.$$

(26) Given the equation $Lu := u'' + a(x)u' + b(x)u = 0$ has solutions $u_1(x) = \sin x$, $u_2(x) = \tan x$, find the general solution of the inhomogeneous equation

$$Lu = \frac{\cos x}{1 + \sin^2 x}.$$

(27) (a) If $Lu := a_2u'' + a_1u' + a_0u$ and $L^*v := (a_2v)'' - (a_1v)' + a_0v$, prove the *Lagrange identity*

$$vLu - uL^*v = \frac{d}{dx}[a_2(u'v - v'u) + (a_1 - a_2')uv],$$

where the functions a_j are assumed to be sufficiently differentiable. The operator L^* is the *adjoint* of L.

(b) Show that L is *self-adjoint*, $L = L^*$, if and only if $a_2' = a_1$. Write the Lagrange identity in this case.

(c) If $c_1u_1(x) + c_2u_2(x)$ is the general solution of the equation $Lu = 0$ find the general solution of the adjoint equation $L^*v = 0$. [Answer: $v = \dfrac{c_3u_1 + c_4u_2}{u_1u_2' - u_1'u_2}$].

(d) Let u be a twice differentiable function which vanishes at α and β. Show the adjoint operator L^* has the property that for all such functions u and v,

$$\langle v,\, Lu\rangle = \langle L^*v,\, u\rangle$$

where

$$\langle f,\, g\rangle := \int_\alpha^\beta f(x)g(x)\,dx.$$

(28) (a) Let L be a *self-adjoint* operator, $L = L^*$. If $LX_1 = \lambda_1X_1$ and $LX_2 = \lambda_2X_2$, where λ_1 and λ_2 are real number, $\lambda_1 \neq \lambda_2$, prove X_1 and X_2 are orthogonal

$$\langle X_1,\, X_2\rangle = 0.$$

[Hint: Compare $\langle X_2,\, LX_1\rangle = \lambda_1\langle X_2,\, X_1\rangle$ with $\langle LX_2,\, X_1\rangle = \lambda_2\langle X_2,\, X_1\rangle$].

(b) Let $L = \frac{d^2}{dx^2}$. For what values of λ can you find a non-zero solution u of the equation $Lu = \lambda u$ where u satisfies the boundary conditions $u(0) = u(\pi) = 0$?

(c) Apply parts a) and b) as well a Ex. 27d to prove

$$\langle \sin nx,\, \sin mx\rangle = \int_0^\pi \sin nx \sin mx\,dx = 0,$$

where n and m are unequal integers.

(29) . Consider the *boundary value problem*

$$Lu := u'' + u = f, \qquad u(0) = 0,\, u(\pi) = 0,$$

where f is continuous in $[0, \pi]$.

a). Show that if a solution exists, it is not unique.

b). Show a solution exists if and only if

$$\int_0^\pi f(x)\sin x\,dx = 0.$$

[Hint: First find the general solution of the homogeneous equation].

Remark: In the notation of Ex. 27, we have $L = L^*$. Moreover, $\mathcal{N}(L^*) = \operatorname{span}\{\sin x\}$. The conclusions of b) states that $\mathcal{R}(L) = \mathcal{N}(L^*)^\perp$, and illustrates how Theorem 34, p. 431, is used in infinite dimensional spaces.

(30) . A proof of Theorem 3. Since $a_2(x) \neq 0$, the equation can be written as

$$u'' + a(x)u' + b(x)u = 0.$$

If

$$a(x) = \sum_{n=0}^\infty \alpha_n x^2, \qquad b(x) = \sum_{n=0}^\infty \beta_n x^n,$$

let

$$u(x) = \sum_{n=0}^\infty c_n x^n, \quad \text{where} \quad u(0) = c_0,\, u'(0) = c_1,$$

(a) Imitate the example to prove the remaining c_n's must satisfy

$$c_{n+2} = -\sum_{k=0}^n \frac{[\alpha_{n-k}(k+1)c_{k+1} + \beta_{n-k}c_k]}{(n+2)(n+1)}.$$

Show that if c_0 and c_1 are known, then the remaining c_n's are determined inductively by the above formula.

(b) Because the series for $a(x)$ and $b(x)$ converge for $|x| < r$, if R is any number less than r, there is a constant M such that for all n, $|\alpha_n| \le \frac{M}{R^n}$ and $|\beta_n| \le \frac{M}{R^n}$ (cf. p. 72, line 2). Define constants C_n as

$$C_0 = |c_0|\,,\, C_1 = |c_1|\,,$$

and for $n \ge 0$

$$C_{n+2} = \frac{\frac{M}{R^n}\sum_{k=0}^n [(k+1)C_{k+1} + C_k]R^k + MC_{n+1}R}{(n+2)(n+1)}.$$

(i) Prove $|c_n| \le C_n, \quad n = 0, 1, 2, 3, \ldots$

(ii) Prove

$$\left|\frac{C_{n+1}x^{n+1}}{C_nx^n}\right| = \frac{n(n-1) + MnR + MR^2}{R(n+1)n}\,|x|\,.$$

(iii) Prove $\sum_{n=0}^\infty C_nx^n$ converges for $|x| < R$, where R is any number less than r.

(iv) Prove $\sum_{n=0}^\infty c_nx^n$ converges for $|x| < R$, where R is any number less than r.

(31) .

(a) Let $u(x)$ and $v(x)$ be solutions of the equations $L_1 u := u'' + a(x)u = 0$, and $L_2 v := v'' + b(x)v = 0$ respectively, in some interval, where a and b are continuous. If $b(x) \geq a(x)$ throughout the interval, prove there must be a zero of v between any two zeroes of u. This is the *Sturm oscillation theorem.* [Hint: Suppose α and β are consecutive zeroes of u and $u > 0$ in (α, β). Prove

$$0 = \int_\alpha^\beta (vL_1 u - uL_2 v)\, dx = vu'\Big|_\alpha^\beta - \int_\alpha^\beta (b-a)uv\, dx,$$

and show, because $u'(\alpha) > 0$, $u'(\beta) < 0$, there is a contradiction if v does not vanish somewhere in (α, β).]

(b) Let $u_1(x)$ and $u_2(x)$ be two linearly independent solutions of $u'' + a(x)u = 0$. Prove between any two zeroes of u_1, there is a zero of u_2 and vice verse. Thus, the zeroes interlace.

(c) Apply b) to the solutions $\sin \gamma x$ and $\cos \gamma x$ of the equation $u'' + \gamma^2 u = 0$ to conclude a well-known fact.

(d) If $b(x) \geq \delta > 0$, where δ is a constant, prove every solution of $v'' + b(x)v = 0$ must have an infinite number of zeros by comparing v with a solution of $u'' + \gamma^2 u = 0$, where γ is an appropriate constant.

(e) Apply d) to prove every solution of

$$v'' + (1 - \frac{3}{4x^2})v = 0,$$

has an infinite number of zeroes for $x \geq 1$.

(f) Let $u_1(x)$ be a solution of the first order Bessel equation. Take $v(x) = u_1(x)\sqrt{x}$ and show that v satisfies the equation in e). Deduce that $J_1(x)$ has infinitely many zeroes.

(32) Let L_1 and L_2 be linear constant coefficient differential operators with characteristic polynomials $p_1(\lambda)$ and $p_2(\lambda)$ respectively.

(a) If there is a function $u(x)$, $u(x) \not\equiv 0$, which satisfies both $L_1 u = 0$ and $L_2 u = 0$, prove the polynomials p_1 and p_2 have a common root.

(b) If p_1 and p_2 have no common roots, prove the solution of $L_1 L_2 u = 0$ are exactly all functions of the form $c_1 u_1 + c_2 u_2$ where u_1 is a solution of $L_1 u_1 = 0$, and u_2 of $L_2 u_2 = 0$. Thus $\mathcal{N}(L_1 L_2)$ may be decomposed into the two complementary subspaces $\mathcal{N}(L_1)$ and $\mathcal{N}(L_2)$, $\mathcal{N}(L_1 L_2) = \mathcal{N}(L_1) \oplus \mathcal{N}(L_2)$.

(33) Imitate Exercise 30 and prove Theorem 3. Make sure to observe the trouble in trying to find the solution corresponding to the lower root of the indicial polynomial if the roots differ by an integer.

(34) The purpose of this exercise is to show that an equation with an *irregular* singular point may have a *formal* power series at that point which does *not converge* to the solution.

Try to find a solution of the form (18) for the following equation which has an irregular singularity at $x = 0$,

$$x^6 u'' + 3x^5 u' - 4u = 0.$$

What happened? Two linearly independent solutions for $x \neq 0$ are

$$u_1(x) = e^{-1/x^2} \quad \text{and} \quad u_2(x) = e^{1/x^2}.$$

How does this explain the situation (cf. p. 95-6)?

(35) Consider the equation $2x^2 u'' + 3xu' + u = \sqrt(x)$. Two linearly independent solutions of the homogeneous equation are $x^{-1/2}$ and x^{-1}. Find the general solution of the homogeneous equation.

(36) Consider the equation $u'' + b(x)u' + c(x)u = 0$, where b and c are continuous functions and $c(x) < 0$. Prove that a solution cannot have a positive maximum or negative minimum.

6.4 First Order Linear Systems

Quite often in applications you must consider *systems* of differential equations. We shall consider a *linear system* of the form

$$\frac{du_1}{dx} + a_{11}(x)u_1 + a_{12}(x)u_2 + \cdots + a_{1n}(x)u_n = f_1(x) \tag{6-22}$$

$$\frac{du_2}{dx} + a_{21}(x)u_1 + a_{22}(x)u_2 + \cdots + a_{2n}(x)u_n = f_2(x) \tag{6-23}$$

$$\vdots \qquad\qquad \vdots \tag{6-24}$$

$$\frac{du_n}{dx} + a_{n1}(x)u_1 + a_{n2}(x)u_2 + \cdots + a_{nn}(x)u_n = f_n(x)\ , \tag{6-25}$$

where the functions $a_{ij}(x)$ and $f_j(x)$ are continuous. If we anticipate the next chapter and write the derivative of a vector $U = (u_1, \ldots, u_n)$ as the derivative of its components,

$$\frac{d}{dx}U(x) = \left(\frac{du_1}{dx}, \frac{du_2}{dx}, \cdots, \frac{du_n}{dx}\right),$$

then the above system can be written in the clean form

$$\frac{dU}{dx} + A(x)U = F(x), \tag{6-26}$$

where,

$$A(x) = ((a_{ij})), \qquad F = (f_1, f_2, \ldots, f_n)$$

and

$$U(x) = (u_1, u_2, \ldots, u_n).$$

The initial value problem for the system of differential equations (22) is to find a vector $U(x)$ which satisfies the equation as well as the initial condition

$$U(x_0) = U_0, \tag{6-27}$$

where U_0 is a vector of constants.

It is useful to observe that the initial value problem for a single linear equation of order n

$$u^{(n)} + a_{n-1}(x)u^{(n-1)} + \cdots + a_0(x)u = f(x)$$
$$u(x_0) = \alpha_1,\, u'(x_0) = \alpha_2, \ldots, u^{(n-1)}(x_0) = \alpha_n,$$

can be transformed to the conceptually simpler problem (22)-(23). Let $u_1(x) := u(x)$, $u_2(x) := u'(x), \ldots$, and $u_n(x) = u^{(n-1)}(x)$. Then the components of the vector $U(x) = (u_1, u_2, \ldots, u_n)$ must obviously satisfy the relations

$$\begin{aligned}
\frac{du_1}{dx} &= u_2 \\
\frac{du_2}{dx} &= u_3 \\
&\cdot \\
&\cdot \\
&\cdot \\
\frac{du_{n-1}}{dx} &= u_n \\
\frac{du_n}{dx} &= -a_0u_1 - a_1u_2 - \cdots - a_{n-1}u_n + f(x),
\end{aligned}$$

which may be written as

$$U' = MU + F,$$

where

$$M(x) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & \vdots & & \vdots & \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{pmatrix},$$

and

$$F = (0, 0, \ldots, 0, f).$$

The initial conditions read

$$U(x_0) = (\alpha_1, \alpha_2, \ldots, \alpha_n).$$

Conversely, if U is any solution of this system of equations with the proper initial conditions, then the first component $u_1(x)$ is a solution of the single nth order equation. Thus, the general theory of a single nth order linear O.D.E. is completely subsumed as a portion of the theory of a system of first order linear O.D.E.'s. You should be warned that this generalization is mainly of theoretical value and is of little use if you are seeking an explicit solution.

Both the existence and uniqueness theorems are true for systems, and supply an example where the theoretical advantages of systems become clear. To illustrate this, we shall prove the uniqueness theorem. Our proof is patterned directly after the uniqueness proof for a single equation (Theorem 1).

Theorem 6.9 *(Uniqueness). Let $A(x)$ be a matrix whose coefficients $a_{ij}(x)$ are bounded $|a_{ij}(x)| \le M$ for x in some interval, and let $F(x)$ be a continuous function. Then there is at most one solution $U(x)$ of the initial value problem*

$$U' + AU = F, \qquad U(x_0) = U_0.$$

REMARK: The existence theorem states, if A is nonsingular and each element is integrable there is at least one solution. Thus, there is then exactly one solution.

PROOF: Assume U_1 and U_2 are both solutions. Let

$$W = U_1 - U_2.$$

Then W satisfies the homogeneous equation and is zero at x_0,

$$W' + AW = 0, \quad W(x_0) = 0.$$

Take the scalar product of this with W,

$$\langle W, W'\rangle + \langle W, AW\rangle = 0.$$

But

$$\begin{aligned}
\langle W, W'\rangle &= w_1w_1' + ww_2' + \cdots + w_nw_n' \\
&= \frac{1}{2}\frac{d}{dx}(w_1^2 + w_2^2 + \cdots + w_n^2) \\
&= \frac{1}{2}\frac{d}{dx}\|W\|^2.
\end{aligned}$$

Thus,

$$\frac{1}{2}\frac{d}{dx}\|W\|^2 = -\langle W, AW\rangle.$$

By Theorem 17, p. 173 and the hypothesis $|a_{ij}(x)| \le M$, we know

$$\big|\langle W, AW\rangle\big| \le \Big[\sum_{i,j=1}^{n} |a_{ij}|^2\Big]^{1/2}\|W\|^2 \le nM\|W\|^2.$$

so that

$$\frac{1}{2}\frac{d}{dx}\|W\|^2 \le nM\|W\|^2.$$

Therefore, as on p. 462-3

$$\frac{d}{dx}(\|W\|^2) - 2nM\|W\|^2 \le 0,$$

or

$$e^{2nMx}\frac{d}{dx}[e^{-2nMx}\|W\|^2] \le 0.$$

Because e^{2nMx} is always positive, by the mean value theorem the quantity [ ] is a decreasing function. Its value for $x > x_0$ is then less than at x_0,

$$e^{-2nMx}\|W(x)\|^2 \le e^{-2nMx_0}\|W(x_0)\|^2, \qquad x \ge x_0$$

Consequently

$$\|W(x)\| \le e^{nM(x-x_0)}\|W(x_0)\|, \qquad x \ge x_0.$$

Since $W(x_0) = 0$ and the norm is non negative, we have

$$0 \le \|W(x)\| \le 0, \qquad x \ge x_0,$$

which implies

$$\|W(x)\| = 0, \qquad x \ge x_0.$$

Therefore,

$$W(x) \equiv 0 \qquad x \ge x_0.$$

By replacing x with $-x$ in the original equation, the same statement is true for $x \le x_0$. Thus, throughout the interval where $|a_{ij}(x)| \le M$, we have proved $W(x) \equiv 0$, that is, $U_1(x) \equiv U_2(x)$, so the solution is indeed unique.

Because a single linear nth order O.D.E. can be replaced by an equivalent system of equations, this theorem implies the uniqueness theorem for a single O.D.E. of order n if the coefficients $a_j(x)$ are bounded in some interval - which is certainly true in every interval if the a_j's are continuous.

With this theorem, a short section closes. Further developments in the theory of systems of linear O.D.E.'s make elegant use of linear operators in general and matrices in particular. As you might well accept, the exercises contain a few of the more accessible results.

Exercises

(1) . Find functions $u_1(x), u_2(x)$ which satisfy

$$u_1' = u_1$$

$$u_2' = u_1 - u_2,$$

with the initial conditions $U(0) := (u_1(0), u_2(0) = (1, 0)$. Find the general solution too. [Hint: Solve the equation $u_1' = u_1$ first, then substitute. Answer: General solution is $U(x) = (\gamma_1 e^x, \frac{\gamma_1}{2}e^x + \gamma_2 e^{-x})$].

(2) Consider the system

$$u_1' = 2u_1 - u_2$$

$$u_2' = 3u_1 - 2u_2,$$

that is,

$$U' = AU, \quad \text{where} \quad A = \begin{pmatrix} 2 & -1 \\ 3 & -2 \end{pmatrix}.$$

Let $\phi_1(x) = au_1 + bu_2$, $\phi_2(x) = cu_1 + du_2$, where a, b, c and d are constants. Thus,

$$\Phi = SU,$$

where

$$S = \begin{pmatrix} a & b \\ c & c \end{pmatrix}, \qquad \Phi = (\phi_1, \phi_2).$$

(a) By direct substitution, find the differential equations satisfied by the ϕ_j's and show they can be written as

$$\Phi' = SAS^{-1}\Phi.$$

(b) Pick the coefficients of S so the matrix SAS^{-1} is a diagonal matrix,

$$SAS^{-1} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \equiv \Lambda$$

(c) Solve the resulting equation $\Phi' = \Lambda\Phi$. [Solution: $\phi_1 = \alpha e^x$, $\phi_2 = \beta e^{-x}$—you might have ϕ_1 and ϕ_2 interchanged].

(d) Use this solution to solve the original equations for U. [HINT: Recall $U = S^{-1}\Phi$].

(3) By only a slight modification of Exercise 2, solve

$$v_1'' = 2v_1 - v_2$$

$$v_2'' = 3v_1 - 2v_2.$$

[Hint: Everything, even the algebra, is identical. The only difference is in part c) you have to solve $\Phi'' = \Lambda\Phi$. Then $V = S^{-1}\Phi$ as before].

(4) A bathtub initially contains Q_1 gallons of gin and Q_2 gallons of vermouth, where $Q_1 + Q_2 = Q$, Q being the capacity of the tub. Pure gin enters from one faucet at a constant rate of R_1 gallons per minute, while pure vermouth enters from another faucet at a constant rate R_2 gallons per minute. The well stirred mixture of martinis leaves the drain at a rate $R_1 + R_2$ gallons per minute (so the total amount of fluid in the tub remains constant at Q gallons). Let $G(t)$ be the quantity of gin in the tub at time t and $V(t)$ be the quantity of vermouth.

(a) Show

$$\frac{dG}{dt} = R_1 - \frac{G}{Q}(R_1 + R_2)$$
$$\frac{dV}{dt} = R_2 - \frac{V}{Q}(R_1 + R_2).$$

(b) Integrate this simple system of equations to find $G(t)$ and $V(t)$. Also find their ratio $P(t) := G(t)/V(t)$ which is the strength of the martinis at time t.

(c) Prove

$$\lim_{t\to\infty} P(t) = \frac{R_1}{R_2}.$$

Compare this with your intuitive expectations.

(d) If $Q_1 = 20, Q_2 = 0, R_1 = R_2 = 1$ gal/min, how long must I wait to get a perfect martini (for me, perfect is 5 parts gin to 1 part vermouth). [Needless to say, the mathematical model is applicable to many problems in the mixing of chemicals which do not react with each other. If the chemicals do interact, the model must be changed to account for the interaction].

(5) Consider the homogeneous equation $U' = A(x)U$, where A is non-singular (so $\det A \neq 0$). Assuming the validity of the existence theorem, prove there exists n linearly independent vectors $U_1(x), U_2(x), \ldots, U_n(x)$ which are solutions, $U_k' = AU_k$, $k = 1, \ldots, n$. [Hint: Construct n solutions which are linearly independent at $x = x_0$, and then prove a set of n solutions are linearly independent in an interval if and only if they are linearly independent at $x = x_0$, where x_0 is a point in the interval].

(6) Let $LU := U' - A(x)U$ as in Exercise 5. Prove $\dim \mathcal{N}(L) = n$.

(7) Let $LU := U' - A(x)U$. If a basis $U_1, \ldots, U_n$, for $\mathcal{N}(L)$ is known, prove the inhomogeneous equation $LU = F$ can be solved by variation of parameters. That is, seek a particular solution U_p of $LU = F$ in the form

$$U_p = \sum_{i=1}^{n} U_i v_i$$

where the $v_i(x)$ are scalar-valued functions (*not* vectors).

(a) Compute U_p' and substitute into the O.D.E. to conclude U_p is a particular solution if

$$\sum_{i=1}^{n} U_i v_i' = F.$$

(b) Let U be the $n \times n$ matrix whose columns are $U_1, U_2, \ldots, U_n$. Prove U is invertible and show

$$v_i'(x) = (U^{-1}F)_{\text{ith component}}.$$

(c) Show

$$U_p(x) = \sum_{i=1}^{n} U_i(x) \int^x [U^{-1}(s)F(s)]_i \, ds.$$

This may also be written in the form

$$U_p(x) = U(x) \int^x U^{-1}(s)F(s)\, ds$$

(d) Apply this procedure to find the general solution of

$$u_q' = u_1 + e^{2x} \qquad \text{cf. Ex 1}$$
$$u_2' = u_1 - u_2 + 1.$$

6.5 Translation Invariant Linear Operators

This section develops various extensions and applications of the procedure used to solve *linear* ordinary differential equations with *constant* coefficients. The results will be proved as a series of exercises interspersed by various remarks.

DEFINITION: The *translation operator* T_t acting on functions $u(x)$ is defined by the property

$$(T_t u)(x) = u(x - t). \qquad x, t \in \mathbb{R}.$$

A linear operator L is *translation invariant* if

$$LT_t = T_t L$$

for every t, that is, if

$$L(T_t u) = T_t(Lu)$$

for every t and for every function u for which the operators are defined.

EXAMPLE: 1 Let $(Lu)(x) := 3u(x) - 2u(x-1)$. Then

$$[T_t(Lu)](x) = 3u(x-t) - 2u(x-t-1),$$

and

$$[L(T_t u)](x) = Lu(x-t) = 3u(x-t) - 2u(x-t-1).$$

Thus,

$$LT_t = T_t L,$$

so the operator L is translation invariant.

2. Let $(Lu)(x) := 3xu(x)$. Then

$$[T_t(Lu)](x) = 3(x-t)u(x-t),$$

and
$$[L(T_t u)](x) = Lu(x-t) = 3xu(x-t).$$
Thus
$$LT_t \neq T_t L,$$
so this operator is *not* translation invariant.

Exercises

(1) Which of the following linear operators (verify!) are also translation invariant?

(a) $(Lu)(x) := cu(x), \qquad c \equiv constant$

(b) $(Lu)(x) := \frac{u(x+h)-u(x)}{h}, \quad h \equiv \text{constant} \neq 0.$

(c) $(Lu)(x) := \displaystyle\int_{-\infty}^{x} k(x-s)u(s)\,ds$

(d) $(Lu)(x) := (x-1)u(x)$

(e) $(Lu)(x) = \frac{du}{dx}(x).$

(f) Any linear ordinary differential operator with *constant* coefficients,
$$Lu := a_n u^{(n)} + a_{n-1}u^{(n-1)} + \cdots + a_0 u, \qquad a_k \quad \text{constants.}$$

(g) Any linear ordinary differential operator with *variable* coefficients.

(h) $(Lu)(x) = \displaystyle\sum_{k=1}^{n} a_k u(x-\gamma_k), \quad a_k$ and γ_k constants.

[Answers: All but d) and g) are translation invariant].

(2) If L_1 and L_2 are translation invariant operators which map some linear space into itself, then so are

a). $AL_1 + BL_2, \quad A, B$ constants

b). L_1L_2 and L_2L_1

c). If in addition L is invertible, then L^{-1} is also translation invariant.

Theorem 6.10 . *If L is a translation invariant linear operator, then*
$$L(e^{\lambda a}) = \phi(\lambda)e^{\lambda x}.$$

PROOF: We know so little about L that all we can hope to do is compute $T_tL(e^{\lambda x})$ and $LT_t(e^{\lambda x})$ and see what happens. Let $Le^{\lambda x} = \psi(\lambda; x)$, where ψ is some unknown function whose value depends on both λ and x. Then
$$T_tL(e^{\lambda x}) = \psi(\lambda; x-t),$$

while
$$\begin{aligned} LT_te^{\lambda x} - Le^{\lambda(x-t)} &= L(e^{-\lambda t}e^{\lambda x}) \\ &= e^{-\lambda t}Le^{\lambda x} = e^{-\lambda t}\psi(\lambda; x). \end{aligned}$$
Since $T_tL = LT_t$, we find
$$e^{-\lambda t}\psi(\lambda; x) = \psi(\lambda; x-t),$$
or
$$\psi(\lambda; x) = \psi(\lambda; x-t)e^{\lambda t}.$$
Because the left side does not contain t, the right side must not depend on which value of t is chosen. Using this freedom, we let $t = x$ and conclude
$$\psi(\lambda; x) = \psi(\lambda; 0)e^{\lambda x}.$$
By setting $\phi(\lambda) = \psi(\lambda, 0)$, we find
$$Le^{\lambda x} = \psi(\lambda; x) = \phi(\lambda)e^{\lambda x}$$
as desired.

Exercises

(3) By direct substitution, find $\phi(\lambda)$ for those operators in Exercise 1 which are translation invariant. [Answers: a) $\phi(\lambda) = c$, b) $\phi(\lambda) = (e^{-ah}-1)/h$ c) $\phi(\lambda) = \displaystyle\int_{-\infty}^{0} k(-s)e^{\lambda s}\,ds$, d) $\phi(\lambda) = c\lambda$, f) $\phi(\lambda) = \displaystyle\sum_{k=0}^{n} a_k\lambda^k$ (the characteristic polynomial), h) $\phi(\lambda) = \displaystyle\sum_{k=1}^{n} a_k e^{-\lambda\gamma}k$].

(4) With the same assumptions and notation as in the theorem, if $\phi(\lambda) = 0$ is a polynomial equation with N distinct roots $\lambda_1, \lambda_2, \ldots, \lambda_N$, so $\phi(\lambda_j) = 0$, $j = 1, \ldots, N$, prove any linear combination of the function $e^{\lambda jx}$ is in $\mathcal{N}(L)$, that is,
$$Lu = 0 \quad \text{where} \quad u(x) = \sum_{1}^{N} c_j e^{\lambda jx}.$$

(5) Apply the theorem to find the solution of Exercise 4 for the equation $Lu = 0$, where

(a) $Lu := u'' - u' - u$.

(b) $(Lu)(x) = u(x+2) - u(x+1) - u(x)$.

(c) Find a special solution of b) which satisfies the "initial conditions" $u(0) = u(1) = 1$. Compute $u(2), u(3)$ and $u(4)$ directly from b). The integers $u(n)$, $n \in \mathbb{Z}_+$ are called the *Fibonacci sequence*. [Answer: $u(2) = 2, u(3) = 3, u(4) = 5$, and *surprisingly*,

$$u(n) = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right].$$

(6) Solve $u(x) - au(x-1) + b^2u(x-2) = 0$ with the initial conditions $u(1) = a, u(2) = a^2 - b^2$. Compare with Exercise 17, p. 440.

(7) Extend Exercises 5(b - c) and 6 to develop a theory of *second order difference equations with constant coefficients*. Thus

$$Lu := a_2u(x+2) + a_1u(x+1) + a_0u(x), \qquad a_2 \neq 0, \qquad x \in \mathbb{Z}.$$

In particular, you should,

(a) Find two linearly independent solutions of $Lu = 0$. Remember the degenerate case $a_1^2 - 4a_0a_2 = 0$.

(b) Prove there is at most one solution of the initial value problem $Lu = f$, $u(0) = \alpha_0$, $u(1) = \alpha_1$.

(c) Prove dim $\mathcal{N}(L) = 2$.

REMARKS: The ideas presented above generalize immediately to the case where $X \in \mathbb{R}^n$ instead of just $\mathbb{R}^1$, as well as to the case where the u's are vectors and not scalars. These few concepts lie at the heart of any treatment of many linear operators with constant coefficients, especially ordinary and partial differential operators. This mildly abstract formulation manages to penetrate through the obscuring details of particular cases to observe a rather simple structure unifying many seemingly different problems.

6.6 A Linear Triatomic Molecule

A molecule composed of three atoms is called a *triatomic*. Consider a triatomic molecule whose *equilibrium configuration* is a straight line with two atoms of equal mass m situated on either side of a central atom of mass M.

A FIGURE GOES HERE

To simplify the situation further, we shall only consider the motion along the straight line (axis) of these atoms, and shall *assume* the inter-atomic forces can be *approximated* by springs with equal spring constants k. $u_1(t), u_2(t)$ and $u_3(t)$ will denote the displacements of the atoms (see fig.) from their equilibrium position.

Newton's second law, $m\ddot{u} = \sum F$, will give the equations of motion. The atom on the left only "feels" the force due to the spring attached to it, the force being equal to the spring constant k times the amount that spring is stretched, $u_2 - u_1$. Thus,

$$m\ddot{u}_1 = k(u_2 - u_1).$$

The central atom "feels" two forces, one from each side, with the resulting equation of motion

$$M\ddot{u}_2 = -k(u_2 - u_1) + k(u_3 - u_2).$$

In the same way, the equation of motion for the remaining atom is

$$m\ddot{u}_3 = -k(u_3 - u_2).$$

Collecting our equations, we have

$$\begin{aligned}
\ddot{u}_1 &= -\frac{k}{m}u_1 + \frac{k}{m}u_2 \\
\ddot{u}_2 &= \frac{k}{M}u_1 - \frac{2k}{M}u_2 + \frac{k}{M}u_3 \\
\ddot{u}_3 &= \frac{k}{m}u_2 - \frac{k}{m}u_3.
\end{aligned}$$

These are a *system* of three linear ordinary differential equations with constant coefficients. They cannot be integrated as they stand since each equation involves functions from the other equations, that is, the equations are *copied* (not surprising since we are considering *coupled oscillators*. Now we can integrate such a system immediately if they are in the simple form

$$\begin{aligned}
\ddot{\phi}_1 &= \lambda_1\phi_1 \\
\ddot{\phi}_2 &= \lambda_2\phi_2 \\
\ddot{\phi}_3 &= \lambda_3\phi_3
\end{aligned}$$

by integrating each equation separately. By using an important method, we will be able to place our system in this special form.

Before doing so, it is suggestive to rewrite the system in matrix form

$$\begin{pmatrix} \ddot{u}_1 \\ \ddot{u}_2 \\ \ddot{u}_3 \end{pmatrix} = \begin{pmatrix} -\frac{k}{m} & \frac{k}{m} & 0 \\ \frac{k}{M} & -\frac{2k}{M} & \frac{k}{M} \\ 0 & \frac{k}{m} & -\frac{k}{m} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}.$$

Letting A denote the 3×3 matrix, our hope is to somehow change A into a *diagonal matrix* (one with zeroes everywhere except along the principal diagonal), for then the differential equations will be in a form mentioned above which can be immediately integrated.

The trick is to replace the basis u_1, u_2, u_3 by some other basis in which the matrix assumes a diagonal form. The differential equation can be written in the form

$$\ddot{U} = AU,$$

where $U = (u_1, u_2, u_3)$, and the derivative of a vector being defined as the derivative of each of its components. Let $\phi_1(t), \phi_2(t)$, and $\phi_3(t)$ be three other functions - which we plan to use as a new basis. Then the ϕ_j's can be written as a linear combination of the u_j's,

$$\phi = s_{11}u_1 + s_{12}u_2 + s_{13}u_3$$

$$\phi_2 = s_{21}u_1 + s_{22}u_2 + s_{23}u_3$$

$$\phi_3 = s_{31}u_1 + s_{32}u_2 + s_{33}u_3,$$

where s_{ij} are *constants*. Writing $S = ((s_{ij}))$ and $\Phi = (\phi_1, \phi_2, \phi_3)$, this last equation reads

$$\Phi = SU.$$

Taking the derivative of both sides (or going back to the equations defining ϕ_j in terms of the u_k's), we find

$$\ddot{\Phi} = S\ddot{U}.$$

Because both u_1, u_2 and u_3 as well as ϕ_1, ϕ_2, and ϕ_3 are bases for the solution, the matrix S must be non-singular (its inverse expresses the $\phi_j's$ in terms of the u_j's). Thus

$$\ddot{\Phi} = SAS^{-1}\Phi.$$

The problem has been reduced to *finding a matrix S such that the matrix SAS^{-1} is a diagonal matrix,*

$$SAS^{-1} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \equiv \Lambda.$$

Multiply by S^{-1} on the left:

$$AS^{-1} = S^{-1}\Lambda.$$

Since this equation is equally between matrices, their corresponding columns must be equal. Thus, if we denote by $\hat{S}_i$, the ith column of S^{-1}, the above equation then reads

$$A\hat{S}_i = \lambda_i \hat{S}_i,$$

or

$$(A - \lambda_i I)\hat{S}_i = 0.$$

For each i this is a system of three linear algebraic equations for the three components of $\hat{S}_i$. If there is to be a non-trivial solution, we know

$$\det(A - \lambda_i I) = 0.$$

Since

$$\det(A - \lambda_i I) = \begin{vmatrix} -\frac{k}{m} - \lambda_i & \frac{k}{m} & 0 \\ \frac{k}{M} & -\frac{2k}{M} - \lambda_i & \frac{k}{M} \\ 0 & \frac{k}{m} & -\frac{k}{m} - \lambda_i \end{vmatrix},$$

(algebra later)

$$= -\lambda_i(\frac{k}{m} + \lambda_i)[\lambda_i + (\frac{2}{M} + \frac{1}{m})k]$$

We see the three possible values of λ for $\det(A - \lambda_i I) = 0$ are

$$\lambda_1 = 0, \lambda_2 = -\frac{k}{m}, \qquad \lambda_3 = -k(\frac{2}{M} + \frac{1}{m}).$$

These numbers λ_i are the *eigenvalues of A*. The non-trivial solution $\hat{S}_i$ of the homogeneous equations $(A - \lambda_i I)\hat{S}_i = 0$ corresponding to the ith eigenvalue is called the *eigenvalue* of A corresponding to the eigenvalue λ_i. For example, $\hat{S}_2$ is the solution of $(A - \lambda_2 I)S_2 = 0$ corresponding to $\lambda_2 = -k/m$,

$$0\hat{s}_{12} + \frac{k}{m}\hat{s}_{22} + 0\hat{s}_{32} = 0$$

$$\frac{k}{M}\hat{s}_{12} - (\frac{2k}{M} - \frac{k}{m})\hat{s}_{22} + \frac{k}{M}\hat{s}_{32} = 0$$

$$0\hat{s}_{12} + \frac{k}{m}\hat{s}_{22} + 0\hat{s}_{32} = 0.$$

We see $\hat{s}_{22} = 0$ while $\hat{s}_{12} = -\hat{s}_{32}$. Thus, one solution is

$$\hat{S}_2 = (1, 0, -1)$$

Similarly we find one solution for $\hat{S}_1$ is

$$\hat{S}_1 = (1, 1, 1),$$

while one solution for $\hat{S}_3$ is

$$\hat{S}_3 = (1, -\frac{2m}{M}, 1).$$

The computation is over. All that remains is to put the parts together and interpret the solution. If you got lost, presumably this recapitulation will help. We have found a transformation S to new coordinates (ϕ_1, ϕ_2, ϕ_3) such that the differential equations for the ϕ_j's are in diagonal form, $\ddot{\phi}_m = \lambda_j \phi_j$,

$$\ddot{\phi}_1 = 0$$

$$\ddot{\phi}_2 = -\frac{k}{m}\phi_2$$

$$\ddot{\phi}_3 = -k(\frac{2}{M} + \frac{1}{m})\phi_3.$$

The solutions are

$$\phi_1(t) = A_1 + B_1 t,$$
$$\phi_2(t) = A_2 \cos\sqrt{\frac{k}{m}}t + B_2 \sin\sqrt{\frac{k}{m}}t$$
$$\phi_3 = A_3 \cos\sqrt{k(\frac{2}{M}+\frac{1}{m})}t + B_3 \sin\sqrt{k(\frac{2}{M}+\frac{1}{m})}t.$$

Since $\Phi = SU$, and the $\hat{S}_j$ are the columns of S^{-1},

$$S^{-1} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & -\frac{2m}{M} \\ 1 & -1 & 1 \end{pmatrix},$$

we have $U = S^{-1}\Phi$,

$$u_1(t) = \phi_1(t) + \phi_2(t) + \phi_3(t)$$
$$u_2(t) = \phi_1(t) \qquad - \frac{2m}{M}\phi_3(t)$$
$$u_3(t) = \phi_1(t) - \phi_2(t) + \phi_3(t)$$

Although the solutions $\phi_1(t), \phi_2(t)$, and $\phi_3(t)$ can now be substituted into the first set of equations for the u_j's, it is more instructive to leave that step to your imagination and analyze the nature of the solution.

(1) If $\phi_1(t) \neq 0$ but $\phi_2(t) = \phi_3(t) = 0$, then

$$u_1(t) = u_2(t) = u_3(t) = A_1 + B_1 t.$$

Thus all three atoms - the whole molecule - moves with a constant velocity B_1. This is the trivial translation motion of the molecule, simply moving without internal oscillations at all.

(2) If $\phi_2(t) \neq 0$ but $\phi_1(t) = \phi_3(t) = 0$, then

$$u_1(t) = \phi_2(t) = -u_3(t), \quad \text{and} \quad u_2(t) = 0.$$

Thus, the two outside atoms vibrate in opposite directions with frequency $\sqrt{k/m}$ while the center atom remains still:

A FIGURE GOES HERE

(3) If $\phi_3(t) \neq 0$ but $\phi_1(t) = \phi_2(t) = 0$

$$u_1(t) = u_3(t) = \phi_3(t), \quad u_2(t) = -\frac{2m}{M}\phi_3(t).$$

A bit more complicated. The two outside atoms move in the same direction with same frequency $\sqrt{k(\frac{2}{M}+\frac{1}{m})}$, while the center atom moves in a direction opposite to them and with the same frequency but a different amplitude (to conserve linear momentum $m\dot{u}_1 + M\dot{u}_2 + m\dot{u}_3 = 0$). In the figure we take $m = M$.

A FIGURE GOES HERE

These three simple motions are called the *normal modes of oscillation* of the molecule. They are the oscillations determined by the ϕ_1, ϕ_2, and ϕ_3. Every motion of the system is a linear combination of the normal modes of oscillation, the particular oscillation depending on what initial conditions are given. By an appropriate choice of the initial conditions, one or another of the normal modes will result. Otherwise some less recognizable motion will result.

Exercises

Consider the simpler model of a diatomic molecule

A FIGURE GOES HERE

which we will represent as two masses joined by a spring with spring constant k.

(a) Show the equations of motion are

$$m\ddot{u}_1 = k(u_2 - u_1)$$
$$M\ddot{u}_2 = -k(u_2 - u_1)$$

(b) Introduce new variables, $\Phi = SU$,

$$\phi_1 = s_{11} + s_{12}u_2$$
$$\phi_2 = s_{21}u_1 + s_{22}u_2,$$

and find S so that the equation

$$\ddot{\Phi} = SAS^{-1}\Phi$$

is in diagonal form.

(c) Solve the resulting equation and find the normal modes of oscillation. Interpret your results with a diagram.

Chapter 7

Nonlinear Operators: Introduction

7.1 Mappings from $\mathbb{R}^1$ to $\mathbb{R}^1$, a Review

.

The subject of this section is one you presumably know well. Our intention is to briefly review the more important results, stating them in a form which suggests the generalizations we intend to develop.

Consider a function $y = f(x)$, $x \in \mathbb{R}$. This function assigns to each number x another real number y. Thus we may write

$$f : \mathbb{R} \to \mathbb{R}.$$

f is a scalar-valued function of a scalar. What are the simplest such functions? Linear ones of course,

$$f(x) = ax + b.$$

In keeping with our more sophisticated terminology, this should be called an "affine" function (mapping, operator, ...) since it is linear only if $b = 0$. We shall, however, be abusive and refer to such functions as linear mappings. The study of linear functions in one variable, x, is carried out in elementary analytic geometry.

At an early age we enlarged our vocabulary of functions from linear ones to a more general class which includes, for example,

$$f_1(x) = ax^2 + bx + c,\ f_2(x) = \sin x,\ f_3(x) = \sqrt{x}.$$

These functions are all examples of nonlinear functions. They map the reals (only the positive reals in the case of f_3) into the reals. The portion of the reals for which they are defined is called their *domain of definition*, $\mathcal{D}(f)$. Thus

$$\mathcal{D}(f_1) = \mathbb{R}^1,\ \mathcal{D}(f_2) = \mathbb{R}^1,\ \mathcal{D}(f_3) = \{\, x \in \mathbb{R}^1 : x > 0 \,\}.$$

The class of all real valued functions of a real variable is too large to consider. For most purposes it is sufficient to restrict oneself to the class of continuous or sufficiently differentiable functions.

Here is an outline of the basic definitions and theorems from elementary calculus. In our prospective generalization from the simplest case of a function (operator) f which maps numbers to numbers, $f\colon \mathbb{R}^1 \to \mathbb{R}^1$, to the case of a function from vectors to vectors $f\colon \mathbb{R}^n \to \mathbb{R}^m$, all of these concepts and results will need to be extended.
DEFINITION: a_k converges to a, a_k and $a \in \mathbb{R}^1$.
DEFINITION: Continuity.

Theorem 7.1 *The set of continuous functions forms a linear space.*

DEFINITION: The derivative: limit of difference quotient.

Theorem 7.2 *1.* $\frac{d}{dx}(af+bg) = a\frac{df}{dx}f + b\frac{dg}{dx}$ *(linearity)*
2. $\frac{d}{dx}(fg) = f\frac{dg}{dx} + (\frac{df}{dx})g$ *(Product rule)*
3. $\frac{d}{dx}(f\circ g) = \frac{df}{dg}\frac{dg}{dx}$ *(Chain rule)*

Theorem 7.3 *The Mean Value Theorem.*

DEFINITION: The integral.

Theorem 7.4 *1.* $\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx$
2. $\int_1^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx$
3. $\int_a^b [\alpha f(x) + \beta g(x)]\,dx = \alpha\int_a^b f(x)\,dx + \beta\int_a^b g(x)\,dx$ *(linearity)*
4. $\int_a^b (f\circ\phi)(x)\frac{d\phi}{dx}\,dx = \int_{\phi(a)}^{\phi(b)} f(x)\,dx$ *(Change of variable in an integral)*

Theorem 7.5 *1.* $\int_a^b \frac{df}{dx}(x)\,dx = f(b) - f(a)$
2. $\frac{d}{dx}\int_a^x f(t)\,dt = f(x)$
3. $\int_a^b f(x)\frac{dg}{dx}\,dx = fg\,\big|_a^b - \int_a^b \frac{df}{dx}g(x)\,dx$ *(Integration by parts).*

REMARK: These theorems contain essentially all of elementary calculus. What are missing are specific formulas for the derivatives and integrals of the basic functions as well as the application of these theorems to compute maxima, area, etc.

Exercises

(1) Use the definition of the derivative (as the limit of a difference quotient) to compute the derivatives of the following functions at the given point.
a). $3x^2 - x + 1$, $x_0 = 2$
b). $\frac{1}{x+1}$, $x_0 = 2$
c). $\frac{x}{1+x}$, $x_0 = 2$
d). $\frac{x}{1-x}$, $x = x_0 \neq 1$.

(2) Use the definition of the integral to evaluate
$$\int_0^2 x^2\,dx.$$
You should approximate the area by rectangular strips and evaluate the limit as the width of the thickest strip tends to zero. [Hint: $1^2+2^2+3^2+\cdots+n^2 = \frac{n(n+1)(2n+1)}{6}$].

(3) Prove that
$$.6 < \log 2 < .8 \qquad (\log 2 = 0.693)$$
by using the definition of the integral to find upper and lower bounds for
$$\log 2 = \int_1^2 \frac{1}{x}\,dx.$$

(4) Find the equation of the straight line which is tangent to the curve $f(x) = x^{7/3} + 1$ at $x = 1$. Draw a sketch indicating both the curve and tangent line. Use the tangent line to approximately evaluate $(1.01)^{7/3}$. Find some estimate for the error in your approximation.

7.2 Generalities on Mappings from $\mathbb{R}^n$ to $\mathbb{R}^m$.

A function, or operator, F which maps $\mathbb{R}^n$ to $\mathbb{R}^m$, is a rule which assigns to each vector X in $\mathbb{R}^n$ another vector $Y = F(X)$ in $\mathbb{R}^m$. It is a function from vectors to vectors, a vector-valued function of a vector. We have already discussed the case when F is an affine operator,
$$Y = F(X) = b + LX$$
or in coordinates,
$$\begin{aligned} y_1 &= b_1 + a_{11}x_2 + \cdots + a_{1n}x_{1n} \\ y_2 &= b_2 + a_{21}x_2 + \cdots + a_{2n}x_n \\ &\cdot \\ &\cdot \\ &\cdot \\ y_m &= b_m + a_{m1}x_2 + \cdots + a_{mn}x_n \end{aligned}$$

Linear algebra can be thought of as the study of higher dimensional analytic geometry, the affine transformations taking the role of the straight line $y = b + cx$.

But now it is time to consider more complicated mappings from $\mathbb{R}^n$ to $\mathbb{R}^m$. Here is an

EXAMPLE: $\begin{cases} y_1 = x_1 + x_2 \sin \pi x_3 \\ y_2 = e^{1-x_1} - \sqrt{x_2}. \end{cases}$

This transformation maps vectors $X = (x_1, x_2, x_3) \in \mathbb{R}^3$ to vectors $Y = (y_1, y_2) \in \mathbb{R}^2$. Note the second function is only defined for $x_2 \geq 0$. Thus the domain of the transformation F is

$$\mathcal{D}(F) = \left\{ X \in \mathbb{R}^3 \colon x_2 \geq 0 \right\}.$$

For example, F maps the point $(1, 4, \frac{1}{6})$ into the point $(3, -1)$.

It is usual to write a transformation F which maps a set $A \subset \mathbb{R}^n$ to a set $B \subset \mathbb{R}^m$ in terms of its *components*,

$$\begin{aligned} y_1 &= f_1(x_1, \ldots, x_n) = f_1(X) \\ y_2 &= f_2(x_1, \ldots, x_n) = f_2(X) \\ &\cdot \\ &\cdot \\ &\cdot \\ y_m &= f_m(x_1, \ldots, x_n) = f_m(X), \end{aligned}$$

or more concisely as

$$Y = F(X).$$

To discuss continuity etc. for nonlinear mappings from $\mathbb{R}^n$ to $\mathbb{R}^m$, it is necessary that the distance between points be defined. We shall use the Euclidean norm - although any other norm could also be used. If $X = (x_1, \ldots, x_k)$ is a point (or vector, if you like) in $\mathbb{R}^k$, then $\|X\| = \sqrt{x_1^2 + \cdots + x_k^2}$. To review briefly, a *sequence* of points X_j in $\mathbb{R}^k$ *converges* to a point X in $\mathbb{R}^k$ if, given any $\epsilon > 0$, there is an integer N such that

$$\|X_j - X\| < \epsilon \qquad textforall \quad j \geq N.$$

An *open ball* in $\mathbb{R}^k$ of radius r about the point X_0 is the set $B(X_0; r) = \{ X \in \mathbb{R}^k \colon \|X - X_0\| < r \}$.
A *closed ball* in $\mathbb{R}^k$ is

$$\bar{B}(X_0; r) = \{ X \in \mathbb{R}^k \colon \|X - X_0\| \leq r \}.$$

The only difference is the open ball does not contain the boundary of the ball. In two dimensions, $\mathbb{R}^2$, the names open and closed *disc* are often used.

A set $D \subset \mathbb{R}^k$ is *open* if each point $X \in D$ is the center of some ball contained entirely within D. The radius may be very tiny. Every open ball is open, as can be seen in the figure. A closed ball is not open since there is no way of placing a small ball about a point on the boundary in such a way that the small ball is inside the larger one. A set A is

closed if it contains *all* of its *limit points*, that is, if the points $X_j \in A$ converge to a point X, $X_j \to X$, then X is also in A. An open ball is not closed, for a sequence of points in the ball may converge to a point on the boundary, and the boundary points are not in the ball. For the special case of $\mathbb{R}^1$, these notions coincide with those of open and closed intervals. Again, sets - like doors - may be neither open nor closed.

A point set D is *bounded* if it is contained in some ball (of possibly large radius). The point X is *exterior* to D if X does not belong to D and if there is some ball about X none of whose point are in D. X is *interior* to D if X belongs to D and there is some ball about X all of whose points are in D. X is a *boundary point* of D if it is neither interior nor exterior to D. Note that a boundary point of D may or may not belong to D. For example, the boundaries of the open and closed balls $B(0; r)$, $\bar{B}(0; r)$ are the same. The boundary of a set D is denoted by ∂D. It is evident that a set is open if and only if every point is an interior point, and a set is closed if and only if it contains all of its boundary points.

DEFINITION: Let A be a set in $\mathbb{R}^n$ and C a set in $\mathbb{R}^m$. The function $F \colon A \to C$ is *continuous at the interior point* $X_0 \in A$ if, given any radius $\epsilon > 0$, there is a radius $\delta > 0$ such that

$$\|F(X) - F(X_0)\| < \epsilon \quad textforall \quad \|X - X_0\| < \delta.$$

[Observe the norm on the left is in $\mathbb{R}^m$ while that on the right is in $\mathbb{R}^n$].

It is easy to prove

Theorem 7.6 . *An affine mapping $F(X) = b + LX$ from $\mathbb{R}^n$ to $\mathbb{R}^m$ is continuous at every point $X_0 \in \mathbb{R}^m$.*

PROOF: First, $F(X) - F(X_0) = b + LX - b - LX_0 = L(X - X_0)$. Thus,

$$\|F(X) - F(X_0)\| = \|L(X - X_0)\|.$$

Let $((a_{ij}))$ be a matrix representing L with respect to some bases for $\mathbb{R}^n$ and $\mathbb{R}^m$. Then by Theorem 17, p. 373

$$\|L(X - X_0)\|^2 = \langle L(X - X_0), L(X - X_0)\rangle \leq k\|L(X - X_0)\| \, \|X - X_0\|,$$

where

$$k^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2.$$

Therefore

$$\|L(X - X_0)\| \leq k\|X - X_0\|.$$

It is now clear that if $X \to X_0$, then $L(X - X_0) \to 0$. More formally, given any $\epsilon > 0$, if $\delta = \frac{\epsilon}{k+1}$, we have

$$\|F(X) - F(X_0)\| < \epsilon \quad textforall \quad \|X - X_0\| < \delta.$$

The following theorems have the same proofs as were given earlier for special cases. (See a first year calculus book and our Chapter 0).

Theorem 7.7 . *Let F_1 and F_2 map $A \subset \mathbb{R}^n$ into $C \subset \mathbb{R}^m$. If F_1 and F_2 are continuous at the interior point $X_0 \in A$, then*

1. *$aF_1 + bF_2$ is continuous at X_0.*
2. *$\langle F_1, F_2 \rangle$ is continuous at X_0.*

Theorem 7.8 . *Let $F = (f_1, \dots, f_m)$ map $A \subset \mathbb{R}^n$ into $C \subset \mathbb{R}^m$. Then F is continuous at the interior point $X_0 \in A$ if and only if each of the f_j, $j = 1, \dots, m$, is continuous at X_0.*

Theorem 7.9 . *Let $F\colon A \to C$, where A is a closed and bounded (= compact) set. If F is continuous at every point of A, then it is bounded; that is, there is a constant M such that $\|F(x)\| \le M$ for all $X \in A$. Moreover, if M_0 is the least upper bound, then there is a point $X_0 \in A$ such that $\|F(X_0)\| = M_0$. Similarly, if m_0 is the greatest lower bound for $\|F\|$, then there is a point $X_1 \in A$ such that $\|F(X_1)\| = m_0$.*

There is nothing better than to close this otherwise unauspicious section with one of the crown jewels of mathematics - the Fundamental Theorem of Algebra, all of whose proofs require the non-algebraic notion of continuity. Let

$$p(z) = a_0 + a_1 z + \cdots + a_n z^n, \qquad (n \ge 1),$$

where the a_j's are complex numbers and a_n is not zero.

For every complex number z, the value of the function $p(z)$ is a complex number. Thus $p\colon \mathbb{C} \to \mathbb{C}$. We want to prove there is at least one $z_0 \in \mathbb{C}$ such that $p(z_0) = 0$.

Lemma 7.10 . *$p(z)$ is a continuous function for every $z \in \mathbb{C}$.*

PROOF: Identical to the proof that a real polynomial is continuous everywhere.

Lemma 7.11 *Let D be a set in the complex plane in which $p(z) \ne 0$. The minimum modulus of $p(z)$, that is, the minimum value of $|p(z)|$, cannot occur at an interior point of D. It must occur on the boundary ∂D of D.*

PROOF: Let z_0 be any interior point of D. Rewrite $p(z)$ in the form

$$p(z) = b_0 + b_1(z - z_0) + \cdots + b_n(z - z_0)^n.$$

Since $p(z_0) \ne 0$, we know $b_0 \ne 0$. Also, because p is not identically constant, at least one coefficient following b_0 is not zero. Take b_k to be the first such coefficient. We must write b_0, b_k and $z - z_0$ in polar form,

$$b_0 = \rho_0 e^{u\alpha} \qquad b_k = \rho_1 e^{i\beta} \qquad z - z_0 = \rho e^{i\theta},$$

where $\rho_0 = |p(z_0)|$, ρ_1 and ρ are positive real numbers. Here we are restricting z to a point on a circle of radius ρ about z_0, after taking ρ small enough to insure this circle is interior to D. Then

$$p(z) = \rho_0 e^{i\alpha} + \rho_1 e^{i\beta} \rho^k e^{ik\theta} + b_{k+1}(z - z_0)^{k+1} + \cdots + b_n(z - z_0)^n$$

$$= \rho_0 e^{i\alpha} + \rho_1 \rho^k e^{i(\beta + k\theta)} + (z - z_0)^{k+1}[b_{k+1} + \cdots + b_n(z - z_0)^{n-k-1}].$$

Pick the particular point $\hat{z}$ on the circle whose argument θ is given by $\beta + k\theta = \alpha + \pi$. Then $e^{i(\beta+k\theta)} = e^{i(\alpha+\pi)} = -e^{i\alpha}$, so

$$p(\hat{z}) = (\rho_0 - \rho_1 \rho^k) e^{i\alpha} + (\hat{z} - z_0)^{k+1}[b_{k+1} + \cdots + b_n(\hat{z} - z_0)^{n-k-1}].$$

By the triangle inequality we find

$$|p(\hat{z})| \le \left|\rho_0 - \rho_1 \rho^k\right| + \rho^{k+1}[|b_{k+1}| + \cdots + |b_n| \rho^{n-k-1}].$$

Choose the radius ρ so small that $\rho_0 - \rho_1 \rho^k \ge 0$. Then

$$|p(\hat{z})| \le \rho_0 - \rho_1 \rho^k + \rho^{k+1}[|b_{k+1}| + \cdots + |b_n| \rho^{n-k-1}].$$

By choosing ρ smaller yet, if necessary, we can make the term $\rho[|b_{k+1}| + \cdots + |b_n| \rho^{n-k-1}] < \frac{1}{2}\rho_1$. Consequently,

$$\begin{aligned} |p(\hat{z})| &\le \rho_0 - \rho_1 \rho^k + \frac{1}{2}\rho_1 \rho^k = \rho_0 - \frac{1}{2}\rho_1 \rho^k \\ &< \rho_0 = |p(z_0)|. \end{aligned}$$

Thus, if z_0 is any *interior* point of a domain D in which p does not vanish, then there is a point $\hat{z}$ also interior to D such that $|p(z)| < |p(z_0)|$. Therefore, the minimum of $|p(z)|$ must occur on the boundary of any set in which p does not vanish.

Lemma 7.12 . *Given any real number M, there is a circle $|z| = R$ on which $|p(z)| > M$ for all z, $|z| = R$.*

PROOF: For $z \ne 0$, we can write the polynomial $p(z)$ as

$$\frac{p(z)}{z^n} = a_n + \frac{a_{n-1}}{z} + \cdots + \frac{a_0}{z^n}.$$

From the triangle inequality written in the form $|f_1 + f_2| \ge |f_1| - |f_2|$, we find

$$\left|\frac{p(z)}{z^n}\right| \ge |a_n| - \left|\frac{a_{n-1}}{z} + \cdots + \frac{a_0}{z^n}\right|.$$

If $|z|$ is taken large enough, $|z| \ge R_0$, it is possible to make the second term on the right less than $|a_n|/2$,

$$\left|\frac{a_{n-1}}{z} + \cdots + \frac{a_0}{z^n}\right| < \left|\frac{a_n}{2}\right|, \quad texton \quad |z| = R > R_0$$

Therefore, for $|z| = R \ge R_0$

$$\left|\frac{p(z)}{z^n}\right| \ge |a_n| - \left|\frac{a_n}{2}\right| = \frac{1}{2}|a_n|,$$

so

$$|p(z)| \ge \frac{1}{2}|a_n| R^n, \quad texton \quad |z| = R.$$

It is now clear that by choosing R sufficiently large, $|p(z)|$ can be made to exceed any constant M on the circle $|z| = R$.

Theorem 7.13 *(Fundamental Theorem of Algebra). Let*

$$p(z) = a_0 + a_1 z + \cdots + a_n z^n, \qquad a_n \neq 0, n \geq 1,$$

be any polynomial with possibly complex coefficients, $a_0, a_1, \ldots, a_n$. Then there is at least one number $z_0 \in \mathbb{C}$ such that $p(z_0) = 0$. In other words, every polynomial has at least one complex root.

PROOF: By Lemma 3, we can find a large circle $|z| = R$, on which $|p(z)| > 2|a_0|$ for all $|z| = R$. Since $p(z)$ is a continuous function, by Theorem 4 there is a point z_0 in the closed and bounded disc $|z| \leq R$ for which $|p|$ attains its minimum value m_0, $|p(z_0)| = m_0$. If $p(z_0) = 0$, we are done. However if p does not vanish inside the closed disc, by the important Lemma 2 its minimum value is attained only on the boundary, so z_0 is on the circle $|z_0| = R$. But on the circle we know $|p(z_0)| > 2|a_0| = 2|p(0)|$, so the minimum is not at z_0 after all. The assumption that p does not vanish in the disc $|z| \leq R$ had led us to a contradiction. Notice the proof does not give a procedure for finding the root whose existence has been proved.

Exercises

1. Prove Theorem 2, part 1.
2. Use the Fundamental Theorem of Algebra along with the "factor theorem" of high school algebra to prove that a polynomial of degree n has exactly n roots (some of which may be repeated roots).

7.3 Mapping from $\mathbb{R}^1$ to $\mathbb{R}^n$

.

As a particle moves along a curve γ in $\mathbb{R}^n$ its position $F(t)$ at time t can be specified by a vector

$$X = F(t) = (f_1(t), f_2(t), \ldots, f_n(t)),$$

where $x_j = f_j(t)$ is the jth *coordinate* of the position at time t. Thus, the curve is specified by $F(t)$, a mapping from numbers to vectors, $F\colon A \subset \mathbb{R}^1 \to \mathbb{R}^n$, where A is the domain of definition of F.

For example, the mapping

$$F\colon t \to (\cos \pi t, \sin \pi t, t), \quad t \in (-\infty, \infty)$$

which may also be written as

$$F(t) = (\cos \pi t, \sin \pi t, t)$$

can be thought of as describing the motion of a particle along a helix.

It is natural to ask about the velocity, which means derivative must be defined.

DEFINITION: Let $F(t)$ define a curve γ for t in the interval $A = [a, b]$. Consider the difference quotient

$$\frac{F(t+h) - F(t)}{h}, \qquad t \quad textand \quad t+h \quad \text{in} \quad A,$$

where t is fixed. If this vector has a limit as h tends to zero, then F is said to have a *derivative* $F'(t)$ at t,

$$F'(t) = \lim_{h \to 0} \frac{F(t+h) - F(t)}{h},$$

while the curve has *slope* $F'(t)$ at t. Some other common notations are

$$\dot{F}(t), \quad \frac{dF}{dt}, \quad D_t F.$$

The curve γ is called *smooth* if i) the derivative $F'(t)$ exists and is continuous for each t in $[a, b]$, and if ii) $\|F'(t)\| \neq 0$ for any point t in $[a, b]$.

If t represents time, then $F'(t)$ is the *velocity* of the particle at time t while $\|F'(t)\|$ is the *speed*.

If $F(t)$ is given in terms of coordinate functions, $F\colon t \to (f_1(t), \ldots, f_n(t))$, how can the derivative of F be computed?

Theorem 7.14 . *If $F(t) = (f_1(t), \ldots, f_n(t))$ is a differentiable mapping of $A \subset \mathbb{R}^1$ into $\mathbb{R}^n$, then the coordinate functions are differentiable and*

$$\frac{dF}{dt} = \left(\frac{df_1}{dt}, \frac{df_2}{dt}, \ldots, \frac{df_n}{dt} \right).$$

Conversely, if the coordinate functions are differentiable, then so is $F(t)$ and the derivative is given by the above formula.

PROOF: If t and $t+h$ are both in A, then

$$\frac{F(t+h) - F(t)}{h} = \frac{1}{h}[(f_1(t+h), \ldots, f_n(t+h)) - (f_1(t), \ldots, f_n(t))]$$

$$= \left(\frac{f_1(t+h) - f_1(t)}{h}, \ldots, \frac{f_n(t+h) - f_n(t)}{h} \right)$$

Since the limit as $h \to 0$ of the expression on the left exists if and only if all of the limits

$$\lim_{h \to 0} \frac{f_j(t+h) - f_j(t)}{h}, \quad j = 1, \ldots, n$$

exist, the theorem is proved.

EXAMPLES:

(1) If $F\colon t \to (\cos \pi t, \sin \pi t, t)$, $t \in (-\infty, \infty)$, F is differentiable for all t since each of the coordinate functions are differentiable. Also,

$$F'(t) = (-\pi \sin \pi t, \pi \cos \pi t, 1).$$

In addition, the curve - a helix - which F defines is smooth since F' is continuous and

$$\|F'(t)\| - \sqrt{\pi^2 \sin^2 \pi t + \pi^2 \cos^2 \pi t + 1} = \sqrt{\pi^2 + 1} \neq 0,$$

(2) Let $F\colon t \to (a_1 + b_1 t, a_2 + b_2 t, a_3 + b_3 t) = P + Qt$ where $P = (a_1, a_2, a_3)$ and $Q = (b_1, b_2, b_3)$ are constant vectors. Then the curve F defines is a straight line which passes through the point $P = (a_1, a_2, a_3)$ at $t = 0$. F is differentiable for all t, since each of the coordinate functions are differentiable. Furthermore,

$$F'(t) = Q = (b_1, b_2, b_3),$$

a *constant vector* pointing in the direction $Q = (b_1, b_2, b_3)$, as is anticipated for a straight line. Because

$$\|F'(t)\| = \|Q\| = \sqrt{b_1^2 + b_2^2 + b_3^2},$$

this curve is smooth except in the degenerate case $b_1 = b_2 = b_3 = 0$, that is, $Q = 0$, when the curve degenerates to a single point, $F(t) = (a_1, a_2, a_3) = P$.

(3) The curve defined by the mapping $F\colon t \to (t, |t|)$ is differentiable everywhere and $\|F'(t)\| \neq 0$ except at $t = 0$. It is not differentiable there since the second coordinate function, $f_2(t) = |t|$ is not differentiable at $t = 0$. Thus, the curve is smooth except at $t = 0$.

(4) The curve defined by the mapping $F\colon t \to (t^3, t^2)$ is differentiable everywhere, and

$$F'(t) = (3t^2, 2t).$$

However, $\|F'(t)\| = \sqrt{9t^4 + 4t^2}$, so the curve is smooth everywhere except at $t = 0$, which corresponds to a cusp at the origin in the x_1, x_2 plane.

It is elementary to compute the derivative of the sum of two vectors. The derivative of a product can be defined for the inner product, and for the product with scalar-valued function.

Theorem 7.15 . *If $F(t)$ and $G(t)$ both map an interval $A \subset \mathbb{R}^1$ into $\mathbb{R}^n$, and are both differentiable there, then for all $t \in A$,*

1. $\frac{d}{dt}[aF + bG] = a\frac{dF}{dt} + b\frac{dG}{dt}$ *(linearity of the derivative).*
2. $\frac{d}{dt}\langle F, G\rangle = \langle F', G\rangle + \langle F, G'\rangle$, *(in "dot product" notation: $\frac{d}{dt}(F \cdot G) = F' \cdot G + F \cdot G'$).*

Proof: Since these are identical to the proofs of the corresponding statements for scalar-valued functions, we prove only the second statement.

$$\frac{d}{dt}\langle F(t), G(t)\rangle = \lim_{h \to 0} \frac{1}{h}[\langle F(t+h), G(t+h)\rangle - \langle F(t), G(t)\rangle]$$

$$= \lim_{h \to 0} \frac{1}{h}[\langle F(t+h) - F(t), G(t+h)\rangle + \langle F(t), G(t+h) - G(t)\rangle]$$

$$= \lim_{h \to 0} \left[\frac{\langle F(t+h) - F(t), h\rangle}{G(t+h)} + \langle F(t), \frac{G(t+h) - G(t)}{h}\rangle\right]$$

$$= \langle F'(t), G(t)\rangle + \langle F(t), G'(t)\rangle.$$

An interesting and simple consequence is the fact that if a particle moves on a curve $F(t)$ which remains a fixed distance from the origin, $\|F(t)\| \equiv$ constant $= c$, then the velocity vector F' is always orthogonal to the position vector F. This follows from

$$c^2 = \|F(t)\|^2 = \langle F(t), F(t)\rangle,$$

so taking the derivative of both sides we find

$$0 = \langle F', F\rangle + \langle F, F'\rangle = 2\langle F, F'\rangle.$$

Thus $\langle F, F'\rangle = 0$ for all t, an algebraic statement of the orthogonality. As a particular example, the mapping

$$F(t) = (\cos \frac{\pi}{1+t^2}, \sin \frac{\pi}{1+t^2})$$

has the property $\|F(t)\| = 1$ for all t. You can see the path of the particle in the figure. At $t = 0$ the particle is at $(-1, 0)$. As time increases, the particle moves along an arc of the unit circle toward $(1, 0)$, reaching $(0, 1)$ at $t = 1$. The velocity at time t is

$$F'(t) = \frac{2\pi t}{(1+t^2)^2}(\sin \frac{\pi}{1+t^2}, -\cos -\frac{\pi}{1+t^2}).$$

From this expression, it is evident the particle slows down as it approaches $(1, 0)$. In fact, the particle never does manage to reach $(1, 0)$.

We would like to define the notion of a straight line which is *tangent* to a smooth curve at a given point. There is one touchy issue. You see, the curve may intersect itself, thus having two or more tangents at the same point. Once acknowledged, the difficulty is resolved by realizing that for each value of t, there is a unique point $F(t)$ on the curve. X_0 is a double point if $F(t_1) = F(t_2) = X_0$.

By picking one value of t, there will be a unique tangent line to the curve for this value of t. Thus, we define the tangent line for $t = t_1$ to the curve defined by a differentiable function $F(t)$ as the straight line whose equation is

$$A(t) = F(t_1) + F'(t_1)(t - t_1).$$

At $t=t_1$, the curves defined by $F(t)$ and $A(t)$ have the same value $F(t_1)=X_0$ and the same derivative (slope), $F'(t)$.

EXAMPLE: Consider the curve defined by the mapping $F\colon t\to(3+t^3-t, t^2-t)$, $t\in(-\infty,\infty)$. The point $(3,0)$ is a double point since $F\colon 0\to(3,0)$ and $F\colon 1\to(3,0)$. Thus, the line tangent to the point $(3,0)$ when $t=1$ is defined by

$$A(t)=(3,0)+(2,1)(t-1)=(3,0)+(2(t-1),\ (t-1))$$

or

$$A(t)=(1,-1)+(2t,t).$$

Since we are still working with functions $F(t)$ of one real variable t, the mean value theorem and chain rule follow immediately by applying the corresponding theorems for scalar valued functions to each of the components $f_1(t),\ldots,f_n(t)$ of $F(t)$.

Theorem 7.16 *(Approximation Theorem and Mean Value Theorem). If the vector valued function $F(t)$ is continuous for $t\in[a,b]$ and differentiable for $t\in(a,b)$ then for $t_0\in(a,b)$,*

1. $F(t)=F(t_0)+\frac{dF}{dt}\big|_{t_0}(t-t_0)+R(t,t_0)\,|t-t_0|$ *where*

$$\lim_{t\to t_0}\|R(t,t_0)\|=0.$$

2. *There is a point τ between t and t_0 such that*

$$\|F(t)-F(t_0)\|\le\|F'(\tau)\|\,|t-t_0|\,.$$

3. *If $F=(f_1,\ldots,f_n)$, there are points $\tau_1,\ldots,\tau_n$ between t and t_0 such that*

$$F(t)=F(t_0)+L(t-t_0),$$

where L is the linear transformation

$$L=(f_1'(\tau_1),f_2'(\tau_2),\ldots,f_n'(\tau_n))$$

REMARK: Although 1 and 3 follow from the one variable case $f(t)$—and will be proved again in greater generality later on - the proof of 2 is difficult under our weak hypothesis. If the stronger assumption, F is continuously differentiable, is made, then 2 becomes easy, and the factor $\|F'(\tau)\|$ can be replaced by a constant $M=\max_{\tau\in[a,b]}\|F'(\tau)\|$, since a continuous function $\|F'(\tau)\|$ does assume its maximum if τ is in a closed and bounded set, $\tau\in[a,b]$.

Corollary 7.17 . *If F satisfies the hypotheses of Theorem 8 and if $F'(t)\equiv0$ for all $t\in[a,b]$, then F is a constant vector.*

PROOF: Just look at 2 or 3 above to see that for any points t,t_0 in $[a,b]$, we have $F(t)=F(t_0)$.

Theorem 7.18 *(Chain Rule). Consider the vector-valued function $F(t)$ which is differentiable for $t\in(a,b)$, and the scalar valued function $\phi(s)$ which is differentiable for $s\in(\alpha,\beta)$. If the range of ϕ is contained in (a,b), $\mathcal{R}(\phi)\subset(a,b)$, then the composed function $G(s)=(F\circ\phi)(s)=F(\phi(s))$ is differentiable as a function of s for all s in (α,β) and*

$$G'(s)=F'(\phi(s))\phi'(s),$$

that is,

$$\frac{dG}{ds}(s)=\frac{dF}{d\phi}(\phi)\frac{d\phi}{ds}(s)=\frac{dF}{dt}(t)\Big|_{t=\phi(s)}\frac{d\phi}{ds}(s).$$

If $F(t)=(f_1(t),\ldots,f_n(t))$, then

$$G(s)=F((s))=(f_1(\phi(s)),\ldots,f_n(\phi(s)),\quad \textit{and}$$

$$\begin{aligned}G'(s)&=)f_1'(\phi)\phi'(s),\ldots,f_n'(\phi)\phi'(s))\\&=(f_1'(\phi),\ldots,f_n'(\phi))\phi'(s).\end{aligned}$$

Proof not given here. It is the same as that given in elementary calculus for $n=1$. A more general theorem containing this one is proved later (p. 701).

EXAMPLES:

1. If $F(t)=(1-t^2,t^3-\sin\pi t)$ and $\phi(s)=e^{-s}$, then $G(s)=(F\circ\phi)(s)=(1-e^{-2s},e^{-3w}-\sin\pi e^{-2})$. We compute $G'(s)$ in two distinct ways, using the chain rule, and directly from the formula for $G(s)$. By the chain rule:

$$\begin{aligned}G'(s)&=F'(t)\big|_{t=\phi(s)}\phi'(s)\\&=(-2t,3t^2-\pi\cos\pi t)\big|_{t=e^{-s}}(-)e^{-s}\\&=-(-2e^{-s},3e^{-2s}-\pi\cos\pi e^{-s})e^{-s},\end{aligned}$$

In particular, at $s=0$, since $t=1$ when $s=0$, we find

$$G'(0)=-(-2,3+\pi)=(2,-3,-\pi)$$

Directly from the formula for $G(s)=(1-e^{-2s},-3e^{-3s}-\sin\pi e^{-s})$, we find

$$G'(s)=(2e^{-2s},-3e^{-3s}+\pi e^{-s}\cos\pi e^{-2}),$$

which agrees with the chain rule computation.

Since the derivative $F'(t)$ of a function $F(t)$ from numbers to vectors, $F\colon\mathbb{R}^1\to\mathbb{R}^n$, is also a function of the same type, the second and higher order derivatives can be defined inductively;

$$\frac{d^2}{dt^2}F(t):=\frac{d}{dt}F'(t),\ \frac{d^{k+1}}{dt^{k+1}}F(t):=\frac{d}{dt}F^{(k)}(t).$$

EXAMPLE: If $F\colon t \to (\cos \pi t, \sin \pi t, t)$, then

$$\begin{aligned} F''(t) &= \frac{d}{dt}(-\pi \sin \pi t, \pi \cos \pi t, 1) \\ &= (-\pi^2 \cos \pi t, -\pi^2 \sin \pi t, 0). \end{aligned}$$

If $F(t)$ represents the position of a particle at time t, then $F''(t)$ is the *acceleration* of the particle at time t. All of these ideas were used in the last two sections in Chapter 6 where *linear* systems of ordinary differential equations were encountered. Time permitting, a second application to a *non-linear* system of O.D.E.'s will be treated in Section of Chapter. There another of the crown jewels in the intellectual history of mankind will be discussed: Newton's incredible solution of "the two body problem", that is, to determine the motion of the heavenly bodies.

Recall that the *length* of a curve is defined to be the limit of the lengths of inscribed polygons which approximate the curve as the length of the longest subinterval tends to zero - if the limit does exist. Let the curve γ, which we assume is smooth, be determined by the function $F(t), t \in [a,b]$. Then the length of the straight line joining $F(t_j)$ to $F(t_j + \Delta t_j), t_{j+1} = t_j + \Delta t_j$, is

$$\|F(t_j + \Delta t_j) - F(t_j)\| = \|\frac{F(t_j + \Delta t_j) - F(t_j)}{\Delta t_j}\|\Delta t_j$$

Adding up the lengths of these segments and letting the largest Δt_j tend to zero, we find the length of γ is given by

$$\mathcal{L}(\gamma) = \int_1^b \|F'(t)\|\, dt.$$

If the function F is defined through coordinates, $F(t) = (f_1(t), \ldots, f_n(t))$, this formula reads

$$\mathcal{L}(\gamma) = \int_a^b \sqrt{f'^2_1 + f'^2_2 + \cdots + f'^2_n}\, dt.$$

You will recognize the special case where $F(t) = (x(t), y(t))$

$$\mathcal{L}(\gamma) = \int_a^b \sqrt{\dot{x}^2 + \dot{y}^2}\, dt.$$

EXAMPLE: Find the length of the portion of the helix γ defined by $F(t) = (\cos t, \sin t, t)$, for $t \in [0, 2\pi]$. This is one "hoop" of the helix. Since $F'(t) = (-\sin t, \cos t, 1)$, we have $\|F'(t)\| = \sqrt{\sin^2 t + \cos^2 t + 1} = \sqrt{2}$, so the length is

$$\mathcal{L}(\gamma) = \int_0^{2\pi} \sqrt{2}\, dt = 2\pi\sqrt{2}.$$

For each $t \in [a,b]$, we can define an *arc length function* $s(t)$, the arc length from a to t, by

$$s(t) = \int_a^t \|F'(\tau)\| d\tau.$$

Note we are using a dummy variable of integration τ. By the fundamental theorem of calculus, we have

$$\frac{ds}{dt} = \|F'(t)\|$$

Since ds/dt can be thought of as the rate of change of arc length with respect to time, it is the *speed* of a particle moving *along the curve*, the *tangential speed.*

The integral used in arc length is the integral of a scalar- valued function $\|F'(t)\|$. How can we define the integral of a vector-valued function $F(t) = (f_1(t), \ldots, f_n(t))$? Just integrate each component, assuming they are all integrable of course,

$$\int_a^b F(t)\, dt := (\int_a^b f_1(t)\, dt, \ldots, \int_a^b f_n(t)\, dt).$$

For example, if $F(t) = (t - 3t^2, 1 - \sqrt{2t}, e^{3t})$, then

$$\begin{aligned} \int_0^2 F(t)\, dt &= (\int_0^2 (t - 3t^2)\, dt, \int_0^2 (1 - \sqrt{2t})\, dt, \int_0^2 e^{3t}\, dt) \\ &= (-4, -\frac{2}{3}, e^6 - 1). \end{aligned}$$

We give no physical interpretation of the integral (as an area or the like) except in the case where $F(t)$ represents the velocity of a particle. Then $\int_a^b F(t)\, dt$ is the vector pointing from the position at $t = a$ to the position at $t = b$.

Exercises

(1) (a) Describe and sketch the images of the curves $F\colon \mathbb{R}^1 \to \mathbb{R}^2$ defined by
   (i) $F(t) = (2t, 3 - t)$
   (ii) $F(t) = (2t, |3 - t|)$
   (iii) $F(t) = (t^2, 1 + t^2)$
   (iv) $F(t) = (2t, \sin t)$
   (v) $F(t) = (t^2, 1 + t^4)$
   (b) Which of the above mappings are differentiable and for what value(s) of t? Find the derivatives if the functions are differentiable. Which of the curves defined by these mappings are smooth, and where are they not smooth?

(2) Use the *definition* of the derivative to find $F'(t)$ at $t = 2\pi$ for the functions
   a). $F(t) = (2t, 3 - t) \qquad t \in (-\infty, \infty)$.
   b). $F(t) = (1 + t^2, \sin 2t). \quad t \in (-\infty, \infty)$.

(3) Find the lengths of the curves γ defined by the mappings

a). $F(t) = (a_1 + b_1 t, a_2 + b_2 t, \ldots, a_n + b_n t), = P + Qt, \quad t \in [0,1]$.

b). $F(t) = (\sin 2t, 1 - 3t, \cos 2t, 2t^{3/2}), t \in [-\pi, 2\pi]$

(4) Consider the curve defined by the equation

$$F(t) = (t - t^2, t^4 - t^2 + 1), \quad t \in (-\infty, \infty)$$

a). Sketch the curve.

b). Where does the curve intersect itself?

c). Find the line tangent to the curve at the image of $t = 1$.

(5) If $F\colon A \subset \mathbb{R}^1 \to \mathbb{R}^n$ is twice continuously differentiable and $F''(t) \equiv 0$ for all $t \in A$, what can you conclude? Please prove your assertion. [Hint: First consider the special case where $F\colon \mathbb{R}^1 \to \mathbb{R}^1$].

(6) Let $F(t)$ be a twice differentiable function which maps a set in $\mathbb{R}^1$ into $\mathbb{R}^n$ and satisfies the ordinary differential equation $F'' + \mu F' + kF = 0$, where k and μ are positive constants. Define the energy as

$$E(t) = \frac{1}{2}\|F'\|^2 + \frac{1}{2}k\|F\|^2$$

(a) Prove $E(t)$ is a non-increasing function of t (energy is dissipated). [Hint: $dE/dt = ?$].

(b) If $F(0) = 0$ and $F'(0) = 0$, prove $E(t) \equiv 0$.

(c) Prove there is at most one function which satisfies the given differential equation as well as the initial conditions $F(0) = A$, $F'(0) = B$, where A and B are given vectors.

(7) If $F(t) = (1 - e^{2t}, t^3, \frac{1}{1+t^2})$, and $\phi(x) = \frac{1}{1+x}, \quad x > -1$, compute $\frac{d}{dx}(F \circ \phi)(x)$ by using the chain rule.

(8) Compute d^2F/dt^2 for the function $F(t)$ in Exercise 7.

(9) (a) Show that the equation of a straight line which passes through the point P_1 at $t = 0$ and P_2 at $t = 1$ is

$$F(t) = P_1 + (P_2 - P_1)t.$$

(b) Find the equation of a straight line which passes through the point $P_1 = (1, 2, 3)$ at $t = 0$ and $P_2 = (1, -5, 0)$ at $t = 1$.

(c) Find the equation of a straight line which passes through the point P_1 at $t = t_1$ and P_2 at $t = t_2$.

(d) Apply this to find the equation of a straight line which passes through $P_1 = (-3, 1, -2)$ at $t = -1$ and $P_2 = (0, 2, 1)$ at $t = 2$. What is the slope of this line?

(10) Given a smooth curve all of whose tangent lines pass through a given point, prove that the curve is a straight line.

(11) Let $F\colon \mathbb{R}^1 \to \mathbb{R}^n$ define a smooth curve which does not pass through the origin. Show that the position vector $F(t)$ is orthogonal to the velocity vector at the point of the curve which is closest to the origin. Apply this to prove anew the well known fact that the radius vector to any point on a circle is perpendicular to the tangent vector at that point. [Hint: Why is it sufficient to minimize $\varphi(t) = \langle F(t), F(t)\rangle$?]

Chapter 8

Mappings from $\mathbb{R}^n$ to $\mathbb{R}$: The Differential Calculus

8.1 The Directional and Total Derivatives

.

Throughout this and the next chapter we shall consider functions which map $\mathbb{R}^n$ or a portion of it A, into $\mathbb{R}$. By the statement

$$f\colon A \to \mathbb{R}, \qquad A \subset \mathbb{R}^n$$

we mean that to every vector X in A, the function (operator, map, transformation) assigns a unique real number w. Thus $w = f(X)$ in this case is a map from vectors to numbers.

Two particular examples prove helpful in thinking conceptually about mappings of this type.

(1) *The temperature function.* $f\colon A \to \mathbb{R}$, where the set $A \subset \mathbb{R}^3$ is the room in which you are sitting. To every point X in the room, A, this function f assigns a number - the temperature $f(X)$ at X, $w = f(X)$.

(2) *The height function.* $f\colon A \to \mathbb{R}$, where the set A is some set in the plane $\mathbb{R}^2$. To every point X in A, this function f assigns a number - the height $f(X)$ of a surface (or manifold) M above that point. Thus, the set of all pairs $(X, f(x))$, $X \in A$, defines a portion of a surface, a surface in $\mathbb{R}^2 \times \mathbb{R} \cong \mathbb{R}^3$.

From the second example, it is clear that every function $f\colon A \subset \mathbb{R}^n \to \mathbb{R}$ may be regarded as the *graph* of a surface in $\mathbb{R}^n \times \mathbb{R} \cong \mathbb{R}^{n+1}$, the surface being regarded as all points in $\mathbb{R}^{n+1}$ of the form $(X, f(X))$, where $X \in A$. For example, the temperature function can be thought of as the graph of a surface in $\mathbb{R}^4$, the height of the surface $w = f(X)$ above X being the temperature at X. (Compare with the discussion from p. 322 bottom, to p. 324).

In concrete situations, the point $X \in \mathbb{R}^n$ is specified by giving its coordinates with respect to some fixed bases for $\mathbb{R}^n$ and $\mathbb{R}$. The particular coordinate system used depends on the geometry of the problem at hand. Rectangular symmetry calls for the standard rectangular coordinates, while polar coordinates are well suited to problems with circular symmetry. We shall meet these issues head-on a bit later.

If $X = (x_1, \ldots, x_n)$ with respect to some coordinates for $\mathbb{R}^n$, then we write $w = f(X) = f(x_1, \ldots, x_n)$. The points $(X, f(X))$ on the *graph* are $(x_1, \ldots, x_n,\ f(x_1, \ldots, x_n))$, which we may also write as $(x_1, \ldots, x_n, f)$ or else as $(x_1, \ldots, x_n, w)$. For low dimensional spaces, $\mathbb{R}^2$ or $\mathbb{R}^3$, it is convenient to avoid subscripts. In these situations we shall write $w = f(x, y)$ and $w = f(x, y, z)$ for mappings with domains in $\mathbb{R}^2$ and $\mathbb{R}^3$, respectively. We now examine some more specific examples.

EXAMPLES:

(1) $w = -\frac{1}{2}x + y - 1$. This function assigns to every point $X = (x, y)$ in $\mathbb{R}^2$ a number w in $\mathbb{R}$. We can represent the function, an affine mapping from $\mathbb{R}^2 \to \mathbb{R}$, as the graph of a plane in $\mathbb{R}^3$. The linear nature of the plane reflects the fact that the mapping is an affine mapping - a linear mapping except for a translation of the origin. More generally, the function $w = \alpha + a_1x_1 + a_2x_2 + \cdots + a_nx_n$, an affine mapping from $\mathbb{R}^n \to \mathbb{R}$, represents a plane in $\mathbb{R}^{n+1}$. In fact, this can be taken as the algebraic definition of a plane in $\mathbb{R}^{n+1}$. These affine functions are the simplest functions which map $\mathbb{R}^n$ into $\mathbb{R}$. Although we shall not, it is customary to abuse the nomenclature and refer to affine mappings as being linear. This is because they share most of the algebraic and geometric properties of proper linear mappings, as opposed to the honestly nonlinear mappings we will be treating as in the next examples.

(2) $w = x^2 + y^2$. This function assigns to every point $X = (x, y)$ in $\mathbb{R}^2$ a real number $w \in \mathbb{R}$. We can represent the function as the graph of a *paraboloid* of revolution, obtained by rotating the parabola $w = x^2$ about the w axis. If this paraboloid is cut by a plane parallel to the x, y plane, say $w = 2$, the intersection of these two surfaces is the circle $x^2 + y^2 = 2$.

(3) $w = -x^2 + y^2$. This function can be represented as the graph of a very fancy surface - a *hyperbolic paraboloid*. If this surface is cut by a plane parallel to the x, y plane, $w = c$, the intersection is the curve $c = -x^2 + y^2$. For $c > 0$, this curve is a hyperbola which opens about the y axis, while if $c < 0$, the curve is a hyperbola which opens about the x axis. For $c = 0$ we obtain two straight lines, $x = \pm y$ (see fig). The intersection of the surface with the plane $x = c$ is a parabola which opens upward in the y, w plane. Similarly, the intersection of the surface with the plane $y = c$ is a parabola which opens downward in the xw plane. This curve is rightly called a *saddle*, and the origin $(0, 0, 0)$ a *saddle point* (or mountain pass) since a particle can remain at rest at that point, or ii) move on the surface in one direction and go up, or iii) move on the surface in another direction and go down.

Let $f(X)$ be a function from vectors to numbers,

$$f\colon A \subset \mathbb{R}^n \to \mathbb{R}.$$

How can we define the notion of derivative for such functions? The derivative should measure the rate of change of $f(X)$ as X moves about. But if you think of $f(X)$ as the temperature function, it is clear that the temperature will change at different rates depending which direction you move. Thus, if you move across the room in the direction of the door, the temperature may decrease, while if you move up to the ceiling, the temperature will likely increase. Thus, the natural notion of a derivative is the rate of change in a particular direction - a *directional derivative.*

Let X_0 denote your position and $f(X_0)$ the temperature there. Take η to be a free vector, which we shall think of as pointing from X_0 to $X_0 + \eta$. We want to define the rate at which the temperature changes as you move from X_0 in the direction η toward $X_0 + \eta$. Since all points on the line joining X_0 to $X_0 + \eta$ are of the form $X_0 + \lambda\eta$, where λ is a real number, the difference $f(X_0 + \lambda\eta) - f(X_0)$ is the difference between the temperatures at $X_0 + \lambda\eta$ and at X_0.

DEFINITION: Let $f\colon A \subset \mathbb{R}^n \to \mathbb{R}$. The *derivative* of f at the interior point $X_0 \in A$ *with respect to the vector* η is

$$f'(X_0; \eta) = \lim_{\lambda \to 0} \frac{f(X_0 + \lambda\eta) - f(X_0)}{\lambda},$$

if the limit exists.

In the special case when $\eta = e$ is a *unit vector,* $\|e\| = 1$, we see that $\lambda = \|\lambda e\|$. Then $D_e f(X_0) := f'(X_0; e)$ is the instantaneous rate of change of f *per unit length* as X moves from X_0 toward $X_0 + 3$. This normalization to using only unit vectors is necessary to have a meaningful definition of a directional derivative. Thus, the *directional derivative* of f at $X_0 \in A$ in the direction of the *unit vector* e is the derivative with respect to the unit vector e. It measures how f changes as you move from X_0 to a point on the unit sphere about X_0. For theoretical purposes, the derivative of f with respect to any vector η is useful, while for practical purposes, the more restrictive notion of the directional derivative is needed.

EXAMPLE: 1 Find the directional derivative of $f(X) = x_1^2 - 2x_1x_2 + 3x_2$ at $X_0 = (1, 0)$ in the direction $\eta = (-1, 1)$. Note that η is not a unit vector. The unit vector is $e = \frac{\eta}{\|\eta\|} = (-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. Then

$$X_0 + \lambda e = (1, 0) + \lambda(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) = (1 - \frac{\lambda}{\sqrt{2}}, \frac{\lambda}{\sqrt{2}}),$$

so

$$f(X_0 + \lambda e) = (1 - \frac{\lambda}{\sqrt{2}})^2 - 2(1 - \frac{\lambda}{\sqrt{2}})(\frac{\lambda}{\sqrt{2}}) + 3(\frac{\lambda}{\sqrt{2}})$$

$$= 1 - \frac{\lambda}{\sqrt{2}} + \frac{3}{2}\lambda^2.$$

Thus,

$$\frac{f(X_0 + \lambda e) - f(X_0)}{\lambda} = \frac{1 - \frac{\lambda}{\sqrt{2}} + \frac{3}{2}\lambda^2 - 1}{\lambda} = \frac{-1}{\sqrt{2}} + \frac{3}{2}\lambda.$$

Therefore, the directional derivative $D_e f$ is

$$D_e f(X_0) = \lim_{\lambda \to 0} \frac{f(X_0 + \lambda e) - f(X_0)}{\lambda} = \frac{-1}{\sqrt{2}}.$$

In words, the rate of change of f at X_0 in the direction of the unit vector e is $-\frac{1}{\sqrt{2}}$. One qualitative conclusion we arrive at is that $f(X)$ decreases as X moves from X_0 in the direction e.

2. Compute $f'(X;\eta)$ if $f(X) = \langle X,\, AX\rangle$, where A is a self-adjoint transformation.

$$f(X + \lambda\eta) = \langle X + \lambda\eta,\, A(X + \lambda\eta)\rangle$$
$$= \langle X,\, AX\rangle + \lambda\langle \eta,\, AX\rangle + \lambda\langle X,\, A\eta\rangle + \lambda^2\langle \eta,\, A\eta\rangle$$

and since A is self-adjoint,

$$= \langle X,\, AX\rangle + 2\lambda\langle AX,\, \eta\rangle + \lambda^2\langle \eta,\, A\eta\rangle.$$

Thus,

$$f'(X,\eta) = \lim_{\lambda \to 0} \frac{f(X + \lambda\eta) - f(X)}{\lambda} = 2\langle AX,\, \eta\rangle.$$

In particular when $A = I$ is the identity operator, $f(X) = \|X\|^2$, we find $f'(X;\eta) = 2\langle X,\, \eta\rangle$

The directional derivatives of f in the particular direction of the coordinate axes $e_1 = (1,0,\ldots)$, $e_2 = (0,1,0,\ldots)$ have special names. They are called the *partial derivatives* of f. For example, the partial derivative of $f(X) = f(x_1, x_2, \ldots, x_n)$ at X_0 with respect to x_2 is

$$\frac{\partial f}{\partial x_2}(X_0) := f'(X_0; e_2) = \lim_{\lambda \to 0} \frac{f(X_0 + \lambda e_2) - f(X_0)}{\lambda}$$

There are many other competing notations, all of them being used. We shall list them shortly, after observing there is a simple way to compute these partial derivatives. Consider $f(X) = f(x-1, x_2, x_3)$. Then

$$\frac{\partial f}{\partial x_2}(X) = \lim_{\lambda \to 0} \frac{f(X + \lambda e_1) - f(X)}{\lambda}$$

Since $X + \lambda e_1 = (x_1, x_2, x_3) + \lambda(1,0,0) = (x_1 + \lambda, x_2, x_3)$ we have

$$\frac{\partial f}{\partial x_2}(X) = \lim_{\lambda \to 0} \frac{f(x_1 + \lambda, x_2, x_3) - f(x_1, x_2, x_3)}{\lambda}.$$

But this is the ordinary derivative of f with respect to the single variable x_1, while holding the other variables x_2 and x_3 fixed. Thus, $\partial f/\partial x_1$ can be computed by merely taking the ordinary one variable derivative of f with respect to x_1, pretending the other variables are constants.

EXAMPLE: If $f(X) = x_1^2 + x_1 e^{x_1 x_2}$, find the rate of change of f at the point X_0 in the directions $e_1 = (1,0)$ and $e_2 = (0,1)$. Thus, we want to compute $\frac{\partial f}{\partial x_1}(X_0)$ and $\frac{\partial f}{\partial x_2}(X_0)$.

$$\frac{\partial f}{\partial x_1} = 2x_1 + e^{x_1 x_2} + x_1 x_2 e^{x_1 x_2}$$

$$\frac{\partial f}{\partial x_2} = x_1^2 e^{x_1 x_2}$$

At the pint $X_0 = (2,-1)$, we have

$$\left.\frac{\partial f}{\partial x_1}\right|_{2,-1} = 4 - e^{-2}, \quad \left.\frac{\partial f}{\partial x_2}\right|_{(2,-1)} = 4e^{-2}.$$

Some common notation. If $w = f(x_1, x_2)$, then

$$\frac{\partial w}{\partial x_1} = \frac{\partial f}{\partial x_1} = D_1 f = f_1 = f_{x_1} = w_{x_1}$$

$$\frac{\partial w}{\partial x_2} = \frac{\partial f}{\partial x_2} = D_2 f = f_2 = f_{x_2} = w_{x_2}$$

If $f\colon A \subset \mathbb{R}^n \to \mathbb{R}$, then $\partial f/\partial x_j$ is another function of $X = (x_1, x_2, \ldots, x_n)$. It is then possible to take further partial derivatives.

EXAMPLE: Let $w = f(X) = x_1^2 + x_1 e^{x_1 x_2}$ as in the previous example. Then

$$w_{11} = f_{11} = f_{x_1 x_1} = \frac{\partial^2 f}{\partial x_1^2} = \frac{\partial}{\partial x_1}(\frac{\partial f}{\partial x_1}) = 2 + x_2 e^{x_1 x_2} = x_2 e^{x_1 x_2} + x_1 x_2^2 e^{x_1 x_2}$$

$$w_{12} = f_{12} = f_{x_1 x_2} = \frac{\partial^2}{\partial x_1 \partial x_2} = \frac{\partial}{\partial x_2}(\frac{\partial f}{\partial x_1}) = x_1 e^{x_1 x_2} = x_1 e^{x_1 x_2} + x_1^2 x_2 e^{x_1 x_2}$$

$$w_{21} = f_{21} = f_{x_2 x_1} = \frac{\partial^2 f}{\partial x_2 \partial x_1} = \frac{\partial}{\partial x_1}(\frac{\partial f}{\partial x_2}) = 2x_1 e^{x_1 x_2} = x_1^2 x_2 e^{x_1 x_2} = f_{12}$$

$$w_{22} = f_{22} = f_{x_2 x_2} = \frac{\partial^2 f}{\partial x_2^2} = \frac{\partial}{\partial x_2}(\frac{\partial f}{\partial x_2}) = x_1^3 e^{x_1 x_2}.$$

And even higher derivatives can be computed too, like

$$f_{221} = f_{x_2 x_2 x_1} = \frac{\partial^3 f}{\partial x_2^2 \partial x_1} = \frac{\partial}{\partial x_1}(\frac{\partial^2 f}{\partial x_2^2}) = 3x_1^2 e^{x_1 x_2} + x_1^3 x_2 e^{x_1 x_2}.$$

REMARK: From this one example, it appears possible that we always have $f_{12} = f_{21}$, that is $\frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial^2 f}{\partial x_2 \partial x_1}$. This is indeed the case *if* the second partial derivatives of f are continuous, but for lack of time we shall not prove it (see Exercise 6).

So far we have defined the directional derivative of a function $f\colon \mathbb{R}^n \to \mathbb{R}$ and called particular attention to those in the direction of the coordinate axes - the partial derivatives of f. Although the actual computation of the partial derivatives has been reduced to the formal procedure of computing ordinary derivatives, the computation of the directional derivative in an arbitrary direction must still be done by using the definition: the limit of a difference quotient. We shall now reduce the computation of all directional derivatives to a simple formal procedure. In order to do so, we shall introduce the concept of the *total derivative* for functions $f\colon A \subset \mathbb{R}^n \to \mathbb{R}^1$. This derivative will not be a directional derivative, but rather a more general object.

The motivating idea here is the important one of approximating a non-linear function f at a point X_0 by a linear function. If we think of the function $f(X)$ as defining a surface M in $\mathbb{R}^{n+1}$ with point $(X, f(X))$, then the picture is that of approximating the surface M near X_0 by a plane (or hyperplane) tangent to the surface at X_0. We want to write

$$f(X) \sim f(X_0) + L(X - X_0),$$

where L is a linear operator, $L\colon \mathbb{R}^n \to \mathbb{R}$, which may depend on the "base point" X_0. Of course, as $X \to X_0$ we want the accuracy to improve in the sense that the tangent plane should be a better approximation the closer X is to X_0. At $X = X_0$, the tangent plane $f(X_0) + L(X - X_0)$ and surface M touch since they both pass through the point $(X_0, f(X_0))$. Notice that the function $f(X_)) + L(X - X_0)$ is affine, so it does represent a plane surface.

Motivated by the above considerations, we can now make a reasonable

Definition: Let $f\colon A \subset \mathbb{R}^n \to \mathbb{R}$ and X_0 be an interior point of A. f is *differentiable at* X_0 if there exists a linear transformation $L\colon \mathbb{R}^n \to \mathbb{R}$ such that

$$\lim_{\|h\| \to 0} \frac{\|f(X_0 + h) - f(X_0) - Lh\|}{\|h\|} = 0,$$

for any vector h in some small ball about X_0 (so $f(X_0 + h)$ is defined). The operator L will usually depend on the base point X_0. If f is differentiable at X_0, we shall use the notation

$$\frac{df}{dX}(X_0) = f'(X_0) = L_{(X_0)} = L,$$

and refer to $f'(X_0)$ as the *total derivative* of f at X_0. [The notation $\nabla f(X_0)$ and grad $f(X_0)$, for *gradient*, are also used]. If $L = f'(X_0)$, a linear operator from $\mathbb{R}^n$ to $\mathbb{R}$, exists and depends continuously on the base point X_0 for all $X_0 \in A$, then f is said to be *continuously differentiable in* A, written $f \in C^1(A)$.

Remark: The condition that f be differentiable at X_0 can also be written in the following useful form:

$$f(X_0 + h) = f(X_0) + Lh + R(X_0, h)\|h\|, \tag{8-1}$$

where the remainder $R(X_0, h)$ has the property

$$\lim_{\|h\| \to 0} R(X_0, h) = 0.$$

This abstract operator L has the delightful property that it can be computed easily. But before telling you how, we should first prove for a given f there can be at most one linear operator L which is the total derivative.

Theorem 8.1 . *(Uniqueness of the total derivative). Let $f\colon A \to \mathbb{R}$ be differentiable at the interior point $X_0 \in A$. If L_1 and L_2 are linear operators both of which satisfy the conditions for the total derivative of f at X_0, then $L_1 = L_2$.*

Proof: Let $L = L_1 - L_2$. We shall show L is the zero operator. Since

$$Lh = L_1 h - L_2 h = [f(X_0 + h) - f(X_0) - L_2 h] - [f(X_0 + h) - f(X_0 - L_1 h],$$

by the triangle inequality we have

$$\|Lh\| \le \|f(X_0 + h) - f(X_0) - L_2 h\| + \|f(X_0 + h) - f(X_0) - L_1 h\|.$$

Consequently,

$$\lim_{\|h\| \to 0} \frac{\|Lh\|}{\|h\|} = 0.$$

To complete the proof, a trick is needed. Fix $\eta \neq 0$. If λ is a constant, $\lambda \to 0$, then $\|\lambda\eta\| \to 0$ so

$$\lim_{\|\lambda\| \to 0} \frac{\|L(\lambda\eta)\|}{\|\lambda\eta\|} = 0.$$

But since L is linear, $\|L\lambda\eta\| = \|\lambda L\eta\| = |\lambda|\ \|L\eta\|$, so the factor λ can be canceled in numerator and denominator. Thus the last equation is independent of λ, so $\|L\eta\|/\|\eta\| = 0$. Because $\eta \neq 0$, this implies $\|L\eta\| = 0$. Therefore L must be the zero operator.

Next, we give a method for computing L. Not only that, but we also find an easy way to compute the directional derivatives.

Theorem 8.2 . *Let $f\colon A \to \mathbb{R}$ be differentiable at the interior point $X_0 \in A$. Then a) the directional derivative of f at X_0 exists for every direction e and is given by the formula*

$$D_e f(X_0) = Le.$$

b) Moreover, if f is given in terms of coordinates, $f(X) = f(x_1, \ldots, x_n)$, then L is represented by the $1 \times n$ matrix

$$f'(X_0) = L = (f_{x_1}(X_0), \ldots, f_{x_n}(X_0)).$$

c) Consequently, the directional derivative is simply the product of this matrix L with the unit vector e, which can also be thought of as the scalar product of the $1 \times n$ matrix, a vector, and the vector e,

$$D_e f(X_0) = \langle f'(X_0),\, e\rangle.$$

PROOF: This falls out of the definitions. First

$$D_e f(X_0) = \lim_{\lambda\to 0} \frac{f(X_0+\lambda e) - f(X_0)}{\lambda}.$$

$$= \lim_{\lambda\to 0} \frac{f(X_0+\lambda e) - f(X_0) - L(\lambda e) + L(\lambda e)}{\lambda}.$$

Since $L(\lambda e) = \lambda Le$ and $\|\lambda e\| = \lambda$

$$= \lim_{\|\lambda e\|\to 0} \frac{f(X_0+\lambda e) - f(X_0) - L(\lambda e)}{\|\lambda e\|} + Le.$$

Because f is differentiable at X_0, the first term tends to zero. Thus proving the first part. To prove the last part, it is sufficient to observe that if $e = e_j$ is one of the coordinate vectors, then by definition $D_{e_j} f(X_0) := f_{x_j}$. Thus, if h is any vector $h = (h_1, \dots, h_n) = h_1 e_1 + \cdots + h_n e_n$, by the linearity of L we have

$$\begin{aligned} Lh = L(h_1 e_1 + \cdots + h_n e_n) &= h_1 Le_1 + \cdots + h_n Le_n \\ &= h_1 f_{x_1}(X_0) + \cdots + h_n f_{x_n}(X_0) \\ &= (f_{x_1}(X_0), \dots, f_{x_n}(X_0)) \begin{pmatrix} h_1 \\ \cdot \\ \cdot \\ \cdot \\ h_n \end{pmatrix}. \end{aligned}$$

Since h is any vector, we have shown L is represented by the given matrix.

REMARK: The theorem states that *if* f is differentiable, then all the partial derivatives exist and $f'(X_0) := L$ is represented by the above matrix. It does *not* state that if the partial derivatives exist, then f is differentiable. This is false (see Exercise 16). However, if the partial derivatives of f exist and are continuous, then f is differentiable. The last statement will be proved as Theorem 3.

EXAMPLE: The same one worked before (p. 573). Find the directional derivative of $f(X) = x_1^2 - 2x_1x_2 + 3x_2$ at $X_0 = (1, 0)$ in the direction $\eta = (-1, 1)$.

Since η is not a unit vector, we let $e = \frac{\eta}{\|\eta\|} = (-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. Now at a point X,

$$L = f'(X) = (f_{x_1}, f_{x_2}) = (2x_1 - 2x_2, -2x_1 + 3).$$

In particular, at $X = X_0 = (1, 0)$,

$$L = (2, -2+3) = (2, 1).$$

Therefore

$$D_e f(X_0) = Le = (2, 1)\begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = -\frac{1}{\sqrt{2}},$$

which checks with the answer found previously.

Consider the mapping $w = f(X)$, $X \in A \subset \mathbb{R}^n$, $w \in \mathbb{R}$ as defining a surface $M \subset \mathbb{R}^{n+1}$. It is now evident how to define the tangent plane to M at the point $(X_0, f(X_0))$, where $X_0 \in A$.

DEFINITION: Let $F\colon A \subset \mathbb{R}^n \to \mathbb{R}$ be a differentiable mapping, thus defining a surface M with points $(X, f(X))$, $X \in A$. The *tangent plane* to M at the point $(X_0, f(X_0))$, where $X_0 \in A$, is the surface defined by the affine mapping

$$\Phi(X) = f(X_0) + f'(X_0)(X - X_0).$$

or

$$\Phi(X) = f(X_0) + L(X - X_0), \quad \text{where} \quad L = f'(X_0),$$

Thus, the tangent plane to the surface defined by f is merely the "affine part" of f at X_0.

EXAMPLE: Consider the function $w = f(X) = 3 - x_1^2 - x_2^2$. This function defines a paraboloid (see fig.). Let us find the tangent plane to this surface at $(X_0, f(X_0))$, where $X_0 = (1, -1)$, so $f(X_0) = 3 - 1^2 - (-1)^2 = 1$. Also

$$f_{x_1}(X) = -2x_1, f_{x_2}(X) = -2x_2.$$

Thus

$$f'(X_0) = (f_{x_1}(X_0), f_{x_2}(X_0)) = (-2, 2).$$

Since $X - X_0 = (x_1, x_2) - (1, -1) = (x_1 - 1, x_2 + 1)$ we find the equation of the tangent plane is

$$\Phi(X) = 1 + (-2, 2)\begin{pmatrix} x_1 - 1 \\ x_2 + 1 \end{pmatrix} = 1 - 2(x_1 - 1) + 2(x_2 + 1),$$

or

$$\Phi(X) = 5 - 2x_1 + 2x^2.$$

This tangent plane is the unique plane with the property

$$\Phi(X_0) = f(X_0), \quad \text{and} \quad \Phi'(X_0) = f'(X_0).$$

Although we have given necessary conditions that a function be differentiable (all directional derivatives exist, in particular, all partial derivatives exist), we have not given sufficient conditions. The next theorem gives sufficient conditions for a function to be continuously differentiable.

Theorem 8.3 . *Let $f\colon A \subset \mathbb{R}^n \to \mathbb{R}$, where A is an open set. Then f is continuously differentiable throughout A if and only if all the partial derivatives of f exist and are continuous.*

PROOF: $\Rightarrow$ If f is continuously differentiable, then the partial derivatives exist by Theorem 2. Furthermore, for any X and Y in A,

$$f_{x_i}(X) - f_{x_i}(Y) = \langle f'(X),\, e_i\rangle - \langle f'(Y),\, e_i\rangle = \langle f'(X) - f'(Y),\, e_i\rangle.$$

Thus, applying the Schwartz inequality we find

$$|f_{x_i}(X) - f_{x_i}(Y)| \le \|f'(X) - f'(Y)\|.$$

The statement f is continuously differentiable means the vector $f'(X)$ is a continuous function of X. Therefore, given any $\epsilon > 0$ is a $\delta > 0$ such that $\|f'(X) - f'(Y)\| < \epsilon$ for all $\|X - Y\| < \delta$. For any $\epsilon > 0$, the inequality above shows $|f_{x_i}(X) - f_{x_i}(Y)|$ is also less than ϵ for the same δ. Consequently, f_{x_i} is continuous.

$\Leftarrow$. A little more difficult. The idea is to use the mean value theorem for functions of one variable. Let X and Y be points on A. To prove continuity at X, it is sufficient to restrict Y to being in some ball about X which is entirely in A (some ball does exist since A is open). For notational convenience, we take $n = 2$. Then

$$f(Y) - f(X) = f(Y) - f(Z) + f(Z) - f(X),$$

where Z is a point in A whose coordinates, except the first, are the same as X and whose coordinates, except the second, are the same as Y. By the one variable mean value theorem, there is a point $\tilde{X}$ between X and Z and point $\hat{X}$ between Y and Z such that

$$f(Z) - f(X) = \frac{\partial f}{\partial x_1}(\tilde{X})(y_1 - x_1),\ f(Y) - f(Z) = \frac{\partial f}{\partial x_2}(\hat{X})(y_2 - x_2).$$

Therefore

$$f(Y) - f(X) = \frac{\partial f}{\partial x_1}(\tilde{X})(y_1 - x_1) + \frac{\partial f}{\partial x_2}(\hat{X})(y_2 - x_2),$$

so

$$\begin{aligned} &f(Y) - f(X) - [f_{x_i}(X)(y_1 - x_1) + f_{x_2}(X)(y_2 - x_2)] \\ &= [f_{x_i}(\tilde{X}) - f_{x_i}(X)](y_1 - x_1) + [f_{x_2}(\hat{X}) - f_{x_2}(X)](y_2 - x_2). \end{aligned}$$

Therefore

$$\|f(Y) - f(X) - L(Y - X)\| \le \left|f_{x_i}(\tilde{X}) - f_{x_i}(X)\right|\ |y_1 - x_1| \quad \left|f_{x_2}(\hat{X}) - f_{x_2}(X)\right|\ |y_2 - x_2|$$

where we have written $L = (f_{x_1}(X), f_{x_2}(X))$. Since $|y_j - x_j| \le \|Y - X\|$, we see that

$$\frac{\|f(X) - f(Y) - L(Y - X)\|}{\|Y - X\|} \le \left|f_{x_1}(\tilde{X}) - f_{x_1}(X)\right| + \left|f_{x_2}(\hat{X}) - f_{x_2}(X)\right|.$$

Because f_{x_1} and f_{x_2} are continuous and $\|\tilde{X} - X\| < \|Y - X\|$, $\|\hat{X} - X\| < \|Y - X\|$, by making $\|Y - X\|$ sufficiently small the right side of the above inequality can be made arbitrarily small. This proves the limit as $\|Y - X\| \to 0$ of the expression on the left - exists and is zero. Since L is linear, the proof that f is differentiable is complete. The continuous differentiability is an immediate consequence of the linearity of L and the continuity of its components - the partial derivatives f_{x_i}.

Exercises

(1) i) Use the definition of the directional derivative to compute the given directional derivatives, ii) Check your answer by computing the directional derivative using the procedure of the Corollary to Theorem I.

(a) $f(x_1, x_2) = 1 - 2x_1 + 3x_2$, at $(2, -1)$ in the direction $(3, 4)$. [Answer: $+\frac{6}{4}$].
(b) $f(x, y) = e^{x+2y}$, at $(3, -2)$ in the direction $(1, 1)$. [Answer: $3e^{-1}/\sqrt{2}$].
(c) $f(u, v, w) = 3uv + uw - v^2$, at $(1, 1, 1)$ in the direction $(1, -2, 2)$.
(d) $f(x, y) = 1 - 3y + xy$ at $(0, 6)$ in the direction $(\frac{3}{5}, -\frac{4}{5})$.

(2) i) Compute *all* of the first and second partial derivatives for the following functions.

(a) $f(x_1, x_2) = x_1 + x_1 \sin 2x_1$
(b) $f(x_1, x_2, x_3) = x_1^2 x_2 + 2x_1\sqrt{x_3} - x_3$
(c) $f(x, y) = x^y$
(d) $f(x_1, x_2, \ldots, x_n) = a + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n$.
(e) $f(x_1, x_2, \ldots, x_n) = \sum_{i,j=1}^{n} a_{ij} x_i x_j = \langle X,\, AX\rangle$, where $a_{ij} = a_{ji}$ (first try the cases $n = 2$ and $n = 3$ to see what is happening).

ii) Find the $1 \times n$ matrix $f'(x)$.

(3) For the surfaces defined by the functions $f(X)$ listed below, find the equation of the tangent plane to the surface at the point $(X_0, f(X_0))$. Draw a sketch showing the surface and its tangent plane.

(a) $f(X) = x_1^2 + 3x_2^2 + 1, \quad X_0 = (0, 0)$.
(b) $f(X) = e^{x_1 x_2}, \quad X_0 = (0, 1)$
(c) $f(X) = x_1^2 \sin \pi x_2, \quad X_0 = (-1, \frac{1}{2})$
(d) $f(X) = -\frac{1}{2}x_1 + x_2 + 1, \quad X_0 = (2, 1)$
(e) $f(X) = x_1^2 + 2x_2^2 - x_1 x_3 + x_1, \quad X_0 = (1, -2, -1)$.
Why can't you sketch the surface defined by this function?

(4) Let $f(X)$ and $g(X)$ both map $A \subset \mathbb{R}^n \to \mathbb{R}^1$. If f and g are differentiable for all $X \in A$, prove

(a) $\frac{d}{dX}[af(X)+bg(X)] = a\frac{df}{dX}(X) + b\frac{dg}{dX}(X)$ (Linearity), where a and b are constants.

(b) $\frac{d}{dX}[f(X)g(X)] = f(X)\frac{dg}{dX}(X) + g(X)\frac{df}{dX}(X)$

(c) $\frac{d}{dX}\left[\frac{f(X)}{g(X)}\right] = \frac{g(X)f'(X)-f(X)g'(X)}{g^2(X)}$, if $g(X) \neq 0$.

(5) Use the rules (a-c) of Exercise 4 to compute $\frac{d}{dX}[2f-3g]$, $\frac{d}{dX}[f\cdot g]$, and $\frac{d}{dX}[\frac{f}{g}]$, where $f(X) = f(x_1,x_2) = 1 - x_1 + x_1x_2$, and $g(X) = g(x_1,x_2) = e^{x_1-x_2}$.

(6) Let $f(X) = f(x,y) = \begin{cases} \frac{xy(x^2-y^2)}{x^2+y^2}, & X=(x,y)\neq 0 \\ 0 & X = 0 \end{cases}$

Prove

(a) f, f_x, f_y are continuous for all $X \in \mathbb{R}^2$. [Hint: Prove and use $2xy \le x^2+y^2$].

(b) f_{xy} and f_{yx} exist for all $X \in \mathbb{R}^2$, and are continuous *except* at the origin.

(c) $f_{xy}(0) = 1, f_{yx}(0) = -1$, so $f_{xy}(0) \neq f_{yx}(0)$ (cf. Remark p. 577).

(7) Let $f\colon A \subset \mathbb{R}^n \to \mathbb{R}$ be a differentiable map. Prove it is necessarily continuous. [Hint: This is a simple consequence of the definition in the form (1)].

(8) Let $f\colon A \subset \mathbb{R}^n \to \mathbb{R}$ be a continuous map. We say f has a *local maximum* at the point X_0 interior to A if $f(X_0) \ge f(X)$ for all X in some sufficiently small ball about X_0. If we assume f is continuously differentiable, more can be said.

(a) If f as above has a local maximum at the point X_0, prove $\langle f'(X_0,\, X-X_0)\rangle + R(X_0,X)\|X-X_0\| \le 0$ for all X is some small ball about X_0.

(b) Use the property of $R(X_0,X)$ to conclude the stronger statement

$$\langle f'(X_0),\,(X-X_0)\rangle \le 0.$$

for all X in some small ball about X_0.

(c) Observe the statement must also hold for the vector $X_0 - X$, which points in the direction opposite to $X - X_0$, to conclude

$$\langle f'(X_0),\,(X-X_0)\rangle \ge 0,$$

and hence that in fact

$$\langle f'(X_0),\, Z\rangle = 0,$$

for all vectors $Z = X - X_0$.

(d) Finally, show that at a maximum,

$$f'(X_0) = 0.$$

(9) (a) Find the equation of the plane which is tangent at the point $X_0 = (2,6,3)$ to the surface consisting of the points $(X, f(X))$, where

$$f(X) = f(x,y,z) = (x^2+y^2+z^2)^{1/2}.$$

(b) Use the tangent plane found above to find the approximate value of

$$((2.01)^2 + (5.98)^2 + (2.99)^2)^{1/2}.$$

(10) Assume the continuously differentiable function $f(X)$ has a zero derivative, $f'(X) \equiv 0$, for X in some ball in $\mathbb{R}^n$. Prove that $f(X) \equiv$ constant throughout the ball.

(11) (a) Show the following functions satisfy the two dimensional *Laplace equation*

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

i) $u(x,y) = x^2 - y^2 - 3xy + 5y - 6$
ii) $u(x,y) = \log(x^2+y^2)$, except at the origin, $(x,y) = 0$.
iii) $u(x,y) = e^x \sin y$

(b) Show the following functions satisfy the one (space) dimensional *wave equation*

$$u_{tt} = c^2 u_{xx}, \quad c \equiv \text{constant}$$

[Here t is time and x is space; c is the velocity of light, sound, etc.]
i) $u(x,y) = e^{x-ct} - 2e^{x+ct}$
ii) $u(x,y) = 2(x+ct)^2 + \sin 2(x-ct)$.

(12) Let $f\colon A \subset \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable throughout A. If $X_0 \in A$ is not a critical point of f, so $f'(X_0) \neq 0$, prove the directional derivative at X_0 is greatest in the direction $e_{\max} := f'(X_0)/\|f'(X_0)\|$, and least in the opposite direction, $e_{\min} := -e_{\max}$. [Hint: Use the Schwarz inequality.]

(13) Consider the function $f(X) = f(x,y) = \begin{cases} \frac{xy}{x^2+y^2}, & X=(x,y)\neq 0 \\ 0, & X = 0 \end{cases}$

Since F is the quotient of two continuous functions, it is continuous except possibly at the origin, where the denominator vanishes. Show that $f(X)$ is *not* continuous at the origin by finding $\lim f(X)$ as $X \to 0$ along paths 1 and 2, and showing that

$$\lim_{\substack{X\to 0 \\ \text{path1}}} f(X) \neq \lim_{\substack{X\to 0 \\ \text{path2}}} f(X).$$

(14) Let L be the partial differential operator defined by

$$Lu = \frac{\partial^2 u}{\partial x^2} - 5\frac{\partial^2 u}{\partial x \partial y} + 6\frac{\partial^2 u}{\partial y^2}.$$

Show that

$$L[e^{\alpha x+\beta y}] = p(\alpha,\beta)e^{\alpha x+\beta y},$$

where $p(\alpha,\beta)$ is a polynomial in α and β. Find a solution of the linear homogeneous partial differential equation $Lu = 0$. Find an infinite number of solutions of $Lu = 0$, one for each value of α, by choosing α to depend on β in a particular way. [Answer: $e^{2\beta x+\beta y}$ and $e^{3\beta x+\beta y}$ are solutions for any β].

(15) The two equations

$$x = e^u \cos v$$

$$y = e^u \sin v$$

define $u = f(x,y)$ and $v = g(x,y)$. Find the functions f and g for $x > 0$. Compute $f'(X)$ and $g'(X)$ and show $f'(X) \perp g'(X)$.

(16) This exercise gives an example in which the first partial derivatives of a function exist but the function is not continuous, let alone differentiable. Let

$$f(X) = f(x,y) = \begin{cases} \frac{xy^2}{x^2+y^4}, & X = (x,y) \neq 0 \\ 0, & X = 0. \end{cases}$$

(a) If $\cos\alpha \neq 0$, prove the directional derivative at the origin in the direction $e = (\cos\alpha, \sin\alpha)$ exists and is

$$D_e f(0) = \frac{2\sin^2\alpha}{\cos\alpha}, \qquad \cos\alpha \neq 0$$

while if $\cos\alpha = 0$,

$$D_e f(0) = 0, \qquad \cos\alpha = 0.$$

(b) Prove f is discontinuous at the origin by showing $\lim_{X\to 0} f(X)$ has two different values along the two paths in the figure. Then appeal to exercise 7 to conclude f is not differentiable.

(17) (a) Let $P(X), X \in \mathbb{R}^n$, be a polynomial of degree N, that is,

$$P(X) = \sum_{k_1+k_2+\cdots+k_n \leq N} a_{k_1,\ldots,k_n} x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n},$$

where $k_1, k_2, \ldots, k_n$ are all non-negative integers. Prove $P(\alpha)$ is continuously differentiable. [Hint: How do you prove a polynomial in one variable is continuously differentiable.]

(b) Let $R(X), X \in \mathbb{R}^n$, be a rational function - that is, the quotient of two polynomials. Prove $R(X)$ is continuously differentiable whenever the denominator is not zero.

(18) If $f\colon \mathbb{R}^1 \to \mathbb{R}^1$, show that the definition of differentiability on page 578 coincides with the usual one.

8.2 The Mean Value Theorem. Local Extrema.

Although the full "chain rule" will not be proved until Chapter 10, we shall need a very special and elementary case to develop the main features of the theory of mappings from $\mathbb{R}^n$ to $\mathbb{R}$. Let $f\colon A \subset \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function at all interior points of A. Take X and Z to be fixed interior points of A. Let $\phi(t) = f(X+tZ)$. We want to compute

$$\frac{d}{dt}\phi(t) = \frac{d}{dt}f(X+tZ)$$

that is, the rate of change of $f(X)$ at the point $X+tZ$ as X varies along the line joining X to Z.

Theorem 8.4 . *Let $f\colon A \to \mathbb{R}$ be a differentiable function throughout A. If X and Z are two interior points of A, and if the line segment joining them is in A, then*

$$\frac{d}{dt}f(X+tZ) = f'(X+tZ)Z, \quad t \in (0,1).$$

By the product $f'(Y)Z$ we mean matrix multiplication.

PROOF: For fixed X and Z, the function $\phi(t) := f(X+tZ)$ an ordinary scalar valued function of the one variable t. Thus

$$\begin{aligned} \frac{d}{dt}\phi(t) &= \lim_{\lambda\to 0}\frac{\phi(t_\lambda)-\phi(t)}{\lambda} \\ &= \lim_{\lambda\to 0}\frac{f(X+tZ+\lambda Z)-f(X+tZ)}{\lambda} \\ &= \lim_{\lambda\to 0}\frac{f(X+tZ+\lambda Z)-f(X+tZ)-f'(X+tZ)(\lambda Z)+f'(X+tZ)(\lambda Z)}{\lambda} \end{aligned}$$

Since f is differentiable at $X+tZ$, then as $\lambda \to 0$ the first three terms tend to zero. The factor λ in the last term cancels. Therefore

$$\frac{d}{dt}f(X+tZ) = \lim_{\lambda\to 0} f'(X+tZ)Z = f'(X+tZ)Z,$$

as claimed.

An easy consequence is

Theorem 8.5 *(The Mean Value Theorem). Let $f\colon A \to \mathbb{R}$, where A is an open convex set in $\mathbb{R}^n$, that is, if X and Y are any points in Z, then the straight line segment joining X and Y is in A too. If f is differentiable in A, there is a point Z on the segment joining X and Y such that*

$$f(Y) - f(X) = f'(Z)(Y-X).$$

If, moreover, f' is bounded by some constant C, $\|f'(X)\| \leq C$ for all $X \in A$, then

$$|f(Y) - f(X)| \leq C\|Y-X\|$$

A FIGURE GOES HERE

PROOF: Every point on the segment joining X and Y is of the form $X+t(Y-X)$, where $t\in[0,1]$. Consider the function $\phi(t)$ of one variable,

$$\phi(t)=f(X+t(Y-X)).$$

Theorem 4 states ϕ is differentiable. Therefore, by the one variable mean value theorem, there is a number t_0 in the interval $(0,1)$ such that $\phi(1)-\phi(0)=\phi'(t_0)$. But $\phi(1)=f(Y)$, $\phi(0)=f(X)$ and, by Theorem 4, $\phi'(t_0)=f'(X+t_0(Y-X))(Y-X)$. Letting $Z=X+t_0(Y-X)$, a point on the segment joining X to Y, we conclude

$$f(Y)-f(X)=f'(Z)(Y-X).$$

The second part of the theorem follows by applying the Schwarz inequality to the function $f'(Z)(Y-X)$ which can be written as $\langle f'(Z),\,(Y-X)\rangle$. Then

$$\langle f'(Z),\,Y-X\rangle\le\|f'(Z)\|\,\|Y-X\|.$$

Therefore if $\|f'(Z)\|\le C$ for all $Z\in A$, we find

$$|f(Y)-f(X)|\le C\|Y-X\|.$$

Corollary 8.6 *Let $f\colon A\to\mathbb{R}$ be a differentiable map and A an open connected set in $\mathbb{R}^n$ (by a connected open set we mean it is possible to join any two points in A by a polygonal curve contained in A). If $f'(X)\equiv 0$ for every $X\in A$, that is, if $f_{x_1}(X)=\ldots=f_{x_n}(X)=0$, then $f(X)\equiv c$, c a constant.*

PROOF: If A is convex, say a ball, this is an immediate consequence of the second part of the mean value theorem, for $\|f'(X)\|=0$ so $|f(Y)-f(X)|=0$. Thus $f(Y)=f(X)=$ constant for any two points X and Y. The requirement that A is connected is to exclude the possibility that A consists of two (or more) disjoint sets, in which case, all we can conclude is that f is constant on each connected part, but not necessarily the same constant. However, if A is connected, then any two points in A can be joined by a polygonal curve which is contained in A. Consider some straight line segment in this curve. By the mean value theorem, f must be constant on it. In particular, it has the same value at both end points. Checking the beginning and end of the whole polygonal curve, we find that $f(X)=f(Y)$. Because X and Y were any points, we are done.

It is not at all difficult to generalize the mean value theorem to Taylor's theorem and then to power series for functions of several variables. The only problem is one of notation, and that is a problem. As a compromise, we will prove the Taylor theorem - but only the first two terms for functions of three variables $f(x,y,z)$.

Just as in the mean value theorem, the idea is to reduce the problem to a function $\phi(t)$ of one real variable, because we do know the result for these functions. Let f be differentiable in some open set $A\subset\mathbb{R}^3$ and X_0 a point in A. If X_0+h is also in A, we would like to express $f(X_0+h)$ in terms of f and its derivatives at X_0. Fix X_0 and h and consider the real valued function $\phi(t)$ of one variable defined by

$$\phi(t)=f(X_0+th),\qquad t\in[0,1].$$

Then by Theorem 4,

$$\phi'(t)=f'(X_0+th)h=f_x(X_0+th)h_1+f_y(X_0+th)h_2+f_z(X_0+th)h_3,$$

where $h=(h_1,h_2,h_3)$. Since each of the partial derivatives are maps from A to $\mathbb{R}$, they can be differentiated in the same way f was. So can a sum of such functions. Thus

$$\begin{aligned}\phi''(t)&=\frac{d}{dt}[f_x(X_0+th)h_1+\cdots+f_z(X_0+th)h_n]\\&=f_{xx}(X_0+th)h_1h_1+f_{xy}(X_0+th)h_1h_2+f_{xz}(X_0+th)h_1h_3\\&+f_{yx}(X_0+th)h_2h_1+f_{yy}(X_0+th)h_2h_2+f_{yz}(X_0+th)h_2h_3\\&+f_{zx}(X_0+th)h_3h_1+f_{zy}(X_0+th)h_3h_2+f_{zz}(X_0+th)h_3h_2.\end{aligned}$$

If we introduce a matrix $H(X)$, the *Hessian matrix*, whose elements are $\dfrac{\partial^2 f(X)}{\partial x_i\partial x_j}$, $\phi''(t)$ can be written as

$$\phi''(t)=\langle h,\,H(X_0+th)h\rangle.$$

We remark that if f is sufficiently differentiable (two continuous derivatives is enough), then the Hessian matrix is self-adjoint since $f_{x_ix_j}=f_{x_jx_i}$, as we mentioned - but did not prove - earlier. If $\phi(t)$ is twice differentiable, by Taylor's theorem for functions of one variable, we know that

$$\phi(1)=\phi(0)+\phi'(0)+\frac{1}{2!}\phi''(\tau),\quad \tau\in(0,1).$$

Substituting into this formula, we find

$$f(X_0+h)=f(X_0)+f'(X_0)h+\frac{1}{2!}\langle h,\,H(X_0+\tau h)h\rangle.$$

Let us summarize. We have proved

Theorem 8.7 (TAYLOR'S THEOREM WITH TWO TERMS) . *Let $f\colon A\to\mathbb{R}$, where A is an open connected set in $\mathbb{R}^n$. Assume f has two continuous derivatives - that is, all the second partial derivatives of f exist and are continuous. If X_0 is in A and X_0+h is in a ball about X_0 in A, then*

$$f(X_0+h)=f(X_0)+f'(X_0)h+\frac{1}{2!}\langle h,\,H(X_0+\tau h)h\rangle,$$

where $H(X)=((\dfrac{\partial^2 f}{\partial x_1\partial x_j}))$ is the $n\times n$ Hessian matrix and $\tau\in(0,1)$.

Letting $X=X_0+h$ and $Z=X_0+\tau h$, Z being a point on the line segment joining X_0 to X, this reads

$$f(X)=f(X_0)+f'(X_0)(X-X_0)+\frac{1}{2!}\langle X-X_0,\,H(Z)(X-X_0)\rangle,$$

or, in more detail,

$$f(X)=f(X_0)+\sum_{i=1}^{n}\frac{\partial f(X_0)}{\partial x_i}(x_i-x_i^0)+\frac{1}{2}\sum_{i=j}^{n}\sum_{j=1}^{n}\frac{\partial^2 f(Z)}{\partial x_i\partial x_j}(x_i-x_i^0)(x_j-x_j^0).$$

EXAMPLE: Find the first two terms in the Taylor expansion for the function $f(X) = f(x,y) = 5 + (2x-y)^3$ about the point $X_0 = (1,3)$.

We compute

$$f_x(X) = 6(2x-y)^2, \quad f_y(X) = -3(2x-y)^2$$

$$f_{xx}(X) = 24(2x-y), f_{xy}(X) = f_{yx}(X) = -12(2x-y), f_{yy}(X) = 6(2x-y).$$

Therefore $f(X_0) = 4, f_x(X_0) = 6, f_y(X_0) = -3$, so

$$f(X) = 4 + (6,-3)\begin{pmatrix} x-1 \\ y-3 \end{pmatrix} + \frac{1}{2}(x-1, y-3)\begin{pmatrix} 2\xi - \eta & -12(2\xi-\eta) \\ -12(2\xi-\eta) & 6(2\xi-\eta) \end{pmatrix}\begin{pmatrix} x-1 \\ y-3 \end{pmatrix}$$

where $Z = (\xi, \eta)$ is a point on the segment between $X_0 = (1,3)$ and $X = (x,y)$. Written out, the above equation reads,

$$f(x,y) = 4 + 6(x-1) - 3(y-3) + \frac{1}{2}[f_{xx}(x-1)^2 + 2f_{xy}(x-1)(y-3) + f_{yy}(y-x)^2],$$

where the second derivatives are evaluated at $Z = (\xi, \eta)$.

We are now in a position to examine the extrema of functions of several variables. Finding the maxima and minima of functions is important for several reasons. First of all, there is the vague emotional feeling that all patterns of action should maximize or minimize something. Second, we can investigate a complicated geometrical object by the relatively easy procedure of finding the local maxima and minima. Without further mention, for the balance of this section $f(X)$ will be a twice continuously differentiable function which maps the open set $A \subset \mathbb{R}^n$ into $\mathbb{R}$.

DEFINITION: A function $f \colon A \to \mathbb{R}$ has a *local maximum* at the interior point $X_0 \in A$ if, for all X in some open ball about X_0

$$f(X) \le f(X_0).$$

f has a *local minimum* at X_0 if for all X in some open ball about X_0

$$f(X) \ge f(X_0).$$

If f has a local maximum or minimum at X_0, is $f'(X_0) = 0$? Certainly.

Theorem 8.8 . *If f has a local maximum or minimum at X_0, then $f'(X_0) = 0$. In coordinates, this means all the partial derivatives vanish at X_0,*

$$\frac{\partial f}{\partial x_1 1}(X_0) = \frac{\partial f}{\partial x_2}(X_0) = \cdots = \frac{\partial f}{\partial x_n}(X_0) = 0.$$

PROOF: Let η be any fixed vector. Then the function $\phi(t)$ of one variable

$$\phi(t) = f(X_0 + t\eta)$$

has a local maximum or minimum at $t = 0$. Consequently $\phi'(0) = 0$. But by Theorem 4, $\phi'(0) = f'(X_0)\eta$ which we may write as $\langle f'(X_0), \eta \rangle$. Thus $\langle f'(X_0), \eta \rangle = 0$, so the vector $f'(X_0)$ is orthogonal to η. Since η was any vector, we conclude that $f'(X_0) = 0$.

The derivative $f'(X_0)$ may vanish at points other than maxima or minima. An example is the "saddle point" of the hyperbolic paraboloid at the beginning of Section 1. All points where f' vanishes are called *critical points* or *stationary points* of f. Let us give a precise definition of a saddle point. f has a *saddle point* at X_0 if X_0 is a critical point of f and if every ball about X_0 contains points X_1 and X_2 such that $f(X_1) < f(X_0)$ and $f(X_2) > f(X_0)$. Thus, every critical point is either a local maximum, minimum, or saddle point.

There is a more intuitive way to prove Theorem 7. If e is a unit vector, then by Theorem 2, the directional derivative at X in the direction e is $D_e f(X) = \langle f'(X), e \rangle$. In what way should you move so f increases fastest? By the Schwartz inequality, we find

$$|D_e f(X)| \le \|f'(X)\| \quad \|e\| = \|f'(X)\|,$$

with equality if and only if the vectors e and $f'(X)$ are parallel. Thus, *the directional derivative is largest when e has the same direction as $f'(X)$, and smallest when e has the opposite direction*, $e_{\max} = f'(X)/\|f'(X)\|$, $e_{\min} = -e_{\max}$,

$$D_{e_{\max}} f(X) = \|f'(X)\|, \, D_{e_{\min}} f(X) = -\|f'(X)\|.$$

If X_0 is a local maximum of f, then $f'(X_0)$ must be zero, for otherwise you could move in the direction of $f'(X_0)$ and increase the value of f. Similarly, if X_0 is a local minimum, $f'(X_0)$ must be zero.

Once we know X_0 is a critical point of f, $f'(X_0) = 0$, an effective criterion is needed to determine if X_0 is a local maximum, minimum, or saddle point for f. In elementary calculus, the sign of the second derivative was used. Our next theorem generalizes this test.

The idea is essentially the same as in the one variable case (p. 104a-c). If f has a local maxima or minima, the tangent plane to the surface whose points are $(X, f(X))$ is horizontal, that is, $f'(X_0) = 0$. Thus, near X_0 the quadratic terms - the next lowest power in the Taylor expansion of f about X_0—will determine the behavior of f near X_0. Let X_0 be the origin and take $f(X) = f(x,y)$ to be a function of two variables with $f(0) = 0$. Then near $X_0 = 0$, by Taylor's theorem, we have

$$f(x,y) \sim \frac{1}{2}[ax^2 + 2bxy + cy^2],$$

where $a = f_{xx}(0)$, $b = f_{xy}(0)$, and $c = f_{yy}(0)$. The nature of the quadratic form

$$Q(X) = ax^2 + 2bxy + cy^2$$

has already been determined. If $Q(X)$ is positive definite, then $Q(X) > 0$ for $X \neq 0$. Since $f(x,y) \sim Q(X)$, this means $f(x,y)$ is positive near the origin. Because $f(0,0) = 0$, this implies the origin is a minimum for f.

Instead of completing and rigorously justifying this special case, we shall immediately treat the general situation.

Theorem 8.9 . *Assume the twice continuously differentiable function $f\colon A \to \mathbb{R}$ has a critical point at an interior point X_0 of $A \subset \mathbb{R}^n$, $f'(X_0) = 0$. Let $H(X_0)$ be the Hessian matrix $\left(\left(\frac{\partial^2 f}{\partial x_i \partial x_j}(X_0)\right)\right)$ evaluated at X_0.*

(a) *If $H(X_0)$ is positive definite, then f has a local minimum at X_0.*

(b) *If $H(X_0)$ is negative definite, then f has a local maximum at X_0.*

(c) *If at least two of the diagonal elements of $H(X_0)$, $f_{x_1x_1}(X_0), \dots, f_{x_nx_n}(X_0)$ have different signs, then X_0 is a saddle point.*

(d) *Otherwise the test fails.*

Proof: If X_0 is a critical point for f, then Taylor's theorem (Theorem 6) states

$$f(X_0 + \eta) = f(X_0) + \frac{1}{2}\langle \eta,\, H(Z)\eta \rangle$$

where Z is between X_0 and $X_0 + \eta$. The linear term has been dropped since $f'(X_0) = 0$.

As in the proof of Taylor's theorem, let

$$\phi(t) = f(X_0 + t\eta).$$

Then

$$\phi''(t) = \langle \eta,\, H(X_0 + t\eta)\eta \rangle.$$

Since the second derivatives of f are assumed to be continuous, the function $\phi''(t)$ is a continuous function of t. Consequently, if $\phi''(0)$ is positive then $\phi''(t)$ is also positive for all t sufficiently close to zero (Theorem I p. 29b). Because $\phi''(0) = \langle \eta,\, H(X_0)\eta \rangle$ and $\phi''(\tau) = \langle \eta,\, H(Z)\eta \rangle$, where $Z = X_0 + \tau\eta$, this implies if H is positive definite at X_0, it is also positive definite at Z when Z is close to X_0.

Assuming $H(X_0)$ is positive definite, we see that for all η sufficiently small, $H(Z)$ is positive definite. Therefore,

$$f(X_0 + \eta) - f(X_0) = \frac{1}{2}\langle \eta,\, H(Z)\eta \rangle > 0, \quad \eta \neq 0,$$

that is

$$f(X_0 + \eta) - f(X_0) > 0$$

for all η is some small ball about X_0. Thus f has a local minimum at X_0.

If $H(X_0)$ is negative definite, the same proof with trivial modifications works. Another way to complete the proof is to apply part a) to the function $g(X) := -f(X)$. The Hessian for g at X_0 will be $-H(X_0)$ which is positive definite (since $H(X_0)$ was negative definite). Thus g has a local minimum at X_0 so $f\colon = -g$ has a local maximum at X_0.

If any two of the diagonal elements of $H(X_0)$ have opposite sign, say $f_{x_1x_1}(X_0) > 0$ and $f_{x_2x_2}(X_0) < 0$, then for $\eta = \lambda e_1 = (\lambda, 0, 0, \dots, 0)$, λ any real number, we find $\langle \eta,\, H(X_0)\eta \rangle = \lambda^2 f_{x_1x_1}(X_0) > 0$, while for $\eta = \lambda e_2 = (0, \lambda, 0, \dots, 0)$ $\langle \eta,\, H(X_0)\eta \rangle = \lambda^2 f_{x_2x_2}(X_0) < 0$. Therefore the quadratic form $\langle \eta,\, H(X_0)\eta \rangle$ assumes positive and negative values in any ball about X_0, proving X_0 is a saddle point.

Since this theorem reduces the investigation of the nature of a critical point to testing if a matrix is positive or negative definite, it would do well in this context to repeat Theorem A (p. 386d) which tells us when a 2×2 matrix is positive definite.

Corollary 8.10 . *Let X_0 be a critical point for the function of two variables $f(x,y)$ with Hessian matrix*

$$H(X_0) = \begin{pmatrix} f_{xx}(X_0) & f_{xy}(X_0) \\ f_{xy}(X_0) & f_{yy}(X_0) \end{pmatrix}.$$

(a) *If $\det H(X_0) > 0$ and $f_{xx}(X_0) > 0$, then f has a local minimum at X_0.*

(b) *If $\det H(X_0) > 0$ and $f_{xx}(X_0) < 0$, then f has a local maximum at X_0.*

(c) *If $\det H(X_0) < 0$, then f has a saddle point at X_0 (this is a stronger statement than part c of Theorem 8).*

Proof: Since these merely join Theorem A (p. 386d) with Theorem 8, the proof is done.

Examples:

(1) Find and classify the critical points of the function $w = f(x,y) := 3 - x^2 - 4y^2 + 2x$.

A sketch of the surface with points $(x, y, f(x,y))$, a paraboloid, is at the right. At a critical point $f'(X) = 0$, that is, $f_x = 0$, $f_y = 0$. Since

$$f_x = -2x + 2, f_y = -8y,$$

at a critical point

$$-2x + 2 = 0, \quad -8y = 0.$$

There is therefore only one critical point, $X_0 = (1, 0)$. We look at the Hessian to determine the nature of the critical point. Because $f_{xx} = -2$, $f_{xy} = f_{yx} = 0$, $f_{yy} = -8$,

$$H(X_0) = \begin{pmatrix} -2 & 0 \\ 0 & -8 \end{pmatrix}.$$

Since $\det H(X_0) = 16 > 0$ and $f_{xx}(X_0) = -2 < 0, H(X_0)$ is negative definite so $X_0 = (1, 0)$ is a local maximum for the function, and at that point $f(X_0) = 4$.

(2) Find and classify the critical points of $w = f(x, y) = -x^2 + y^2$.

The surface $(x, y, f(x, y))$ is a hyperbolic paraboloid. We expect a saddle point at the origin. At a critical point

$$f_x = -2x = 0, \qquad f_y = 2y = 0.$$

Thus the origin $(0, 0)$ is the only critical point. Since

$$H(x, y) = \begin{pmatrix} -2 & 0 \\ 0 & 2 \end{pmatrix},$$

and $\det H(0, 0) = -4 < 0$, the origin is a saddle point. This also follows from the observation that the diagonal elements have different signs.

(3) Find and classify the critical points of

$$w = f(x, y) = [x^2 + (y + 1)^2][x^2 + (y - 1)^2].$$

At a critical point,

$$f_x = 2x[x^2 + (y - 1)^2] + 2x[x^2 + (y + 1)^2] = 0$$

and

$$f_y = 2(y + 1)[x^2 + (y - 1)^2] + 2(y - 1)[x^2 + (y + 1)^2] = 0.$$

The first equation implies $x = 0$. Substituting this into the second we find $y = 0, y = 1, y = -1$. Thus there are three critical points

$$X_1 = (0, 0), \quad X_2 = (0, 1), \quad X_3 = (0, -1).$$

We must evaluate the Hessian matrix at these points. Since

$$f_{xx} = 12x^2 + 4y^2 + 4, \quad f_{xy} = 9xy, \quad f_{yy} = 4x^2 + 12y^2 = -4,$$

$$H(X_1) = \begin{pmatrix} 4 & 0 \\ 0 & -4 \end{pmatrix}, \qquad H(X_2) = \begin{pmatrix} 8 & 0 \\ 0 & 8 \end{pmatrix} = H(X_3).$$

Because $\det H(X_1) = -16 < 0$, $X_1 = (0, 0)$ is a saddle point. Because $\det H(X_2) > \det H(X_3) = 64 > 0$ and $f_{xx}(X_2) = f_{xx}(X_3) = 8 > 0$, both $X_2 = (0, 1)$ and $X_3 = (0, -1)$ are local minima. To complete the computation, we find $f(X_1) = 1, f(X_2) = 0, f(X_3) = 0$. A sketch of the surface is at the right.

(4) Find and classify the critical points of

$$w = f(x, y, z) = 1 - 2x + 3x^2 - xy + xz - z^2 + 4z + y^2 + 2yz.$$

At a critical point,

$$f_x = -2 + 6x - y + z, f_y = -x + 2y + 2z, f_z = x - 2z + 4 + 2y.$$

Solving these equations, we find only one critical point, $X_0 = (0, -1, 1)$, where $f(X_0) = 3$. Since

$$f_{xx} = 6, f_{xy} = -1, f_{xz} = 1\, f_{yy} = 2, f_{yz} = 2, f_{zz} = -2,$$

then

$$H(X) = \begin{pmatrix} 6 & -1 & 1 \\ -1 & 2 & 2 \\ 1 & 2 & -2 \end{pmatrix}.$$

Because the diagonal elements $6, 2, -2$ are not all of the same sign, by part c of the theorem, the critical point $X_0 = (0, -1, 1)$ is a saddle point.

(5) Find and classify the critical points of $w = f(x, y) := x^2 y^2$. At a critical point,

$$f_x = 2xy^2 = 0, \qquad f_y = 2x^2 y = 0.$$

Thus the points where either $x = 0$ or $y = 0$ are all critical points. Since

$$f_{xx} = 2y^2, f_{xy} = 4xy, \quad f_{yy} = 2x^2,$$

we find

$$H(X) = \begin{pmatrix} 2y^2 & 4xy \\ 4xy & 2x^2 \end{pmatrix}$$

If either $x = 0$ or $y = 0$, then $\det H = 0$ so none of our tests apply to determine the nature of the critical point. However, a glance at the function $f(x, y) = x^2 y^2$ reveals that all of the points where either $x = 0$ or $y = 0$ are clearly local minima, since $f = 0$ there, while $f > 0$ elsewhere.

Exercises

(1) Find and classify the critical points of the following functions.

(a) $f(x, y) = x^2 - 3x + 2y^2 + 10$
(b) $f(x, y) = 3 - 2x + 2y + x^2 y^2$
(c) $f(x, y) = [x^2 + (y + 1)^2][4 - x^2 - (y - 1)^2]$
(d) $f(x, y) = x^3 - 3xy^2$ (figure on next page)
(e) $f(x, y) = xy - x + y + 2$
(f) $f(x, y) = x \cos y$
(g) $f(x, y, z) = 2x^2 + 3xz + 5z^2 + 4y - y^2 + 7$
(h) $f(x, y, z) = 5x^2 + 4xy + 2y^2 + z^2 - 4z + 31$

(2) Let $X_1, \ldots, X_N$ be N distinct points in $\mathbb{R}^n$. Find a point $X \in \mathbb{R}^n$ such that the function
$$f(X) = \|X - X_1\|^2 + \cdots + \|X - X_N\|^2$$
is a minimum. [Answer: $X = \frac{1}{N}\sum_{j=1}^N X_j$, the center of gravity.]

(3) (a) Find the minimum distance from the origin in $\mathbb{R}^3$ to the plane $2x + y - z = 5$.

(b) Find the minimum distance from the origin in $\mathbb{R}^n$ to the hyperplane $a_1x_1 + a_2x_2 + \cdots + a_nx_n = c$.

(c) Find the minimum distance between the fixed point $X_0 = (\tilde{x}_1, \ldots, \tilde{x}_n)$ and the hyperplane $a_1x_1 + \ldots + a_nx_n = c$.

(d) Find the minimum distance between the two parallel planes $a_1x_1 + \cdots + a_nx_n = c_1$ and $a_1x_1 + \cdots + a_nx_n = c_2$,

(4) If $f(x, y)$ has two continuous derivatives, use Taylor's Theorem (Theorem 6) to prove
$$f(x + h_1, y + h_2) = f(x, y) + f_x(x, y)h_1 + f_y(x, y)h_2$$
$$+\frac{1}{2}[f_{xx}(x, y)h_1^2 + 2f_{xy}(x, y)h_1h_2 + f_{yy}(x, y)h_2^2] + (h_1^2 + h_2^2)R,$$
where R depends on x, y, h_1 and h_2, and $\lim_{\substack{h_1\to 0\\ h_2\to 0}} R = 0$.

(5) (a) If $u(x, y)$ has two continuous derivatives, use the result of Exercise 4 to prove
$$u_{xx}(x, y) = \frac{u(x + h_1, y) - 2u(x, y) + u(x - h_1, y)}{h_1^2} + h_1\tilde{R}$$
and
$$u_{yy}(x, y) = \frac{u(x, y + h_2) - 2u(x, y) + u(x, y - h_2)}{h_2^2} + h_2\hat{R},$$
where $\lim_{h_1\to 0} \tilde{R} = 0$ and $\lim_{h_2\to 0} \hat{R} = 0$.

(b) Use part a) to deduce that if $h_1 = h_2 = h$ then
$$u_{xx}(x, y) + u_{yy}(x, y)$$
$$= \frac{4}{h^2}[u(x, y) - \frac{u(x + h, y) + u(x - h, y) + u(x, y + h) + u(x, y - h)}{4}] + h^2R \tag{8-2}$$
where $\lim_{h\to 0} R = 0$.

(c) Use part b) to deduce that if h is small, the solution of the partial differential equation $u_{xx}+u_{yy} = 0$, Laplace's equation, approximately satisfies the difference equation
$$u(x, y) = \frac{u(x + h, y) + u(x - h, y) + u(x, y + h) + u(x, y - h)}{4}$$

This difference equation states that the value of u at the center of a cross equals the arithmetic mean ("average") of its values at the four ends of the cross. One could use the difference equation to solve Laplace's equation numerically.

(d) Prove that any function which satisfies the above difference equation in some set cannot have a maxima or minima inside that set. [Do not differentiate! Reason directly from the difference equation. No computation is necessary.]

(6) If all the second partial derivatives of a function $f(X)$ vanish identically in some open connected set, prove that f is an affine function.

(7) (The Method of Least Squares). Let $Z_1, \ldots, Z_N$ be N distinct points in $\mathbb{R}^n$, and $w_1, \ldots, w_N$ a set of N numbers. We imagine the points $(Z_j, w_j) \in \mathbb{R}^{n+1}$ to be points on a surface M in $\mathbb{R}^{n+1}$. Find a hyperplane
$$w = \phi(X) = c + \xi_1x_1 + \cdots, +\xi_nx_n \equiv c + \langle\xi,\, X\rangle$$
which most closely approximates the surface M in the sense that the error $E(\xi)$
$$E(\xi) := \sum_{j=1}^N |\phi(Z_n) - w_j|^2 =$$
is minimized. Note that you are to find the coefficients $\xi_1, \ldots, \xi_n$ in the equation of the hyperplane.

(8) (a) Let $u(x, y)$ be a twice continuously differentiable function which satisfies the partial differential equation
$$Lu := u_{xx} + u_{yy} + au_x + bu_y - cu = 0$$
in some open set D, where the coefficients $a(x, y), b(x, y)$, and $c(x, y)$ are continuous functions. If $c > 0$ throughout D, prove that $u(x, y)$ cannot have a positive maximum or negative minimum anywhere in D.

(b) Extend the result of part a) to functions $u(x_1, \ldots, x_n)$ which satisfy
$$Lu := \sum_{i=1}^n \frac{\partial^2 u}{\partial x_1^2} + \sum_{j=1}^n a_j\frac{\partial u}{\partial x_j} - cu = 0,$$
in some open set D, where $c > 0$ throughout D.

(c) If $u(x, y)$ satisfies the equation of part a) and u vanishes on the boundary of D, $u \equiv 0$ on ∂D, prove that $u(x, y) \equiv 0$ throughout D.

(d) Assume $u(x, y)$ and $v(x, y)$ both satisfy the same equation $Lu = 0$, $Lv = 0$, where L is the operator of part a). If $u(x, y) \equiv v(x, y)$ on the whole boundary of D, prove that $u(x, y) \equiv v(x, y)$ throughout the interior of D.

(9) Let f be a twice continuously differentiable function throughout the open set A. Prove that

(a) if f has a local minimum at $X_0 \in A$, then its Hessian $H(X_0)$ is positive definite or semi-definite there.

(b) if f has a local maximum at $X_0 \in A$, then its Hessian $H(X_0)$ is negative definite or semi-definite there.

(10) Let A be a square $n \times n$ self-adjoint matrix and Y a fixed vector in $\mathbb{R}^n$, and let

$$f(X) = \langle X,\, AX \rangle - 2\langle X,\, Y \rangle.$$

(a) If $f(X)$ has a critical point at X_0, prove X_0 satisfies the equation

$$AX_0 = Y.$$

(b) If A is positive definite and X_0 satisfies the equation $AX_0 = Y$, prove $f(X)$ defined above has a minimum at X_0. [The results of this problem remain valid if A is any positive definite linear operator - possibly a differential operator. The nonlinear function $f(X)$ defines a *variational problem* associated with the equation $AX = Y$.]

(11) If $f\colon A \to \mathbb{R}$ has three continuous derivatives in the open set $A \subset \mathbb{R}^2$ containing the origin, state precisely and prove Taylor's Theorem with three terms about the origin. The resulting expression will be

$$f(X) := f(x,y) = f(0) + f_x(0)x + f_y(0)y + \frac{1}{2!}[f_{xx}(0)x^2 + 2f_{xy}(0)xy + f_{yy}(0)y^2]$$

$$+\frac{1}{3!}[f_{xxx}(Z)x^3 + 3f_{xxy}(Z)x^2y + 3f_{xyy}(Z)xy^2 + f_{yyy}(Z)y^3]$$

where Z is on the line segment between 0 and $X = (x,y)$.

(12) If $u(x,y)$ has the property $u_{xy}(x,y) = 0$ for (x,y) in some open set, prove $u(x,y) = \phi(x) + \psi(y)$, where ϕ and ψ are functions of one variable.

(13) Compute the direction(s) at X_0 in which the following functions f

i) increase most rapidly,

ii) decrease most rapidly,

iii) remain constant.

(a) $f(x_1,x_2) = 3 - 2x_1 + 5x_2 \quad$ at $\quad X_0 = (2,1)$

(b) $f(x,y) = e^{2x+y} \quad$ at $\quad X_0 = (1,-2)$

(c) $f(x,y,z) = 2x^2 + 3xy + 5z^2 + 4y - y^2 + 7 \quad$ at $\quad X_0(1,0,-1)$

(d) $f(u,v) = uv - u + v + 2 \quad$ at $\quad (-1,1)$.

8.3 The Vibrating String.

Waves. You have been hearing about them your whole life. Waves are the term used to describe the oscillatory behavior of continuous media; water waves and sound waves being the most familiar. We shall give a mathematical description of a very simple type of wave - those in an oscillating violin string. The resulting mathematical model will be a second order linear partial differential equation - the wave equation - with both initial and boundary conditions.

a) The Mathematical Model

Consider a string of length ℓ stretched along the x axis. Imagine the string vibrating in the plane of the paper and let $u(x,t)$ denote the vertical displacement of the point x at time t. In order to end up with a tractable mathematical model several reasonable simplifying assumptions will be made. We assume the tension τ and density ρ of the string are constant throughout the motion, while the string is taken to be perfectly flexible so the tension force in the string acts along the tangential direction. Dissipative effects (air resistance, heating, etc.) are entirely neglected. One more assumption will be made when needed. It essentially states that the oscillations are small in some sense.

Newton's second law, $ma = \sum F$, is where we begin. Draw your attention to a small segment of the string whose length, at rest, is $\Delta x = x_2 - x_1$. The mass of the segment is $\rho\Delta x$. By Newton's second law the segment moves in such a way that the product of its center of gravity equals the resultant of the forces acting on it. For the vertical component, this means

$$\rho\Delta x\frac{\partial^2 u}{\partial t^2}(\tilde{x},t) = F_v,$$

where $\tilde{x} \in (x, x+\Delta x)$ is the horizontal coordinate of the center of gravity of the segment, and F_v means the vertical component of the resultant force.

There are two types of forces. One is the tension acting at both ends of the segment. The other is gravity acting down with a force equal to the weight of the segment, $\rho g\Delta x$. To evaluate the tension forces, let θ_1 and θ_2 be the angles the string makes with the horizontal at either end of the segment (see figure above). Then the vertical component of the tension force is

$$\tau\sin\theta_2 - \tau\sin\theta_1.$$

The signs indicate one force is up while the other is down. Adding the tension force to the gravitational force and substituting into Newton's second law, we find

$$\rho\Delta x\frac{\partial^2 u}{\partial t^2}(\tilde{x},t) = \tau(\sin\theta_2 - \sin\theta_1) - \rho g\Delta x.$$

The dependence of θ_1 and θ_2 on the displacement can be brought out by using the relation

$$\sin\theta = \frac{u_x}{\sqrt{1+u_x^2}},$$

which follows from the relation $u_x = \tan\theta$ for the slope of the string. Using this, we obtain the equation

$$\rho\Delta x \frac{\partial^2 u}{\partial t^2}(\tilde{x},t) = \tau\left[\left.\frac{u_x}{\sqrt{1+u_x^2}}\right|_{x=x_2} - \left.\frac{u_x}{\sqrt{1+u_x^2}}\right|_{x=x_1}\right] - \rho g \Delta x.$$

A simplifying assumption is badly needed. If the function $u_x/\sqrt{1+u_x^2}$ is expanded in a Taylor series,

$$\frac{u_x}{\sqrt{1+u_x^2}} = u_x - \frac{1}{2}u_x^3 + \cdots,$$

we see that if the slope u_x is small, essentially only the linear term in this series counts. Therefore, we do assume the slope u_x is small (this is the same assumption made in treating the simple pendulum). With this simplification, the equation of motion is

$$\rho\Delta x \frac{\partial^2 u}{\partial t^2}(\tilde{x},t) = \tau[u_x(x_2,t) - u_x(x_1,t)] - \rho g \Delta x.$$

Divide both sides of this equation by $\Delta x = x_2 - x_1$ and let the length of the interval shrink to zero. Since

$$\lim_{(x_2-x_1)\to 0} \frac{u_x(x_2,t)-u_x(x_1,t)}{x_2-x_1} = \frac{\partial}{\partial x}u_x(x,t) = \frac{\partial^2 u}{\partial x^2}(x,t),$$

where x is the limiting value of x_1 and x_2, we find

$$\rho\frac{\partial^2 u}{\partial t^2}(x,t) = \tau\frac{\partial^2 u}{\partial x^2}(x,t) - \rho g$$

Because the length of the interval has been shrunk to one point x, the center of gravity is now at x too.

It is customary to let $\tau/\rho = c^2$. The constant c has units of velocity, and, in fact, is just the speed with which waves travel along the string. Thus

$$Lu := u_{tt} - c^2 u_{xx} = -g.$$

This is the *wave equation*, a second order linear inhomogeneous partial differential equation. As was the case with linear ordinary differential equations, it is easier to attempt first to solve the homogeneous equation

$$Lu := u_{tt} - c^2 u_{xx} = 0.$$

On physical grounds, we expect the motion $u(x,t)$ of the string will be determined if the initial position $u(x,0)$ and initial velocity $u_t(x,0)$ are known, along with the motion of

both end points $u(0,t)$ and $u(\ell,t)$. However the mathematical model must be examined to see if these four facts do determine the subsequent motion (which it should if the model is to be of any use). Thus we must prove that given the

initial position $u(x,0) = f(x), \quad x \in [0,\ell]$
initial velocity $u_t(x,0) = g(x), \quad x \in [0,\ell]$
motion of left end $u(0,t) = \phi(t) \quad t \geq 0$
motion of right end $u(\ell,t) = \psi(t), \quad t \geq 0,$

then a solution $u(x,t)$ of the wave equation

$$u_{tt} - c^2 u_{xx} = 0$$

does exist which has these properties, and there is only one such solution. Existence and uniqueness theorems must therefore be proved.

b) Uniqueness

.

This is almost identical to all uniqueness theorems encountered earlier, especially that for the simple harmonic oscillator in Chapter 4, Section 2.

Theorem 8.11 *(Uniqueness). There exists at most one twice continuously differentiable function $u(x,t)$ which satisfies the inhomogeneous wave equation*

$$Lu := u_{tt} - c^2 u_{xx} = F(x,t)$$

and the subsidiary

initial conditions: $u(x,0) = f(x),\ u_t(x,0) = g(x),\ x \in [0,\ell]$
boundary conditions: $u(0,t) = \phi(t),\ u(\ell,t) = \psi(t),\ t \geq 0,$

where F, f, g, ϕ, and ψ are given functions.

Proof: Assume $u(x,t)$ and $v(x,t)$ both satisfy the same equation and the same subsidiary conditions. Let $w(x,t) = u(x,t) - v(x,t)$. Then $Lw = Lu - Lv = F - F = 0$, so w satisfies the homogeneous equation

$$Lw := w_{tt} - c^2 w_{xx} = 0$$

and has zero subsidiary data

initial conditions: $w(x,0) \equiv 0,\ w_t(x,0) \equiv 0,\ x \in [0,\ell]$
boundary conditions: $w(0,t) \equiv 0,\ w(\ell,t) \equiv 0,\ t \geq 0$

We want to prove $w(x,t) \equiv 0$. Notice that w satisfies the equation for a vibrating string which is initially at rest on the x axis, and whose ends never move. Therefore our desire to prove the string never moves, $w(x,t) \equiv 0$, is certain physically reasonable.

For this function w, *define* the new function $E(t)$

$$E(t) = \frac{1}{2}\int_0^\ell [w_t^2 + c^2 w_x^2]\, dx.$$

We have named the function $E(t)$ since it actually happens to be the *energy* in the string associated with the motion $w(x,t)$ at time t, except for a factor of ρ. Assume it is "legal" to differentiate under the integral sign (it is). Upon doing so, we get

$$\frac{dE}{dt} = \int_0^\ell [w_t w_{tt} + c^2 w_x w_{xt}]\, dx.$$

But an integration by parts reveals that

$$\int_0^\ell w_x w_{xt}\, dx = w_x w_t\Big|_0^\ell - \int_0^\ell w_t w_{xx}\, dx.$$

Because the end points are held fixed, $w(0,t) = 0$ and $w(\ell,t) = 0$, the velocity at those points is zero too, $w_t(0,t) = 0$ and $w_t(\ell,t) = 0$. This drops out the boundary terms in the integration by parts. Substituting the last expression into that for dE/dt, we find that

$$\frac{dE}{dt} = \int_0^\ell w_t[w_{tt} - c^2 w_{xx}]\, dx.$$

But w satisfies the homogeneous wave equation $w_{tt} - c^2 w_{xx} = 0$. Therefore $dE/dt \equiv 0$, so

$$E(t) \equiv \text{constant} = E(0),$$

that is, *energy is conserved*. Now

$$E(0) = \frac{1}{2}\int_0^\ell [w_t^2(x,0) + c^2 w_x^2(x,0)]\, dx.$$

Since the initial position is zero, $w(x,0) = 0$, its slope is also zero, $w_x(x,0) = 0$. The initial velocity $w_t(x,0)$ is also zero, $w_t(x,0) = 0$. Thus

$$E(t) \equiv E(0) \equiv 0,$$

that is,

$$0 = E(t) = \frac{1}{2}\int_0^\ell [w_t^2(x,t) + c^2 w_x^2(x,t)]\, dx.$$

Because the integrand is positive, we conclude $w_t(x,t) \equiv 0$ and $w_x(x,t) \equiv 0$. Consequently $w(x,t) \equiv$ constant. Since $w(0,t) = 0$, that constant is the zero constant,

$$w(x,t) \equiv 0.$$

Therefore

$$u(x,t) - v(x,t) \equiv w(x,t) \equiv 0,$$

so $u(x,t) \equiv v(x,t)$: the solution is unique.

c) Existence

For the simple one (space) dimension wave equation, there are many ways to prove a solution exists. The one to be given here is not the simplest (see Exercise 6 for the result of that method), but it does generalize immediately to many other problems. It makes no difference how we find a solution, for once found, by the uniqueness theorem it is the only possible solution. To avoid complications, we shall consider only the homogeneous equation and assume the end points are tied down. Thus, we want to solve

Wave equations: $u_{tt} - c^2 u_{xx} = 0$.
Initial conditions: $u(x,0) = f(x)$, $u_t(x,0) = g(x)$.
Boundary conditions: $u(0,t) = 0$, $u(\ell,t) = 0$.

The idea is first to find special solutions $u_1(x,t)$, $u_2(x,t), \ldots$, which satisfy the boundary conditions but do not necessarily satisfy the initial conditions. Then, as was done for linear O.D.E.'s, we build the solution which does satisfy the given initial conditions as a linear combination of these special solutions,

$$u(x,t) = \sum A_j u_j(x,t),$$

that is, by superposition.

Let us seek special solutions in the form of a *standing wave*,

$$u(x,t) = X(x)T(t).$$

Here $X(x)$ and $T(t)$ are functions of one variable. Our procedure is reasonably called *separation of variables*. Substitution of this into the wave equation gives

$$\ddot{T}(t)X(x) - c^2 X''(x)T(t) = 0,$$

or

$$\frac{X''(x)}{X(x)} = \frac{1}{c^2}\frac{\ddot{T}(t)}{T(t)}.$$

Since the left side depends only on x, while the right depends only on t, both sides must be constant (a somewhat tricky remark; think it over). Let that constant be $-\gamma$ (using $-\gamma$ instead of γ is the result of hindsight, as you shall see).

$$\frac{X''}{X} = \frac{1}{c^2}\frac{\ddot{T}}{T} = -\gamma.$$

This leads us to the two ordinary differential equations

$$X''(x) + \gamma X(x) = 0, \qquad \ddot{T}(t) + \gamma c^2 T(t) = 0.$$

Since $u(0,t) = 0$ and $u(\ell,t) = 0$ and $u(x,t) = X(x)T(t)$, the function $X(x)$ must also satisfy the boundary conditions

$$X(0) = 0, \qquad X(\ell) = 0.$$

There are several ways to show γ must be positive. Perhaps the simplest is to observe that if $\gamma < 0$ or $\gamma = 0$, the only function $X(t)$ which satisfies the differential equation $X'' + \gamma X = 0$ and boundary conditions $X(0) = X(\ell) = 0$ is the zero function $X(x) \equiv 0$. Since for this function $u(x,t) = X(x)T(t) \equiv 0$, it is devoid of further interest.

Another way to show γ is positive is to multiply the ordinary differential equation $X'' + \gamma X = 0$ by $X(x)$ and integrate over the length of the string,

$$\int_0^\ell [X(x)X''(x) + \gamma X^2(x)]\,dx = 0.$$

Upon integrating by parts, we find that

$$\int_0^\ell ?(x)X'(x)\,dx = XX'\Big|_0^\ell - \int_0^\ell X'^2(x)\,dx.$$

Since $X(0) = X(\ell) = 0$, the boundary terms drop out. Substituting this into the above equation, we find that

$$\int_0^\ell X'^2(x)\,dx = \gamma \int_0^\ell X^2(x)\,dx.$$

If $X(x)$ is not identically zero, this can be solved for γ

$$\gamma = \frac{\displaystyle\int_0^\ell X'^2(x)\,dx}{\displaystyle\int_0^\ell X^2(x)\,dx},$$

and clearly shows $\gamma > 0$.

Enough for that. The solution of $X'' + \gamma X = 0$, $\gamma > 0$, is

$$X(x) = A\cos\sqrt{\gamma}x + B\sin\sqrt{\gamma}x.$$

The boundary condition $X(0) = 0$ implies $A = 0$, while the boundary condition at the other end point $X(\ell) = 0$, implies

$$0 = B\sin\sqrt{\gamma}\ell.$$

If $B = 0$ too, then $X(x) \equiv 0$, so $u(x,t) \equiv 0$. This is of no use to us. The only alternative is to restrict γ so that $\sin\sqrt{\gamma}l = 0$. This means $\sqrt{\gamma}\ell$ is a multiple of π, $\sqrt{\gamma}\ell = n\pi$, $n = 1, 2, \ldots$,

$$\sqrt{\gamma} = \frac{n\pi}{\ell}, \quad n = 1, 2, \ldots.$$

There is then one possible solution $X(x)$ for each integer n,

$$X_n(x) = B_n \sin\frac{n\pi}{\ell}x,$$

where the constants B_n are arbitrary.

Remark: There is a similarity of deep significance for mathematics and physics between the work in these last few paragraphs and that done for the coupled oscillators in Chapter 6. There (p. 528-9), we had an operator A and wanted to find nonzero vectors S_n and numbers λ such that

$$AS_n = \lambda_n S_n.$$

The numbers found λ_n were called the eigenvalues of A, and S_n the corresponding eigenvectors.

Here, we were given the operator $A = -\frac{d^2}{dx^2}$ and wanted to find nonzero functions $X_n(t) \in \{\, X \in C^2[0,\ell]\colon X(0) = X(\ell) = 0 \,\}$ which satisfy the equation

$$AX_n = \gamma_n X_n$$

The numbers found, $\gamma_n = n^2\pi^2/\ell^2$, are also called the *eigenvalues* of A, and the function $X_n(t) = \sin\frac{n\pi}{\ell}x$, the *eigenfunction* of A corresponding to the eigenvalue γ_n.

Associated with each possible eigenvalue γ_n, there is a solution of the time equation, $\ddot{T} + \gamma c^2 T = 0$,

$$T_n(t) = C_n\cos\frac{nc\pi}{\ell}t + D_n\sin\frac{nc\pi}{\ell}t.$$

We therefore have found one special solution, $u_n(x,t) - X_n(t)T_n(t)$, for each value of the index n,

$$u_n(x,t) = \sin\frac{n\pi x}{\ell}(\alpha_n\cos\frac{nc\pi t}{\ell} + \beta_n\sin\frac{nc\pi t}{\ell}).$$

The arbitrary constants have been lumped in this equation. These special solutions are the "natural" vibrations of the string, or *normal modes of vibration.* A snapshot at $t = t_0$ of the string moving in the nth normal mode would reveal the sine curve

$$u_n(x,t_0) = C\sin\frac{n\pi x}{\ell},$$

the constant C accounting for the remaining terms, which are constant for t fixed. In music, the integer n refers to the octave. The fundamental tone is the case $n = 1$, while the tone for $n = 2$, the *second harmonic* or *first overtone*, is one octave higher.

A FIGURE GOES HERE

The *time frequency* $\mathcal{V}_n$ of the nth normal mode is $\mathcal{V}_n = \frac{nc\pi}{\ell}$, this is the number of oscillations in 2π units of time. It is the time frequency which we usually associate with musical *pitch.* The (time) *period* τ_n of the nth normal mode is $2\pi/\mathcal{V}_n$, that is $\tau_n = 2l/nc$. Another name you will want to know is the *wave length* λ_n of the nth normal mode, $\lambda_n = 2\ell/n$ (see figures above). Notice that $\mathcal{V}_n\lambda_n = c$, an important relationship.

Having found the special normal mode solutions, $u_n(x,t)$, we hope that arbitrary constants α_n and β_n can be chosen so a linear combination

$$u(x,t) = \sum_{n=1}^\infty u_n(x,t) = \sum_{n=1}^\infty (\alpha_n\cos\frac{nc\pi t}{\ell} + \beta_n\sin\frac{nc\pi t}{\ell})\sin\frac{n\pi x}{\ell}$$

will satisfy the given initial conditions. Every function $u(x,t)$ of this form automatically satisfies the boundary conditions $u(0,t)=0$, $u(\ell,t)=0$ since each of the u_n's satisfy them.

If $u(x,0)=f(x)$ and $u_t(x,0)=g(x)$, then from the above equation, we must have

$$f(x)=\sum_{n=1}^{\infty}u_n(x,0)=\sum_{n=1}^{\infty}\alpha_n\sin\frac{n\pi x}{\ell}$$

and

$$g(x)=\sum_{n=1}^{\infty}\frac{\partial u_n}{\partial t}(x,0)=\sum_{n=1}^{\infty}\frac{n\pi c}{\ell}\beta_n\sin\frac{n\pi x}{\ell}.$$

Thus, the coefficients α_n are the coefficients in the Fourier sine series for f, while the β_n are essentially the coefficients in the Fourier sine series for g. In fact, this is how Fourier was led to the series bearing his name. These formulas for $u(x,y)$, $f(x)$, and $g(x)$ become easier on the eye if the length of the string is $\pi, \ell=\pi$. Then

$$u(x,y)=\sum_{n=1}^{\infty}(\alpha_n\cos nct+\beta_n\sin nct)\sin nx,$$

while

$$f(x)=\sum_{n=1}^{\infty}u_n(x,0)=\sum_{n=1}^{\infty}\alpha_n\sin nx, \tag{8-3}$$

and

$$g(x)=\sum_{n=1}^{\infty}\frac{\partial^u n}{\partial t}(x,0)=\sum_{n=1}^{\infty}nc\beta_n\sin nx.$$

Finding the coefficients α_n and β_n is particularly simple if f and g can be represented by finite series.

Examples: Find the solution $u(x,t)$ of the wave equation for a string of length π, $l=\pi$, which is pinned down at its end points, $u(0,t)=u(\pi,t)=0$, and satisfies the given initial conditions.

(1) $u(x,0)=f(x)=2\sin 3x$, $u_t(x,0)=g(x)=\frac{1}{2}\sin 4x$. We have to find α_n and β_n for the two series

$$2\sin 3x=\sum_{n=1}^{\infty}\alpha_n\sin nx$$

$$\frac{1}{2}\sin 4x=\sum_{n=1}^{\infty}nc\beta_n\sin nx.$$

For these simple functions, just match coefficients, giving

$$\alpha_3=2,\ \alpha_n=0,\ n\neq 3,\ \text{and}\ \beta_4=\frac{1}{8c},\ \beta_n=0,\ n\neq 4.$$

Therefore, the sum of the two waves

$$u(x,t)=2\cos 3ct\sin 3x+\frac{1}{8c}\sin 4ct\sin 4x$$

is the (unique!) solution of this example.

(2) $u(x,0)=f(x)=\frac{1}{2}\sin 3x-\sin 17x$ and

$$u_t(x,0)=g(x)=-9\sin x+13\sin 973x.$$

We have to find α_n and β_n for the two series

$$\frac{1}{2}\sin 3x-\sin 17x=\sum_{n=1}^{\infty}\alpha_n\sin nx$$

and

$$-9\sin x+13\sin 973x=\sum_{n=1}^{\infty}nc\beta_n\sin nx.$$

By matching again, we find $\alpha_3=\frac{1}{2}$, $\alpha_{17}=-1$, and $\alpha_n=0$ for $n\neq 3$ or 17. Also, $\beta_1=\frac{9}{c}, \beta_{973}=\frac{13}{973c}$, and $\beta_n=0$ for $n\neq 1$ or 973. The (unique) solution is then a sum of four waves

$$u(x,t)=-\frac{9}{3}\sin ct\sin x+\frac{1}{2}\cos 3ct\sin 3x$$
$$-\cos 17ct\sin 17x+\frac{13}{973c}\sin 973ct\sin 973x.$$

Since f and g are not usually given in the simple form of these examples, the full Fourier series is needed. Recall that the string is pinned down at both ends. Therefore both the initial position function $f(x)$ and velocity function $g(x)$ have the property $f(0)=f(\pi)=0$, and $g(0)=g(\pi)=0$, where we have taken the length of the string to be π. It is now possible to extend both f and g, assumed continuous in $[0,\pi]$, to the whole interval $[-\pi,\pi]$ as continuous odd functions,

A FIGURE GOES HERE

that is, if $x\in[0,\pi]$, we can define

$$f(-x)=-f(x)\quad\text{and}\quad g(-x)=-g(x),$$

since the right sides, $-f(x)$ and $-g(x)$, are known functions for $x\in[0,\pi]$.

As odd functions now on the whole interval $[-\pi,\pi]$, the functions f and g have Fourier sine series (cf. p. 252, Exercise 3a).

$$f(x) = \sum_{n=1}^{\infty} b_n \frac{\sin nx}{\sqrt{\pi}}$$

$$g(x) = \sum_{n=1}^{\infty} \tilde{b}_n \frac{\sin nx}{\sqrt{\pi}}$$

where

$$b_n = 2\int_0^{\pi} f(x)\frac{\sin nx}{\sqrt{\pi}}\,dx, \quad \tilde{b}_n = 2\int_0^{\pi} g(x)\frac{\sin nx}{\sqrt{\pi}}\,dx \tag{8-4}$$

Comparing with the previous formulas (3) for f and g, we find

$$\alpha_n = b_n/\sqrt{\pi}, \quad \text{and} \quad \beta_n = \tilde{b}_n/nc\sqrt{\pi}$$

Consequently

$$u(x,t) = \sum_{n=1}^{\infty}\left(b_n \frac{\cos nct}{\sqrt{\pi}} + \frac{\tilde{b}_n}{nc}\frac{\sin nct}{\sqrt{\pi}}\right)\sin nx \tag{8-5}$$

the coefficients b_n and $\tilde{b}_n$ being determined from the initial conditions by equation (4).

Thus, we have almost proved

Theorem 8.12 . *If $f(x)$ is twice continuously differentiable and $g(x)$ once continuously differentiable for $x \in [0,\pi]$ and both functions vanish at $x = 0$ and $x = \pi$, then the function $u(x,t)$ defined by equation (5) is a solution of the homogeneous wave equation*

$$u_{tt} - c^2 u_{xx} = 0$$

and satisfies the

initial conditions: $u(x,0) = f(x)$, $u_t(x,0) = g(x)$, $x \in [0,\pi]$,

as well as the

boundary conditions: $u(0,t) = 0$, $u(\pi,t) = 0$, $t \geq 0$,

where b_n and $\tilde{b}_n$ are determined from f and g through equations (4). Moreover, this solution is unique (by Theorem 9).

Outline of Proof. If it is possible to differentiate the infinite series (5) term by term $u(x,t)$ would satisfy the wave equation since each special solution $u_n(x,t)$ does. In any case, the initial condition $u(x,0) = f(x)$ is clearly satisfied. However, checking the other initial condition $u_t(x,0) = g(x)$ also involves differentiating the infinite series term by term.

Thus, we must only justify the term by term differentiation of an infinite Fourier series. For power series, we found (p. 82-3, Theorem 16) we can always differentiate term by term within its disc of convergence. Such is not the case with Fourier series. For example,

the Fourier series $\displaystyle\sum_{n=1}^{\infty} \frac{\sin n^2 x}{n^2}$ converges for all x, but the series obtain by differentiating formally, $\displaystyle\sum_{n=1}^{\infty} \cos n^2 x$ diverges at $x = 0$. However, if a function is sufficiently smooth, its Fourier series can be differentiated term by term and does converge to the derivative of the function. Since the details of a complete proof are but a rehash of the proof carried out for power series (p. 82ff), we omit it.

EXAMPLE: Find the displacement $u(x,t)$ of a violin string of length π with fixed end points which is plucked at its midpoint to height h. The initial position is then

$$f(x) = \begin{cases} xh, & x \in [0,\pi/2] \\ (\pi - x)h, & x \in [\pi/2,\pi], \end{cases}$$

and the initial velocity, $g(x)$, is zero.

We must find the coefficients b_n and $\tilde{b}_n$ in the series (5). After mentally continuing f and g to the interval $[-\pi,\pi]$ as odd functions, the formulas (4) give us b_n and $\tilde{b}_n$,

$$b_n = 2\int_0^{\pi} f(x)\frac{\sin nx}{\sqrt{\pi}}\,dx = \frac{2h}{\sqrt{\pi}}\left\{\int_0^{\pi/2} x\sin nx\,dx + \int_{\pi/2}^{\pi}(\pi - x)\sin nx\,dx\right\}.$$

Integrating and simplifying, we find that

$$b_n = \frac{4h}{\sqrt{\pi}n^2}\sin\frac{n\pi}{2} = \begin{cases} 0, & n \text{ even} \\ 1, & n = 1,5,9,13,\ldots \\ -1, & n = 3,7,11,15. \end{cases}$$

From $g(x) \equiv 0$, it is immediate that $\beta_n = 0$ for all n. Thus,

$$u(x,t) = \frac{4h}{\pi}\sum_{n=1}^{\infty}\frac{1}{n^2}\sin\frac{n\pi}{2}\cos n\,ct\,\sin nx$$

$$= \frac{4h}{\pi}\left[\frac{\cos 3ct\,\sin x}{1} - \frac{\cos ct\,\sin 3x}{3^2} + \frac{\cos 5ct\,\sin 5x}{5^2} + \cdots\right]$$

is the desired solution.

Exercises

(1) (a) Find a solution $u(x,t)$ of the homogeneous wave equation for a string of length π whose end points are held fixed if the initial position function is

$$u(x,0) = \frac{1}{2}\sin 4x - \sin 7x,$$

while the initial velocity is

$$u_t(x,0) = \sin 3x + \sin 73x.$$

(b) Same problem as a), but

$$u(x,0) = \sin 5x + 12\sin 6x - 7\sin 9x$$

$$u_t(x,0) = -\sin x + 91\sin 273x.$$

(2) Find a solution $u(x,t)$ of the homogeneous wave equation for a string of length π whose end points are held fixed if the string is initially plucked at the point $x = \pi/4$ to the height h.

(3) Consider a vibrating string of length ℓ whose end points are on rings which can slide freely on poles at 0 and ℓ. Then the boundary conditions at the end points are

$$u_x(0,t) = 0,\ u_x(\ell,t) = 0$$

that is, zero slope.

(a) Use the method of separation of variables to find the form of special standing wave solutions. [Answer: $u_n(x,t) = \cos\frac{n\pi x}{\ell}(\alpha_n \cos\frac{nc\pi t}{\ell} + \beta_n \sin\frac{nc\pi t}{\ell})$].

(b) Use these to find a solution with the initial conditions

$$u(x,0) = \cos x - 6\cos 3x \qquad (\text{let } \ell = \pi)$$

$$u_t(x,0) = \frac{1}{2}\cos 2x.$$

(4) Let $u(x,t)$ satisfy the homogeneous wave equation. Instead of keeping the end points fixed, we either put them on rings (cf. Exercise 3) or attach them by elastic bands, in which case the boundary conditions become

$$u_x(0,t) - c_1 u(0,t) = 0,\ u_x(\pi,t) + c_2 u(\pi,t) = 0,\ c_1, c_2 \ge 0.$$

(a) Define the energy as before, and prove that energy is dissipated with these boundary conditions, unless c_1 and c_2 vanish.

(b) Prove there is at most one function $u(x,t)$ which satisfies the inhomogeneous wave equation $u_{tt} - c^2 u_{xx} = F(x,t)$ with initial conditions as before, but with elastic boundary conditions

$$u_x(0,t) - c_1 u(0,t) = \phi(t),\ u_x(\pi,t) + c_2 u(\pi,t) = \psi(t),$$

where c_1 and c_2 are non-negative constants.

(5) To account for the effect of air resistance on a vibrating string, one common assumption is that the resistance on a segment of length Δx is proportional to the velocity of its center of gravity,

$$F_{\text{res}} = -k\Delta x u_t(\tilde{x},t),\ k > 0,$$

where k is a numerical constant. This is analogous to the standard viscous resistance force on a harmonic oscillator.

(a) Find the equation of motion ignoring gravity. [Answer: $\frac{1}{c^2}u_{tt} + ku_t = u_{xx}$]

(b) Find the form of the special standing wave solutions, assuming, the end points are held fixed.

(c) Write a formula giving the probable form for the general solution $u(x,t)$.

(d) If the end points are pinned down, what do you expect the behavior of the string will be as $t \to \infty$? Does the formula found in part c) verify your belief (it should).

(e) Define the energy $E(t)$ as before and show that energy is dissipated if the ends are held fixed.

(f) Use the result of e) to prove $\dot{E}(t) + 2kE(t) \ge 0$, and conclude that $E(t) \ge E(0)e^{-2kt}$ for $t \ge 0$. This shows that the energy is not dissipated too rapidly.

(6) It is possible to write the solution of the homogeneous wave equation for a string of length π with fixed end points in a simple closed form by using the trigonometric identities

$$2\sin nx\cos nct = \sin n(x-ct) + \sin n(x+ct).$$

$$2\sin nx\sin nct = \sin n(x-ct) - \cos n(x+ct).$$

(a) Do this and obtain *d'Alembert's formula*

$$u(x,t) = \frac{f(x-ct) + f(x+ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} g(\xi)d\xi.$$

(b) Solve the example of a plucked string (p. 641) again using this formula. Draw two sketches, one indicating the position of the string at time $t = \frac{\pi}{2c}$ and another at $t = \frac{\pi}{c}$.

(7) (a) Prove the wave operator $L := \frac{\partial^2}{\partial t^2} - c^2\frac{\partial^2}{\partial x^2}$, c a constant, is translation invariant, that is, if $T\colon u(x,t) \to u(x+x_0,\ t+t_0)$, prove $(LT)u = (TL)u$ for all values of x_0 and t_0, and for all functions u for which the operators make sense.

(b) Find the function $\phi(a,b)$ in the formula

$$Le^{ax+bt} = \phi(a,b)e^{ax+bt}.$$

(c) Use part b) to show that if a is any constant, the four functions

$$e^{a(x+ct)}, e^{-a(x+ct)}, e^{a(x-ct)}, e^{-a(x-ct)}$$

are solutions of the homogeneous wave equation $Lu = 0$.

(d) Use the fact that each of the above functions satisfies the ordinary differential equation $v''(x) = a^2 v(x)$ to conclude that if linear combinations of these functions are to satisfy the boundary conditions $v(0) = v(\ell) = 0$, then necessarily $a^2 < 0$, so the constant a is pure imaginary and we can write $a = i\gamma$, where γ is real.

(e) Let $u(x,t)$ be a linear combination of the four functions part c) with $a = i\gamma$. Show that $u(x,t)$ may be written in the form

$$u(x,t) = \sin\gamma x \quad [A\cos\gamma ct + B\sin\gamma ct].$$

(f) If $u(0,t) = u(\ell,t) = 0$, show that $\gamma_n = \frac{n\pi}{\ell}$. Find an infinite set of special solutions $u_n(x,t)$ which satisfy the homogeneous wave equation with zero boundary values [From here on, one proceeds as before to find the general solution. This problem has shown how the idea of translation invariance can also be used to lead one to the special solutions u_n].

(8) (a) By *inspection*, find a particular solution for the solution of the inhomogeneous wave equations

$$Lu := u_{tt} - c^2 u_{xx} = g, \quad g \equiv \text{constant}.$$

(b) How can this particular solution be used to find the solution of the equation $Lu = g$ which has given initial conditions and zero boundary conditions?

(9) Flow of heat in a thin insulated rod on the x axis is governed by the *heat equation*

$$u_t(x,t) = k^2 u_{xx}(x,t),$$

where $u(x,t)$ represents the temperature at the point x at time t, and k^2, the *diffusivity*, is a constant depending on the material. The "energy" in a rod of length ℓ, $0 \le x \le \ell$, is defined as

$$E(t) = \frac{1}{2}\int_0^l u^2(x,t)\,dx.$$

(a) If the ends of the rod have zero temperature, $u(0,t) = u(\ell,t) = 0$, prove "energy" is dissipated, $\dot{E}(t) \le 0$, by showing

$$\frac{dE(t)}{dt} = -k^2\int_0^\ell u_x^2(x,t)\,dx.$$

(b) Given a rod whose ends have zero temperature and whose initial temperature is zero, $u(x,0) = 0$, prove that the temperature remains zero, $u(x,t) \equiv 0$.

(c) Prove the temperature of a rod is uniquely determined if the following three data are known:
initial temperature: $u(x,0) = f(x), \quad x \in [0,\ell]$.
boundary conditions: $u(0,t) = \phi(t), \quad u(\ell,t) = \psi(t)$, $t \ge 0$.

(d) Use the method of separation of variables to find an infinite number of special solutions of the heat equation for a thin rod whose end points have zero temperature for all $t \ge 0$. [Answer: $u_n(x,t) = c_n e^{\frac{-n^2k^2\pi^2}{\ell^2}t}\sin\frac{n\pi}{\ell}x,\ n = 1,2,\ldots$]

(e) If the ends of a rod have zero temperature for all $t \ge 0$, what do you intuitively expect the temperature $u(x,t)$ will be as $t \to \infty$? Is this borne out by the formulas for the special solutions?

(f) Find the temperature distribution in a rod of length π if the ends have zero temperature and if the initial temperature distribution in the rod is

$$u(x,0) = \sin x - 4\sin 7x,$$

(10) If the temperature at the ends of the bar of length ℓ is constant but not necessarily zero, say

$$u(0,t) = \theta_1, \qquad u(\ell,t) = \theta_2,$$

the temperature distribution can be found be splitting the solution into two parts, $u(x,t) = \tilde{u}(x,t) + u_p(x,t)$, where $u_p(x,t)$ is a particular solution having the correct temperature at the ends of the bar and $u(x,t)$ is a general solution which has zero temperature at the ends.

(a) Find a particular solution of the homogeneous heat equation $u_t = k^2 u_{xx}$ which satisfies $u(0,t) = 20^0$, $u(\ell,t) = 50^0$, but does not necessarily satisfy any prescribed initial condition. [Answer: Many possible solutions - for example $u_p(x,t) = 20 + 30\frac{x}{\ell}$, or $u_p(x,t) = 20 + 30\sin\frac{\pi x}{2\ell}$].

(b) Find the temperature distribution in a rod of length π if the initial temperature is $u(x,0) = 2\sin x - \sin 4x$, while the boundary conditions are as in part a).

(11) If the ends of a bar of length ℓ are insulated instead of being kept at zero, the boundary conditions are

$$u_x(0,t) = u_x(\ell,t) = 0.$$

(a) Use the method of separation of variables to find an infinite number of special solutions for the homogeneous heat equation with insulated ends. [Answer: $u_n(x,t) = c_n e^{\frac{-n^2k^2\pi^2}{\ell^2}t}\cos\frac{n\pi x}{\ell}$, $n = 0,1,2,\ldots$].

(b) What is the temperature distribution in a rod whose ends are insulated if the initial temperature distribution is

$$u(x,t) = 3\cos\frac{2\pi x}{\ell} - \frac{1}{5}\cos\frac{5\pi x}{\ell}.$$

(12) In this exercise you will find a quantitative estimate for the rate of decrease of energy for the heat in a rod of length ℓ with zero temperature at the ends.

(a) Use the result of Exercise 9a to prove the differential inequality

$$\frac{dE}{dt} \le -cE(t),$$

where c is a positive constant. [Hint: Look at p. 227 Exercise 15c].

(b) Conclude that

$$E(t) \le E(0)e^{-ct},\ t \ge 0.$$

This is the desired estimate for the decrease of energy in the rod.

(13) The linear partial differential equation

$$u_{xx} - u = u_t$$

governs the temperature distribution in a rod of length ℓ made up of a material which uses up heat to carry out a chemical process. Define the energy $E(t)$ in the rod as in Exercise 9.

(a) Prove that if the ends of the rod have zero temperature, then the energy is dissipated, $\dot{E}(t) \le 0$.

(b) Given a rod whose ends have zero temperature and whose initial temperature $u(x,0)$ is zero, use a) to prove that the temperature remains zero, $u(x,t) \equiv 0,\ t \ge 0$.

(c) Use part b) to prove that the temperature of the rod described above is uniquely determined if the following three data are known

$$u(x,0) \quad \text{for} \quad x \in [0,\ell],\ u(0,t) \quad \text{and} \quad u(\ell,t) \quad \text{for} \quad t \ge 0.$$

(14) In setting up the mathematical model for the vibrating string, we never examined the horizontal components of the forces.

(a) Show that the net horizontal force is

$$F_h = \tau\cos\theta_2 - \tau\cos\theta_1$$

(b) Under our assumption u_x is small, show that the net horizontal force is zero - so there is no horizontal motion of the string. This justifies the statement that the motion of the string is entirely vertical.

(15) Use the formula $\mathcal{V}_n = n\pi c/\ell$ (page 635) for the frequency and the relationship between c, T and ρ (page 624) to derive a formula for $\mathcal{V}_n$ in terms of the physical constants ℓ, T, and ρ for a vibrating string. Interpret the effect on the frequency, $\mathcal{V}_n$, if the physical constants are changed. Does this agree with your experience in tuning stringed instruments?

8.4 Multiple Integrals

How can we extend the notion of integration from functions of one variable to functions of several variables? That is the problem we shall face in this section.

Let $w = f(X) = f(x_1,\ldots,x_n)$ be a scalar-valued function defined in $C \subset \mathbb{R}^n$. For the purposes of this section it will be convenient to think of f as either the height function for a surface M in $\mathbb{R}^{n+1}$ over D, or as the mass density of D. In the first case. $\iint_D f$ should be the volume of the solid contained between M and D (see fig.), whereas in the second case, $\iint_D f$ should be total mass of the set D.

Two problems have to be solved. First, define the integral in $\mathbb{R}^n$. Second, give a reasonable procedure for explicitly evaluating the integral in sufficiently simple situations. More so than for the single integral, the problem of defining the multiple integral bristles with technical difficulties. However, after this is done the evaluation of integrals in $\mathbb{R}^n$ can be reduced to the evaluation of repeated integrals, that is, a sequence of n integrals in $\mathbb{R}^1$, which is in turn effected not by using the definition of the integral, but rather by recourse to the fundamental theorem of calculus.

Before starting the formalities, it is well advised to see where some difficulties lie. Suppose we are given a density function f defined on some domain D and want to find the total mass of D. To make things even simpler, assume for the moment that the density is constant and equal to 1, for all $X \in D \subset \mathbb{R}^n$. Then the mass coincides with the volume of the domain. For the special case of functions of one variable $D \subset \mathbb{R}^1$ is an interval so the "volume" of D (really the length of D) is trivial

A FIGURE GOES HERE

to compute, $\mathrm{Vol}(D) = b - a$. However if D has two or more dimensions, even finding the volume of D (area if $D \subset \mathbb{R}^2$) is itself difficult.

The problem is that a connected set D in $\mathbb{R}^1$ can only be a line segment, whereas a connected open set in $\mathbb{R}^n$, $n \ge Z$ can be much more complicated topologically. In $\mathbb{R}^1$, the closed "cube" and closed "ball" are both intervals $[a,b]$, and every other connected set is also an interval. In $\mathbb{R}^2$, not only do the cube and ball become distinct, but also a slew of other possibilities arise. D may be riddled with holes and its

A FIGURE GOES HERE

boundary wild (contrasted to the boundary of a connected set in $\mathbb{R}^1$ which is always just two points, the end points of the interval). It should be clear that the notion of volume of a set D may only be definable if the boundary of D is sufficiently smooth.

As you should be anticipating, the volume of a set D will be defined by filling it up with little cubes of volume $\Delta x_1 \Delta x_2 \ldots \Delta x_n = \Delta V$, and then proving that as the size of the cubes becomes small, the sum of volumes of the cubes approaches a limit (here is where the smoothness of θD enters). In two dimensions, $D \subset \mathbb{R}^2$, this roughly reads

$$\text{Area}\,(D) = \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \sum\sum \Delta x \Delta y = \int_D dx\,dy.$$

Only after the volume of a domain is defined can the more general notion of mass of a set D for a density function f be defined. The procedure here is straightforward, however it is important that the density f be "essentially" continuous. Using the same approximating cubes, we assign to each little cube its approximate density, say by using the value of the density f at the center of the little cube. Adding up the masses of these little cubes and passing to the limit again, we find the total mass of the solid D with density f. Again, in two dimensions this roughly reads

$$\text{Mass}\,(D) = \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \sum\sum f(x_i, y_j)\Delta x \Delta y = \iint_D f(x,y)\,dx\,dy.$$

Because of the technical complications, we shall only state a series of propositions which give the existence of the integral. The proofs of several crucial - but believable - results will not be carried out, but can be found in many advanced calculus books. For convenience, the geometric language of the plane, $\mathbb{R}^2$, will be used. The ideas extend immediately to higher dimensions. Now some terminology.

Definition: A *shaved rectangle* is a rectangle with its bottom and left sides omitted, that is, a set of the form

$$Q = \{X = (x_1, x_2) \colon a_j < x_j \le b_j,\ j = 1, 2\}.$$

A *rectangular complex* is a finite union of shaved rectangles, which can always be assumed disjoint, that is, non-overlapping. This should more accurately be called a shaved rectangular complex, but is not for the sake of euphony.

If D is a set, the *characteristic function of* D, X_D is defined by

$$X_D(X) = \begin{cases} 1, & X \in D \\ 0, & X \ni D. \end{cases}$$

A *step function* $s(X)$ is a finite linear combination of characteristic functions of shaved rectangles. The graph of this function looks like its name implies.

A FIGURE GOES HERE

A function f has *compact support* if it is identically zero outside some sufficiently large rectangle. The *support* of a particular function f, written supp f, is the smallest closed set outside of which f is zero. Thus, it is the set of all points X where $f(X) \neq 0$ and the limit points of those points.

We take the area of a shaved rectangle Q as a known quantity - the height times base, and *define* the integral as

$$I(X_Q) = \iint_{\mathbb{R}^2} X_Q\,dA = \int_D dA \equiv \text{Area}\,(Q),$$

where the Area (Q) is defined in the natural way as length $\times$ width. You may wish to think of dA as representing an "infinitesimal element of area". We however assign no meaning

to the symbol and use it only as a reminder. Some prefer to do without it altogether and write

$$\iint_{\mathbb{R}^2} X_Q = \text{Area}\,(Q).$$

Our task is to define

$$I(f) \equiv \iint_{\mathbb{R}^2} f\,dA$$

for density functions other than X_Q's. For example, if D is some set, for the function X_D we want to define

$$\text{Area}\,(D) = \iint_{\mathbb{R}^2} X_D\,dA = \iint_D dA$$

But this will not make sense unless it is shown that the set D does have a number associated with it which has the properties of area. It is easy to define the integral of a step function S. Let

$$S(X) = \sum_{j=1}^{n} a_j X_{Q_j}(X),$$

where the Q_j's are disjoint. Then $\iint S\,dA$ should represent the total mass of a plate composed of rectangles $Q_1, \ldots, Q_n$ with respective densities $a_1, \ldots, a_n$. Thus, we define

$$I(S) = \iint S\,dA \equiv a_1\text{Area}\,(Q_1) + \ldots + a_n\text{Area}', (Q_n) = \sum_{j=1}^{n} a_j \int X_{Q_j}\,dA.$$

The integrals of step functions clearly satisfy the following

Lemma 8.13 . *If $S_1(X)$ and $S_2(X)$ are step functions, then*
a). $I(aS_1 + bS_2) = aI(S_1) + bI(S_2)$.
b). $S_1(X) \le S_2(X)$ implies $I(S_1) \le I(S_2)$.
c). If $S(X)$ is bounded by M, $S(X) \le M$, then

$$I(S) \le cM,$$

where c is the area of the support of S.

The integral of any other more complicated function is defined by using step functions.

Definition: A function $f : \mathbb{R}^2 \to \mathbb{R}$ is *Riemann integrable* if given any $\epsilon > 0$, there are step functions s and S with $s(X) \le f(X) \le S(X)$ for all $X \in \mathbb{R}^2$ such that $I(S) - I(s) < \epsilon$, that is

$$\iint_{\mathbb{R}^2} S\,dA - \iint_{\mathbb{R}^2} s\,dA < \epsilon.$$

Intuitively, a function is Riemann integrable if it can be trapped between two step functions S and s in such a way that the integrals of S and s differ by an arbitrarily small amount.

DEFINITION: If f is Riemann integrable, let S_n and s_n be a *trapping sequence*, for f, that is, $s_n(X) \leq f(X) \leq S_n(X)$ and $I(S_n) - I(s_n) < \frac{1}{n}$. Then the *Riemann integral* of f, $I(f)$ is defined as (cf. page 21, for the definition of l.u.b. = least upper bound, and of g.l.b.).

$$I(f) \equiv \text{l.u.b.}_{n\to\infty} I(s_n)$$

We could have equivalently defined $I(f)$ as $I(f) = \text{g.l.b.}_{n\to\infty} I(S_n)$. Since both limits are the same, it is irrelevant. However, it is important to show that $I(f)$ has the same value if any other trapping sequence $\hat{S}_n(X)$, $\hat{s}_n(X)$ is used. This is the content of

Lemma 8.14 . *If f is Riemann integrable, then $I(f)$ does not depend on which trapping sequences are used. Proof not given.*

Now we exhibit a class of functions which are Riemann integrable. The issue boils down to finding functions which can be approximated well by step functions.

Lemma 8.15 . *If f is a continuous function and D is a closed and bounded set, then f can be approximated arbitrarily closely from above and below by step functions S and s throughout D. Thus, given any $\epsilon > 0$, there are step functions S and s such that*

$$0 \leq S(X) - f(x) < \epsilon, \quad \textit{and} \quad 0 \leq f(X) - s(X) < \epsilon \quad \textit{for all} \quad X \in D.$$

Proof not given.

Theorem 8.16 . *If f is a continuous function with compact support, then it is Riemann integrable.*

PROOF: Let $S(X)$ and $s(X)$ be as in the lemma where D is the support of f. Then

$$s(X) \leq f(X) \leq S(X)$$

and

$$S(X) - s(X) = [S(X) - f(X)] + [f(X) - s(X)] < 2\epsilon.$$

Thus by Lemma 1,

$$I(S) - I(s) = I(S - s) < 2c\epsilon,$$

where c is the area of the set (supp S) $\cup$ (supp s).

Because f has compact support, the constant c is bounded. Therefore the factor $2c\epsilon$ can be made arbitrarily small by choosing ϵ small. This verifies all the conditions for integrability.

We have disposed of the problem of integrating continuous functions with compact support. Notice that the above procedure is identical to that used for functions of one variable (see figure.)

We still do not know how to find the area of a domain D. Although we anticipate that Area$(D) = I(X_D)$, this does not yet make sense (except for rectangular complexes) since the *discontinuous* function X_D is not covered by Theorem 1). Let us remedy this now. The problem is to show the boundary ∂D does not have any area.

DEFINITION: A set in $\mathbb{R}^2$ has *content zero* if it can be enclosed in a rectangular complex whose total area is arbitrarily small. Thus, if a set has content zero, given any $\epsilon > 0$, there is a rectangular complex R containing ∂D such that

$$\text{Area}\,(R) = I(X_R) < \epsilon.$$

It should be clear that any set with a finite number of points has content zero (since each point can be enclosed on a square of side ϵ, so the total area of N such squares is $N\epsilon^2$, which can be made arbitrarily small.) One would also expect that curves will have zero content. This is not necessarily true unless the curve is not too badly behaved.

Lemma 8.17 . *If a curve is composed of a finite number of smooth curves, then it has zero content. In particular, if the boundary ∂D of a bounded domain D is such a curve, it has zero content. Proof not given.*

Theorem 8.18 . *If the boundary ∂D of a domain $D \subset \mathbb{R}^2$ has content zero, then the function X_D is Riemann integrable. Consequently, the area of D is definable and given by*

$$\textit{Area}\,(D) = \iint_{\mathbb{R}^2} X_D \, dA = \iint_D dA.$$

PROOF: Almost identical to that for Theorem 11. Let $\epsilon > 0$ be given and let R be the rectangular complex which encloses the boundary ∂D, where R has area less than ϵ, $I(X_R) < \epsilon$. Then the part of D which is enclosed by R, $D_- = D - R \cap D$, is a rectangular complex as is $D_+ = R \cup D_-$ and $D_+ - D_- = R$. Since $D_+ \supset D \supset D_-$, we have

$$X_{D_-}(X) \leq X_D(X) \leq X_{D_+}(X) \quad \text{for all} \quad X.$$

Also,

$$I(X_{D_+}) - I(X_{D_-}) = I(X_R) < \epsilon.$$

Thus X_D is trapped by the step functions $S = X_{D_+}$ and $s = S_{D_-}$ and $I(S) - I(s) < \epsilon$, proving the theorem.

It is now possible to define

$$\iint_D f \, dA$$

for continuous functions f where D is not necessarily the support of f.

Theorem 8.19 . *If f is continuous in a closed and bounded set D whose boundary ∂D has content zero, then the function f_{X_D} is Riemann integrable and*

$$\iint_D f \, dA \equiv I(f_{X_D}).$$

PROOF: Let R be the rectangular complex which encloses ∂D and has area less than ϵ, $I(X_R) < \epsilon$. Take $D_- = D - R \cap D$ and $D_+ = R \cup D_-$ as in Theorem 12. Further let S_1 and s_1 be step functions which trap f within ϵ for all $X \in D_-$ (this is possible by Lemma 3)

$$0 \le S_1(X) - f(X) < \epsilon,\ 0 \le f(X) - s_1(X) < \epsilon \quad \text{for all} \quad X \epsilon D_-,$$

so

$$0 \le S_1(X) - s_1(X) < 2\epsilon \quad \text{for all} \quad X \in D.$$

Let M be an upper bound for $|f|$ on D, $|f(X)| \le M$ for all $X \in D_-$. Then define

$$S = S_1 + MX_R \quad \text{and} \quad s = s_1 - MX_R.$$

These functions S and s trap f on all of D,

$$s(X) \le f(X) \le S(X) \quad \text{for all} \quad X \in D,$$

that is,

$$s \le fX_D \le S \quad \text{for all} \quad X.$$

Furthermore

$$\begin{aligned} I(S-s) &= I(S_1 - s_1) + 2MI(X_R) \\ &< 2c\epsilon + 2M\epsilon = (2c + 2M)\epsilon, \end{aligned}$$

where c is the area of D_-. Since S and s are step functions which trap f, and since $I(S-s)$ can be made arbitrarily small, the proof that f_{X_D} is Riemann integrable is completed. We follow custom and write

$$I(fX_D) \equiv \iint_D f\,dA.$$

Except for the three unproved lemmas, this completes the proof of the existence of the integral. The next theorem summarizes some important properties of the integral.

Theorem 8.20 . *If f and g are Riemann integrable, then*
a). $I(af + bg) = aI(f) + bI(g)$, $\quad a, b$ constants
b). $f \le g$ implies $I(f) \le I(g)$.
c). $|I(f)| \le I(|f|)$

PROOF:
a) and b) are immediate consequences of the corresponding statements for step functions (Lemma 1) and the definition of the Riemann integral as the limit of step functions. To prove c), we first observe that if f is integrable, so is $|f|$. Since $-|f| \le f \le |f|$, by parts a and b

$$-I(|f|) \le I(f) \le I(|f|),$$

which is equivalent to the stated property.

Although the approximate value of the integral $\iint_D f\,dA$ can be evaluated by using the procedures of the above theorems, we have as yet no routine way of evaluating the integral if f and D are simple. Some notation will suggest the method. Write $dA = dx\,dy$ and think of $dx\,dy$ as the area of an "infinitesimal" rectangle. Then

$$\iint_D f\,dA = \iint_D f(x,y)\,dx\,dy.$$

If D is the domain in the figure, it is reasonable to evaluate the double integral, which we shall think of as the mass of D with density f, by first finding the mass of a horizontal strip

$$g(y) = \int_{\gamma_1}^{\gamma_2} f(x,y)\,dx,$$

and then adding up the horizontal strips to find the total mass

$$\iint_D f(x,y)\,dx\,dy = \int_{\gamma_3}^{\gamma_4} g(y)\,dy = \int_{\gamma_3}^{\gamma_4} \left(\int_{\gamma_1}^{\gamma_2} f(x,y)\,dx \right) dy.$$

The integral on the right is called an *iterated* or *repeated* integral. In a similar way, one could begin with mass of vertical strips

$$h(x) = \int_{\gamma_3}^{\gamma_4} f(x,y)\,dy$$

and add these up

$$\iint_D f(x,y)\,dx\,dy = \int_{\gamma_1}^{\gamma_2} h(x)\,dy = \int_{\gamma_1}^{\gamma_2} \left(\int_{\gamma_3}^{\gamma_4} f(x,y)\,dy \right) dx.$$

For most purposes, it is sufficient to consider domains which are of the two types pictured

A FIGURE GOES HERE

that is, D is bounded on two sides by straight line segments. More complicated domains can be treated by decomposing them into domains of these two types, where one or both of the straight line segments might degenerate to a point.

Theorem 8.21 . *If f is continuous on a domain D_1 (respectively D_2) as above, then the iterated integral*

$$\int_a^b \left(\int_{\phi_1(x)}^{\phi_2(x)} f(x,y)\,dy \right) dx \quad [\textit{resp.} \int_\alpha^\beta \left(\int_{\phi_1(y)}^{\phi_2(y)} f(x,y)\,dx \right) dy]$$

exists and equals

$$\iint_D f\,dA.$$

Proof not given. It is rather technical.

REMARK: If a domain D happens to be of both types (as, for example, rectangles and triangles are) then either iterated integral can be used and yield the same result - since they are both equal $\iint_D f\,dA$. See Examples 1 and 3 below (Example 2 could also have been done both ways).

EXAMPLES:

(1) Evaluate $\iint_D f\,dA$ where $f(x,y) = x^2y$ and D is the rectangle in the figure. We shall integrate with respect to x first.

$$\iint_D f\,dA = \int_1^2 \left(\int_1^3 (x^2+xy)\,dx\right)\,dy.$$

The inner integral is the mass of a strip. Think of y as being the fixed height of the strip. Then

$$\int_1^3 (x^2+xy)\,dx = \frac{x^3}{3} + \frac{x^2y}{2}\Big|_{x=1}^{x=3} = 9 + \frac{9y}{2} - \frac{1}{3} - \frac{y}{2} = \frac{26}{3} + 4y$$

Therefore, adding up all the strips we find

$$\iint_D f\,dA = \int_1^2 (\frac{26}{3} + 4y)\,dy = (\frac{26}{3}y + 2y^2)\Big|_{y=1}^{y=2} = \frac{26}{3} + 6 = \frac{44}{3}$$

Let us evaluate this again, now integrating first with respect to y.

$$\iint_D f\,dA = \int_1^3 \left(\int_1^2 (x^2+xy)\,dy\right)\,dx.$$

First

$$\int_1^2 (x^2+xy)\,dy = (x^2y + \frac{xy^2}{2})\Big|_{y=1}^{y=2} = x^2 + \frac{3}{2}x$$

so

$$\iint_D f\,dA = \int_1^3 (x^2 + \frac{3}{2}x)\,dx = (\frac{x^2}{3} + \frac{3}{4}x^2)\Big|_{x=1}^{x=3} = \frac{44}{3},$$

which agrees with the previous computation. Instead of imagining f as the density of D, one can also take f to be the height function of a surface above D. Then the integral $\iint_D f\,dA$ is the volume of the solid whose base is D and whose "top" is the surface M with points $(x,y,f(x,y))$. In this case, the volume is 44/3.

(2) Evaluate $\iint_D f\,dA$ where $f(x,y) = x^2+xy+2$ and D is the domain bounded by the curves $\phi_1(x) = 2x^2$, $\phi_2(x) = 4+x^2$, and $x=0$.

Integrate first with respect to y. Then y varies between $2x^2$ and $4+x^2$, while x varies between the two straight lines $x=0$ and $x=2$.

$$\iint_D f\,dA = \int_0^2 \left(\int_{2x^2}^{4+x^2} (x^2+xy+2)\,dy\right)\,dx$$
$$= \int_0^2 (x^2y + \frac{xy^2}{2} + 2y)\Big|_{y=2x^2}^{y=4+x^2}\,dy$$
$$\int_0^2 (8 + 8x + 2x^2 + 4x^3 - x^4 - \frac{3}{2}x^5)\,dy = \frac{464}{15}$$

(3) Evaluate $\iint_D f\,dA$ where $f(x,y) = (x-2y)^2$ and D is the triangle bounded by $x=1, y=-2$, and $y+2x=6$.

We shall integrate first with respect to x. Then x varies between $x=1$ and $x=-\frac{1}{2}y+2$, while y varies between the lines $y=-2$ and $y=2$.

$$\iint_D f\,dA = \int_{-2}^2 \int_1^{\frac{1}{2}y+2} (x-2y)^2\,dx\,dy$$

Since

$$\int_1^{-\frac{1}{2}y+2} (x-2y)^2(x-2y)^2\,dx = \frac{1}{3}(x-2y)^3\Big|_{x=1}^{x=-\frac{1}{2}y+2} = \frac{1}{3}(2-\frac{5}{2}y)^3 - \frac{1}{3}(1-2y)^3,$$

we find

$$\iint_D f\,dA = \frac{1}{3}\int_{-2}^2 [(2-\frac{5}{2}y)^3 - (1-2y)^3]\,dy = \frac{164}{3}.$$

One can also integrate first with respect to y. Then y varies between $y=-2$ and $y=-2x+6$, while x varies between the lines $x=1$ and $x=3$.

$$\iint_D f\,dA = \int_1^3 \left(\int_{-2}^{-2x+4} (x-2y)^2\,dy\right)\,dx.$$

Since

$$\int_{-2}^{-2x+4} (x-2y)^2\,dy = -\frac{1}{6}(x-2y)^3\Big|_{y=-2}^{y=-2x+4} = -\frac{1}{6}[(5x-8)^3 - (x+4)^3]$$

we again find

$$\iint_D f\,dA = -\frac{1}{6}\int_1^3 [(5x-8)^3 - (x+4)^3]\,dx = \frac{164}{3}.$$

(4) Find the volume of the pyramid P bounded by the four planes $x = 0, y = 0, z = 0$, and $x + y + z = 1$. The easiest way to do this is to let $z = f(x, y) = 1 - x - y$ be the height function of the tilted plane which we shall take as the top of the pyramid which lies above the triangle D (in the xy plane) which is bounded by the three lines $x = 0, y = 0$, and $x + y = 1$. Then

$$\text{Volume}\,(P) = \iint_D f(x, y)\,dx\,dy$$

One can integrate with respect to either x or y first. We shall do the x integration first.

$$\iint_D f\,dA = \int_0^1 \left(\int_0^{1-y} (1 - x - y)\,dx \right) dy.$$

Since

$$\int_0^{1-y} (1 - x - y)\,dx = -\frac{1}{2}(1 - x - y)^2\Big|_{x=0}^{x=1-y} = \frac{1}{2}(1 - y)^2$$

we find

$$\text{Volume}\,(P) = \iint_D f\,dA = \frac{1}{2}\int_0^1 (1 - y)^2\,dy = -\frac{1}{6}(1 - y)^3\Big|_0^1 = \frac{1}{6}.$$

This agrees with the usual formula for the volume of a pyramid

$$\text{Vol} = \frac{1}{3}\text{altitude} \times \text{area of base}.$$

The identical methods work for triple integrals. All of the theorems and proofs remain unchanged. Again the integral

$$\iiint_D f\,dV$$

can either be interpreted as the mass of a solid D with density f, or as the "volume" of a four dimensional solid whose base is D and top in the surface with points $(x, y, z, f(x, y, z))$. Because of conceptual difficulties, one usually thinks of f as a density. Calculation of triple integrals is done by evaluating three integrals, as

$$\iiint_D f\,dV = \int \left(\int \left(\int f(x, y, z)\,dz \right) dy \right) dx,$$

where the limits in the iterated integral on the right are determined from the domain D. An example should illustrate the idea adequately,

Example: Evaluate $\iiint_D f\,dV$ where $f(x, y, z) \equiv c$ and D is the solid bounded by the two planes $z \equiv 0$, $y \equiv 2$, and the surface $z \equiv -x^2 + y^2$. We have to evaluate $\iiint_D c\,dV$ which is the mass of the solid D with constant density c, that is c times the volume of D. It is convenient to carry out the z integration first, then the x integration

$$\iiint_D c\,dV = c\int_0^2 \left(\int_{-y}^{y} \left(\int_0^{-x^2+y^2} dz \right) dx \right) dy.$$

The x limits of integration have been found by looking at the region of integration in the xy plane beneath the surface $z = -x^2 + y^2$. This region, found by setting $z = 0$, consists of the points between the straight lines $0 = -x^2 + y^2$, that is between the lines $x = y$ and $x = -y$. Then

$$\iiint_D f\,dV = c\int_0^2 \left(\int_{-y}^{y} (-x^2 + y^2)\,dx \right) dy$$

$$= c\int_0^2 (-\frac{x^3}{3} + xy^2)\Big|_{x=-y}^{x=y} dy = c\int_0^2 \frac{4}{3}\,dy = \frac{16}{3}c.$$

By letting $c = 1$, the volume of the solid is seen to be $16/3$.

Exercises

(1) Evaluate $\iint_D xy\,dx\,dy$ for the following domains D in two ways: $\int (\int xy\,dx)\,dy$ and $\int (\int xy\,dy)\,dx$.

(a) D is the rectangle with vertices at $(1, 1)$, $(1, 5)$, $(3, 1)$ and $(3, 5)$.

(b) D is the triangle with vertices at $(1, 1)$, $(3, 1)$ and $(3, 5)$.

(c) D is the region enclosed by the lines $x = 1$, $y = 2$, and the curve $y = x^3$ (a curvilinear triangle).

(d) D is the region enclosed by the curves $y = x^2$ and $y = \sqrt{x}$.

(2) Evaluate

$$\iint_D \sin \pi(2x + y)\,dx\,dy,$$

where D is the triangle bounded by the lines $x = 1$, $y = 2$ and $x - y = 5$.

(3) Evaluate

$$\iint_D (xy - y_3)\,dx\,dy,$$

where D is the region enclosed by the lines $x = -1$, $x = 1$, $y = -2$ and the curve $y = 2 - x^2$.

(4) Evaluate

$$\int_D (xy + z)\, dx\, dy\, dz,$$

where D is the rectangular parallelepiped bounded by the six planes $x = -2$, $y = 1$, $z = 0$, $x = 1$, $y = 2$, $z = 3$.

(5) Evaluate

$$\iiint_D xyz\, dx\, dy\, dz,$$

where D is the solid enclosed by the paraboloid $z = x^2 + y^2$ and the plane $z = 4$.

(6) Find the volume of an octant of the ball $x^2 + y^2 + z^2 \le a^2$ in two ways;

(a) by evaluating

$$\iint_D f(x, y) dx\, dy$$

where f is a suitable function and D a suitable domain

(b) by evaluating

$$\iiint_D dx\, dy\, dz,$$

where D is the ball.

(7) If $f(x, y) > 0$ is the density function of a plate D, the x and y coordinates of the *center of mass* $(\bar{x}, \bar{y})$ are defined by

$$\overline{x} = \frac{\iint_D x f(x, y)\, dx\, dy}{\iint_D f(x, y)\, dx\, dy}, \qquad \overline{y} = \frac{\iint_D y f(x, y)\, dx\, dy}{\iint_D f(x, y)\, dx\, dy}.$$

Find the center of mass of a triangle whose vertices are at the points $(0, 0)$, $(0, 4)$, and $(2, 0)$, and whose density is $f(x, y) = xy + 1$.

(8) The *moment of inertia with respect to a point* $p = (\xi, \eta)$ of a plate D with density $f(x, y)$ is defined by

$$J_p(D) = \iint_D [(x - \xi)^2 + (y - \eta)^2] f(x, y) dx\, dy.$$

(a) Find the moment of inertia of the plate in Exercise 7, with respect to the point $p = (1, 0)$.

(b) If D is any plate (with sufficiently smooth boundary), prove that the moment of inertia is smallest if the point $f = (\xi, \eta)$ is taken to be the center of mass of D. [Hint: Consider J as a function of the two variables ξ and η and show J has a minimum at $(\bar{x}, \bar{y})$.]

(9) (a) Show that

$$\iint_D f_{xy}(x, y) dx\, dy = f(p_1) - f(p_2) + f(p_3) - f(p_4),$$

where D is a rectangle with vertices at p_1, p_2, p_3, p_4 (see fig.).

(b) Use the result of part (a) to again evaluate the integral in Ex. 1a.

(c) If $U(x, y)$ satisfies the partial differential equation $U_{xy} = 0$ for $0 < y < x$ and $U(x, x) = 0$ while $U(x, 0) = x \sin x$, find $U(x, y)$ for all points (x, y) in the wedge $0 < y < x$. [Answer: $U(x, y) = x\, \sin x - y\, \sin y$ for $0 < y < x$].

(10) Let $f(x, y)$ be a bounded function which is continuous except as a set of points of content zero, and suppose f has compact support. Prove that f is Riemann integrable. This again proves Theorem 13.

(11) Let D_1 and D_2 be domains whose boundaries have zero content and whose intersection $D_1 \cap D_2$ has zero content.

(a) If f is continuous on $D_1 \cup D_2$, prove that the integral $\iint_{D_1 \cup D_2} f\, dA$ exists and that

$$\iint_{D_1 \cup D_2} f\, dA = \iint_{D_1} f\, dA + \iint_{D_2} f\, dA.$$

(b) Give an example showing the above equality does not hold if $D_1 \cap D_2$ has non-zero content.

(12) (a) By an explicit construction, show that the region $D = \{(x, y) \epsilon \mathbb{R}^2 : |x| + |y| \le 1\}$ has boundary with zero content.

(b) By an explicit construction, show that the circle $? = \{(x, y) \epsilon \mathbb{R}^2 : x^2 + y^2 = 1\}$ has zero content.

(13) (a) By interchanging the order of integration, show that

$$\int_0^x \Big(\int_0^s f(t)\, dt\Big)\, ds = \int_0^x (x - t) f(t)\, dt.$$

(b) $\int_0^x \Big(\int_0^2 \Big(\int_0^r f(t)\, dt\Big) dr\Big)\, ds = ?$

(14) Let D be a plate in the x, y plane with density f and total mass M. If $p = (\xi, \eta)$ is an arbitrary point in the plane and $\bar{p} = (\bar{x}, \bar{y})$ is the center of mass of D, prove

$$J_p(D) = J_{\bar{p}}(D) + M \|p - p_0\|^2,$$

where the notation of Exercise 8 has been used. This is the *parallel axis theorem*. It again proves the result of Exercise 8b.

Chapter 9

Differential Calculus of Maps from $\mathbb{R}^n$ to $\mathbb{R}^m$.

9.1 The Derivative

.

Now we generalize the ideas of Chapters 7 and 8 and consider nonlinear mappings from a set D in $\mathbb{R}^n$ to $\mathbb{R}^m$, $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$, or $Y = F(X)$, where $X \in D$ and $Y \in \mathbb{R}^m$. In coordinates, these functions look like

$$\begin{aligned} y_1 &= f_1(x_1, \dots, x_n) \\ &\vdots \\ y_m &= f_m(x_1, \dots, x_n) \end{aligned}$$

where the functions f_j are scalar-valued. The special case $n = 1$, m arbitrary, was treated in Chapter 7, section 3, while the special case $m = 1$, n arbitrary, was treated in Chapter 8.

One interpretation of maps $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ is as a geometric transformation from some subset D of $\mathbb{R}^n$ into all or part of $\mathbb{R}^m$.

EXAMPLES.

(1) The affine map $Y = F(X)$ defined by

$$y_1 = 2 + x_1 - 2x^2$$

$$y_2 = 1 + x_1 + x_2$$

maps $\mathbb{R}^2$ into $\mathbb{R}^2$. Under this map, the origin goes into $(2, 1)$, the x_1 axis (i.e. the line $x_2 = 0$) goes into the line $y_1 - y_2 = 1$,

A FIGURE GOES HERE

while the x_2 axis goes into the line $y_1 + 2y_2 = 4$. The shaded region indicates the image of the indicated square.

(2) The map $Y = F(X)$ defined by

$$y_1 = x_1 - x_2$$

$$y_2 = x_1^2 + x_2^2$$

maps all of $\mathbb{R}^2$ onto the upper half y_1y_2 plane (since $y_2 \geq 0$). Let us see what happens to a rectangle under this mapping. Consider the rectangle R in the figure. The x_1 axis, $x_2 = 0$, goes into the parabola $y_2 = y_1^2$, and the line $x_2 = 1$ into $y_2 = 1 + (y_1 + 1)^2$.

A FIGURE GOES HERE

Similarly, the line $x_1 = 1$ is mapped into $y_2 = 1 + (y_1 - 1)^2$, while $x_1 = 2$ is mapped into $y_2 = 4 + (y_1 - 2)^2$. By following the images of the boundary ∂R, we now see that the interior of R is mapped into the shaded curvilinear "parallelogram". This mapping, though injective when restricted to our rectangle, is not injective for all $(x_1, x_2) \in \mathbb{R}^2$, since, for example, the points $X_1 = (1, 2)$ and $X_2 = (-2, -1)$ are both mapped into the same point $(-1, 5)$.

(3) The function $w = x_1^2 + x_2^2$ whose graph is a paraboloid, is a map from $\mathbb{R}^2$ into $\mathbb{R}^1$. It can also be regarded as a map from $\mathbb{R}^2$ into $\mathbb{R}^3$ by a useful artifice. Let $y_1 = x_1$, $y_2 = x_2$, and $y_3 = w = x_1^2 + x_2^2$. Then

$$y_1 = x_1$$

$$y_2 = x_2$$

$$y_3 = x_1^2 + x_2^2$$

is a map F from $\mathbb{R}^2$ into $\mathbb{R}^3$. The image of the unit square (see figure) is then the shaded region in the figure above the image (y_1, y_2) of the square R

A FIGURE GOES HERE

(4) The map $F : \mathbb{R}^2 \to \mathbb{R}^3$ defined by (cf. example 2)

$$y_1 = x_1 - x_2$$

$$y_2 = x_1^2 + x_2^2$$

$$y_3 = x_1 + x_2$$

also represents a surface M. In fact, since $y_1^2 + y_3^2 = 2y_2$, this surface is a paraboloid opening out on the y_2 axis. Again, we investigate where the rectangle R of example 2 is mapped. Since the y_1 and y_2 components of the mapping are the same as before, the image of R will lie on the surface M above the image (y_1, y_2) of (x_1, x_2). Thus the image of the rectangle R is a patch of the surface M.

From these examples, we see it is natural to regard any map $F : D \subset \mathbb{R}^2 \to \mathbb{R}^m$ as an ordinary surface, or two dimensional manifold, embedded in $\mathbb{R}^m$, much as a map $F : D \subset \mathbb{R}^1 \to \mathbb{R}^m$ was regarded as an ordinary curve. In the case $m = 1$, the surface $F : D \subset \mathbb{R}^2 \to \mathbb{R}^1$ was representable as the graph of the function F. For $m = 2$ and higher, this surface is seen as the range of the map. In the same way, an n dimensional surface, or manifold, embedded in $I\mathbb{R}^m$ is a map $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$. You might want to think of n as being the number of "degrees of freedom" on the manifold. In a strict sense, the map $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ is not an n manifold *embedded* in $\mathbb{R}^m$ unless $\mathbb{R}^m$ is big enough to hold an m manifold, i.e. $m \geq n$. However by either using the graph of F, a subset of $\mathbb{R}^{m+n}$, or by using the trick of example 3 we can always think of the map $F : \mathbb{R}^n \to \mathbb{R}^m$ as an n dimensional surface. For $m \geq n$, this surface can be embedded as a subset of $\mathbb{R}^n$.

There are several valuable physical interpretations of these vector valued functions of a vector, $Y = F(X)$. Consider a fluid flowing through a domain D in $\mathbb{R}^3$. The fluid could be air and D as the outside of an airplane, or the fluid could be an organic fluid, and D as some portion of the body.

The velocity V of a particle of fluid is a three vector which depends upon the space coordinate (x_1, x_2, x_3) as well as the time coordinate t of the particle, $V = F(x_1, x_2, x_3, t) = F(X, t)$. This velocity vector $V(X, t)$ at X points in the direction the fluid is moving. Thus, the velocity function is an example of a mapping from space-time $\mathbb{R}^3 \times \mathbb{R}^1 \cong \mathbb{R}^4$ into vectors in $\mathbb{R}^3$. In this case, we think of the velocity vector $V = F(X, t)$ as having its foot at the point $X \in D$ and imagine the mapping as the domain D along with a vector V attached to each point of D (see fig. above). One calls this a *vector field* defined on the domain D, since it assigns a vector to each point of D.

A very common vector field is a field of forces. By this we mean that to every point X of a domain D, we associate a vector $F(X)$ equal to the force an object at X "feels". If the forces are time dependent, then the force field is written $F(X, t)$, $X \in D$. You are most familiar with the force field due to gravity. If e_3 is the direction toward the center of the earth, and say e_1 points east and e_2 north along the surface of the earth (other coordinates must be chosen for the north and south poles), then the gravitational force is usually written as $F = (0, 0, g)$, a constant vector pointing down to the center of the earth. For more precise purposes, one must take into account the fact that g does vary from place to place of the earth's surface. Then $F(x) = (0, 0, g(X))$. In even more accurate experiments - or in outer space - must further account for the effect of the other heavenly bodies. This brings in the other components of force as well as a time dependence due to the motion of the earth, $F(X, t) = (f_1(X, t),\ f_2(X, t),\ f_3(X, t))$. The force field is imagined as a vector attached to each point X in space, the vector having the magnitude and direction of the net force F there.

An entirely different example of a mapping F from $\mathbb{R}^n$ to $\mathbb{R}^m$ is a factory - or an even larger economic system. The vector $X = (x_1, x_2, \dots, x_n)$ might represent the quantities $x_1, x_2, \dots$ of different raw materials needed. $Y = F(X)$ could then represent the output from the factory, the number y_j being the quantity of the jth product produced from the input X.

Turning to the quantitative mathematical aspect of the mappings $F : \mathbb{R}^n \to \mathbb{R}^m$, we define the derivative. The definition will be formal, patterned directly on the definition of the total derivative given previously (p. 578-9).

DEFINITION: Let $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ and X_0 be an interior point of D. F is *differentiable at* X_0, if there exists a linear transformation $L_{(X_0)} : \mathbb{R}^n \to \mathbb{R}^m$, depending on the base point X_0, such that

$$\lim_{\|h\|\to 0} \frac{\|F(X_0+h) - F(X_0) - L_{(X_0)}h\|}{\|h\|} = 0$$

for any vector h in some sufficiently small ball about X_0. If F is differentiable at X_0, we shall use the notations

$$\frac{dF}{dX}(X_0) = F'(X_0) = L_{(X_0)}$$

and refer to them as the derivative of F at X_0. If $F'(X_0)$ depends continuously on the base point X_0 for all X_0 in D, then F is said to be *continuously differentiable* in D, written $F \in C^1(D)$.

Many of the results from Chapter 8 Sections 1 and 2 generalize immediately to the present situation.

Proposition 9.1 . *The function $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at the interior point $X_0 \in D$ if and only if there is a linear operator $L_{(X_0)} : \mathbb{R}^n \to \mathbb{R}^m$ and a function $R(X_0, h)$ such that*

$$F(X_0+h) = F(X_0) + L_{(X_0)}h + R(X_0,h) \quad \|h\|,$$

where the remainder $R(X_0, h)$ has the property

$$\lim_{\|h\|\to 0} \|R(X_0,h)\| = 0.$$

PROOF: $\Leftarrow$ If F is differentiable at X_0, let $L_{(X_0)}$ be the derivative and take $R(X_0,h) = [F(X_0+h) - F(X_0) - L_{(X_0)}h]/\|h\|$. Then this $L_{(X_0)}$ and $R(X_0,h)$ do satisfy the above conditions.

$\Rightarrow$ If $L_{(X_0)}$ and $R(X_0,h)$ are as above, then

$$\lim_{\|h\|\to 0} \frac{\|F(X_0+h) - F(X_0) - L_{(X_0)}h\|}{\|h\|} = \lim_{\|h\|\to 0} \|R(X_0,h)\| = 0.$$

Since $L_{(X_0)}$ is linear, this proves F is differentiable at X_0.

There is at most one derivative operator $L_{(X_0)}$, that is

Proposition 9.2 . *(Uniqueness of the derivative). Let $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ be differentiable at the interior point $X_0 \in D$. If $\hat{L}_{(X_0)}$ and $\tilde{L}_{(X_0)}$ are linear operators both of which satisfy the conditions for the derivative of F and X_0, then $\hat{L}_{(X_0)} = \tilde{L}_{(X_0)}$.*

PROOF: Word for word the same as the proof of Theorem 1, page 579-80.

If the map $F = F(X)$ is given in terms of coordinates,

$$\begin{aligned} y_1 &= f_1(x_1, \dots, x_n) \\ y_2 &= f_2(x_1, \dots, x_n) \\ &\cdot \\ &\cdot \\ &\cdot \\ y_m &= f_m(x_1, \dots, x_n), \end{aligned}$$

how is the derivative computed, and what is its relationship to the derivative of the individual coordinate functions f_j? The answer is contained in

Theorem 9.3 . *Let F map $D \subset \mathbb{R}^n$ into $\mathbb{R}^m$ be given in terms of the coordinate functions $f_j(X), \quad j = 1, \dots, m$*

$$\begin{aligned} y_1 &= f_1(X) \quad f_1(x_1, \dots, x_n) \\ &\cdot \\ &\cdot \\ &\cdot \\ y_m &= f_m(X) \quad = f_m(x_1, \dots, x_m). \end{aligned}$$

(a) Then F is differentiable or continuously differentiable at the interior point $X_0 \in D$ if and only if all of the f_j's are respectively differentiable or continuously differentiable.

(b) Moreover, if F is differentiable at X_0, then the derivative in these coordinates is given by the $m \times n$ matrix of partial derivatives

$$L_{(X_0)} := F'(X_0) = \begin{pmatrix} f_1'(X_0) \\ \cdot \\ \cdot \\ \cdot \\ f_m'(X_0) \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & (X_0), \dots, \frac{\partial f_1}{\partial x_n}(X_0) \\ \cdot & \\ \cdot & \\ \cdot & \\ \frac{\partial f_m}{\partial x_1} & (X_0), \dots, \frac{\partial f_m}{\partial x_n}(X_0) \end{pmatrix}.$$

The matrix is sometimes called the Jacobian matrix.

PROOF: (a) Observe that the limit

$$\lim_{\|h\|\to 0} \frac{\|F(X_0+h) - F(X_0) - L_{(X_0)}h\|}{\|h\|} = 0$$

exists if and only if each of its components tend to zero,

$$\lim_{\|h\|\to 0} \frac{\|f_j X_0 + h) - f_j(X_0) - L_{j(X_0)}h\|}{\|h\|} = 0, \quad j = 1, 2, \dots, m.$$

Thus, if F is differentiable at X_0, each of the coordinate functions f_j are differentiable and have total derivative $L_{j(X_0)}$. Conversely, if each of the coordinate functions are differentiable at X_0, all of the above limits exist so the vector valued function F is also differentiable.

(b) Since the differentiability of F implies that of the coordinate vectors, we have

$$F'(X_0) = \begin{pmatrix} f_1'(X_0) \\ \cdot \\ \cdot \\ \cdot \\ f_m'(X_0) \end{pmatrix}.$$

The result now follows by writing out each of the derivatives

$$f_1'(X_0) = (\frac{\partial f_1(X_0)}{\partial x_1}, \dots, \frac{\partial f_1(X_0)}{\partial x_n})$$

$f_2'(X_0) = \dots$ etc. and then inserting these in the expression for $F'(X_0)$.

Corollary 9.4 . *A function $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable in D if and only if all the partial derivatives of its components $\partial f_i / \partial x_j$ exist and are continuous.*

Proof: This follows from this theorem and Theorem 3, p. 585.

EXAMPLES.

1. Let F be an affine map from $\mathbb{R}^n$ to $\mathbb{R}^m$

$$F(X) = Y_0 + BX,$$

where B is a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$ (which you may choose to think of as an $m \times n$ matrix with respect to some coordinate system) and $Y_0 = F(0)$ is a fixed vector in $\mathbb{R}^m$. Then F is differentiable at every point of $\mathbb{R}^n$ and it given by the eminently reasonable formula

$$F'(X_0) = B,$$

where the operator B does not depend on X_0. For proof, we observe that

$$F(X_0 + h) - F(X_0) = Y_0 + B(X_0 + h) - [Y_0 + BX_0] = Bh.$$

Thus

$$\lim_{\|h\| \to 0} \frac{\|F(X_0 + h) - F(X_0) - Bh\|}{\|h\|} = \lim_{\|h\| \to 0} \frac{0}{\|h\|} = 0.$$

Since B is linear, this shows the derivatives exists and is B. Let us do this again in coordinates. If $B = ((b_{ij}))$ the function F is

$$\begin{aligned} f_1(X) &= y_{01} + b_{11}x_1 + b_{12}x_2 + \dots + b_{1n}x_n \\ f_2(X) &= y_{02} + b_{21}x_1 + \dots + b_{2n}x_n \\ &\cdot \\ &\cdot \\ &\cdot \\ f_m(X) &= y_{0m} + b_{m1}x_1 + \dots + b_{mn}x_n. \end{aligned}$$

Therefore each of the functions f_j is clearly differentiable and

$$\begin{aligned} f_1' &= (\tfrac{\partial f_1}{\partial x_1}, \dots, \tfrac{\partial f_1}{\partial x_n}) &&= (b_{11}, \dots, b_{1n}) \\ &\cdot \\ &\cdot \\ &\cdot \\ f_m' &= (\tfrac{\partial f_m}{\partial x_1}, \dots, \tfrac{\partial f_m}{\partial x_n}) &&= (b_{m1}, \dots, b_{mn}). \end{aligned}$$

Consequently,

$$F'(X_0) = \begin{pmatrix} f_1'(X_0) \\ \cdot \\ \cdot \\ \cdot \\ f_m'(X_0) \end{pmatrix} = \begin{pmatrix} b_{11}, & \dots, & b_{1m} \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ b_{m1}, & \dots, & b_{mn} \end{pmatrix} = B,$$

which agrees with the result obtained without coordinates.

2. Let $F : \mathbb{R}^2 \to \mathbb{R}^3$ be defined by

$$f_1(x_1, x_2) = 2 - x_1 + x_2^2$$

$$f_2(x_1, x_2) = x_1 x_2 - x_2^3$$

$$f_3(x_1, x_2) = x_1^2 - 3x_1 x_2.$$

Since each of the coordinate functions f_j are continuously differentiable, so is F. Because

$$f_1'(X) = (-1, 2x_2), \quad f_2'(X) - (x_2, x_1 - 3x_2^2), \quad f_3'(X) = (2x_1 - 3x_2, -3x_1),$$

we find that at $X_0 = (3, 1)$

$$F'(X_0) = \begin{pmatrix} f_1'(X_0) \\ f_2'(X_0) \\ f_3'(X_0) \end{pmatrix} = \begin{pmatrix} -1 & 2 \\ 1 & 0 \\ 3 & -9 \end{pmatrix}.$$

If X is near X_0, then by Proposition 1 with $h = X - X_0$

$$F(X) = F(X_0) + f'(X_0)(X - X_0) + \text{remainder}$$

$$= \begin{pmatrix} 0 \\ 2 \\ 3 \end{pmatrix} + \begin{pmatrix} -1 & 2 \\ 1 & 0 \\ 3 & -9 \end{pmatrix} \begin{pmatrix} x_1 - 3 \\ x_2 - 1 \end{pmatrix} + \text{remainder},$$

where the remainder term becomes less significant the closer X is to X_0.

Motivated by our previous work, it is natural to formally define the tangent map as follows.

DEFINITION: Let $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ be differentiable at the interior point $X_0 \in D$. The *tangent* map at $F(X_0)$ to the (hyper) surface defined by F is defined to be the affine mapping

$$\Phi(X) = F(X_0) + f'(X_0)(X - X_0).$$

EXAMPLES:

(1) Let F be the function of Example 2 above. Then the tangent map at $X_0 = (3, 1)$ is

$$\Phi(X) = \begin{pmatrix} 0 \\ 2 \\ 3 \end{pmatrix} + \begin{pmatrix} -1 & 2 \\ 1 & 0 \\ 3 & -9 \end{pmatrix} \begin{pmatrix} x_1 - 3 \\ x_2 - 1 \end{pmatrix}.$$

(2) Let F be the function of Example 4 (page 679). Then

$$F'(X) = \begin{pmatrix} 1 & -1 \\ 2x_1 & 2x_2 \\ 1 & 1 \end{pmatrix}.$$

Thus the tangent map at $(2, 1)$ is

$$\Phi(X) = \begin{pmatrix} 1 \\ 5 \\ 3 \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ 4 & 2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 - 2 \\ x_2 - 1 \end{pmatrix}$$

If we let $Y = \Phi(X)$, then the target plane in the tangent space is found from

$$y_1 = 1 + (x_1 - 2) - (x_2 - 1)$$

$$y_2 = 5 + 4(x_1 - 2) + 2(x_2 - 1)$$

$$y_3 = 3 + (x_1 - 2) + (x_2 - 1)$$

By eliminating x_1 and x_2 from these equations, we find $y_2 = -5 + y_1 + 3y_3$. A graph of the surface M and the tangent plane can now be drawn.

A FIGURE GOES HERE

The next result is the generalization of the mean value theorem.

Theorem 9.5 . *(Mean Value Theorem). Let $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ be differentiable at every point of D, where D is an open convex set in $\mathbb{R}^n$. If $F'(X)$ is bounded in D, that is, if there is a constant $\gamma < \infty$ such that $\left|\frac{\partial f_i}{\partial x_j}(X)\right| \le \gamma$ for all $X \in D$ and for all $i = 1, \ldots, m$, and $j = 1, \ldots, n$, then*

$$\|F(X_2) - F(X_1)\| \le c\|X_2 - X_1\|$$

for all X_1 and X_2 in D, where $C = \sqrt{nm}\gamma$.

PROOF: The idea is to use the components of F and to appeal to the similar theorem (p. 597-8) for the function from $\mathbb{R}^n \to \mathbb{R}^1$. By that theorem, if X_1 and X_2 are in D, then there is a point Z_1 on the line segment joining X_1 to X_2 such that

$$f_1(X_2) = f_1(X_1) + f_1'(Z_1)(X_2 - X_1),$$

and similarly for the other components $f_2, f_3, \ldots, f_m$. Thus

$$\begin{pmatrix} f_1(X_2) \\ \cdot \\ \cdot \\ \cdot \\ f_m(X_2) \end{pmatrix} = \begin{pmatrix} f_1(X_1) \\ \cdot \\ \cdot \\ \cdot \\ f_m(X_1) \end{pmatrix} = \begin{pmatrix} f_1'(Z_1) \\ \cdot \\ \cdot \\ \cdot \\ f_m'(Z_m) \end{pmatrix} (X_2 - X_1),$$

where $Z_1, \ldots, Z_m$ are all on the segment joining

A FIGURE GOES HERE

X_1 to X_2. Observe that the $f_j'(Z_j)$'s are all vectors. Let L be the matrix of derivatives in the last term above, that is

$$L = \begin{pmatrix} f_1'(Z_1) \\ \cdot \\ \cdot \\ \cdot \\ f_m'(Z_m) \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_2}(Z_1) & \cdots & \frac{\partial f_1}{\partial x_n}(Z_1) \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ \frac{\partial f_m}{\partial x_1}(Z_m) & \cdots & \frac{\partial f_m}{\partial x_n}(Z_m) \end{pmatrix}.$$

The above equation then reads

$$F(X_2) = F(X_1) + L(X_2 - X_1). \tag{9-1}$$

This equation itself is sometimes referred to as the mean value theorem. Note, however, that the partial derivatives in L are *not* all evaluated at the same point.

Since $\left|\frac{\partial f_i}{\partial x_j}(X)\right| \le \gamma$ for all X, if η is any vector in $\mathbb{R}^n$, by Theorem 17, p. 373. we find that

$$\|L\eta\| \le \sqrt{nm}\gamma\|\eta\|.$$

Taking $\eta = X_2 - X_1$, and using (1), we are led to the inequality

$$\|F(X_2) - F(X_1)\| \le \sqrt{nm}\gamma\|X_2 - X_1\|,$$

which holds for any points X_1 and X_2 in D. With $C = \sqrt{nm}\gamma$, this is the desired inequality.

A few heuristic remarks. We have been considering mappings $F : \mathbb{R}^n \to \mathbb{R}^m$. In the case of linear mappings, $L : \mathbb{R}^n \to \mathbb{R}^m$, it was possible to prove that the range of L, $\mathcal{R}(L)$

had dimension no greater than n, that is, $\dim \mathcal{R}(L) \le n$. Although this does not remain true for an arbitrary nonlinear map F, it is still true if F is differentiable - after a suitable definition of dimension for an arbitrary point set is made (for the range of F will not usually be a linear space, the only sets whose dimension we have so far defined). In the case of differentiable maps F, it is easy to make a reasonable definition of dimension. The idea is to define dimension of the range of F locally, that is, in the neighborhood of every point in the range. If $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ and F is differentiable at $X \in D$, then for all h sufficiently small,

$$F(X+h) = F(X) + L_{(X)}h + \text{remainder}.$$

The *dimension of the range* of F at $F(X)$ is defined to be the dimension of its affine part, which is the same as $\dim \mathcal{R}(L_{(X)})$. Since $L_{(X)}$ is a linear operator, its range has a well defined dimension. Geometrically, we have defined dimension of the range of F at $F(X)$ as the dimension of the tangent plane at $F(X)$. Our definition makes good physical sense for it is exactly the number an insect on the surface would use for the dimension. The illustration below is for a map $F : D \quad \mathbb{R}^2 \to \mathbb{R}^3$ whose range has dimension 2,

A FIGURE GOES HERE

Some special remarks should be made about maps from one space into another of the *same* dimension,

$$F : D \quad \mathbb{R}^n \to \mathbb{R}^n.$$

Let us assume F is differentiable throughout D. Then the dimension of the range of F at $F(X)$, $X \in D$, is the dimension of the range of $L_{(X)} = F'(X)$. If F is to preserve dimension at every point, then we must have $\dim \mathcal{R}(L_{(X)}) = n$ for all $X \in D$. For maps F given in terms of coordinates, this means the determinant of the $n \times n$ matrix $L_{(X)}$ does not vanish,

$$\det L_{(X)} = \det F'(X) \ne 0$$

for all $x \in D$. In more conceptual terms, this states that a map $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ is dimension preserving at $X_0 \in D$ if its "affine part" $\Phi(X_0 + h) = F(X_0) + F'(X_0)h$ is dimension preserving at X_0 (there is no trouble with the constant vector $F(X_0)$ since it only represents a translation of the origin - which does not affect dimensionality).

From the geometric interpretation of determinants as volume, we see that the condition $\det F'(X_0) \ne 0$ means that if a small set $S \subset D$ has non-zero volume, then its image $F(X)$ also has non-zero volume. In fact, we expect that if S is a small set about X, then

$$\mathrm{Vol}\,(F(S)) = \left|\det F'(X_0)\right| \,\mathrm{Vol}\,(S).$$

Our expectation is based upon the realization that if the points of S are all near X_0, then F will behave like its affine part, $(X_0 + h) = F(X_0) + F'(X_0)h$, on the points $X_0 + h \in S$. The above formula is a restatement of the effect of affine maps on volume (Corollary to Theorem 30, page 426). We shall return to this later (Chapter 10, Section 4).

Because of its frequent appearance, $\det F'(X)$ has a name of its own. It is called the *Jacobian determinant* or just the *Jacobian* of F. If F is given in terms of coordinates,

$$\begin{aligned} y_1 &= f_1(x_1, \ldots, x_n) \\ &\cdot \\ &\cdot \\ &\cdot \\ y_n &= f_n(x_1, \ldots, x_n), \end{aligned}$$

then another common notation for the Jacobian is

$$\det F'(X) = \frac{\partial(f_1, f_2, \ldots, f_n)}{\partial(x_1, x_2, \ldots, x_n)}.$$

For these maps F from a space into one of the same dimension, $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$, there is a very special derivative which appears often. It is the sum of the diagonal elements of the derivative matrix $F'(X)$. One writes this expression as $\nabla.\ F$ or $\div F$, the *divergence* of F,

$$\nabla \cdot F(X) = \operatorname{div} F(X) = \frac{\partial f_1(X)}{\partial x_1} + \frac{\partial f_2(X)}{\partial x_2} + \cdots + \frac{\partial f_n(X)}{\partial x_n}$$

For example, if $Y = F(X)$ is defined by

$$y_1 = x_1 + 2x_1x_2$$

$$y_2 = x_1^2 - 3x_2,$$

then

$$F'(X) = \begin{pmatrix} 1 + 2x_2 & 2x_1 \\ 2x_1 & -3 \end{pmatrix}$$

and

$$\nabla \cdot F(X) = \operatorname{div} F(X) = (1 + 2x_2) + (-3) = -2 + 2x_2.$$

The significance of the divergence will become clear later (Chapter 10, Section 2). You will probably find it helpful to think of ∇ as the operator

$$\nabla = (\frac{\partial}{\partial x_1}, \cdots, \frac{\partial}{\partial x_n}).$$

Then $\nabla \cdot F$ is the "scalar product" of the operator ∇ with the vector F.

EXERCISES.

(1) (a) Find the derivative matrix at the given point for the following mappings $Y = F(X)$.

(i) $y_1 = x_1^2 + \sin x_1 x_2$
$y_2 = x_2^2 + \cos x_1 x_2$ at $X_0 = (0, 0)$

(ii) $y_1 = x_1^2 + x_3 e^{x_2} - x_2^3$
$y_2 = x_1 - 3x_2 + x_1 \log x_3$
$y_3 = x_2 + x_3$
$y_4 = 5x_1x_2x_3$ at $X_0 = (2, 0, 1)$

(b) Find the equation of the tangent plane to the above surfaces at the given point.

(2) Consider the following map from $\mathbb{R}^2 \to \mathbb{R}^2$,

$$\begin{cases} u = e^x \cos y \\ v = e^x \sin y \end{cases}$$

(a) Find the image of the following regions
i) $x \geq 0, \quad 0 \leq y \leq \frac{\pi}{4}$
ii) $x \geq 0, \quad 0 \leq y \leq \pi$
iii) $x \leq 0, \quad 0 \leq y \leq 2\pi$
iv) $1 < x < 2, \quad \frac{\pi}{6} \leq y \leq \frac{\pi}{3}$.

(b) Compute the derivative matrix and its determinant.

(3) If $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $X_0 \in D$, prove it is then also continuous at X_0.

(4) Let F and G both map $D \subset \mathbb{R}^n \to \mathbb{R}^m$, so the function $f(X) = \langle F(X), G(X) \rangle$ is defined for all $X \in D$ and $f : D \to \mathbb{R}^1$.

(a) If F and G are differentiable in D, prove f is also, and that

$$f' = F'G + G'F$$

(b) Apply this result to the function

$$f(X) = \langle X, AX \rangle - 2\langle X, Y \rangle,$$

where A is a constant linear operator from $\mathbb{R}^n \to \mathbb{R}^n$ and Y is a constant vector in $\mathbb{R}^n$. How does the result simplify if A is self adjoint?

(5) If $\varphi : D \subset \mathbb{R}^n \to \mathbb{R}^1$ and $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$, then the function $G(X) := \varphi(X)F(X)$ is defined for all $x \in D$ and $G : \mathbb{R}^n \to \mathbb{R}^m$.

(a) Let $\varphi(x_2, x_2) = ax_1 + bx_2$ and $F(x_1, x_2) = (\alpha x_1 + \beta x_2, \gamma x_1 + \delta x_2)$. Let $G = \varphi F$ and compute $G'(X)$.

(b) More generally, prove that if φ and F are differentiable in D, then $G := \varphi F$ is also differentiable and find a formula for G'. If F is expressed in terms of coordinate functions, $F = (f_1, f_2, \ldots, f_m)$, how does your formula read? Check the result with that of part (a).

(6) (a) If $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ is differentiable in the open connected set D, and if $F'(X) \equiv 0$ for all $x \in D$, prove that F is a constant vector.

(b) If F and G map $D \subset \mathbb{R}^n \to \mathbb{R}^m$ are differentiable in the open connected set D, and if $F'(X) \equiv G'(X)$ for all $x \in D$, what can you conclude?

(7) Consider the map $F : Q \to \mathbb{R}^3$ defined by

$$F : \begin{array}{l} x = (a + b\cos\varphi)\cos\theta \\ y = (a + b\cos\varphi)\sin\theta \\ z = b\sin\varphi \end{array}$$

A FIGURE GOES HERE

(a) Compute F'.

(b) Find the equation of the tangent map at $(0, 0)$ and at $(\pi/2, \pi/2)$.

(c) Determine the range of the tangent map at the above two points and indicate your findings in a sketch.

9.2 The Derivative of Composite Maps ("The Chain Rule").

Consider the two mappings

$$F : A \subset \mathbb{R}^n \to \mathbb{R}^m \quad \text{and} \quad G : B \subset \mathbb{R}^m \to \mathbb{R}^r.$$

Then the composite map $H := G \circ F : A \subset \mathbb{R}^n \to \mathbb{R}^r$ is defined if B contains the image of all the points from A, $F(A) \subset B$.

A FIGURE GOES HERE

The map $H = G \circ F$ takes points from $A \subset \mathbb{R}^n$ and sends them into $\mathbb{R}^r$. From knowledge of the derivatives of F and G, it is possible to compute the derivative of the composite map $G \circ F$.

Theorem 9.6 . *Let $F : A \subset \mathbb{R}^n \to \mathbb{R}^m$ and $G : B \subset \mathbb{R}^m \to \mathbb{R}^r$ be differentiable maps defined in the open sets A and B, respectively, with $F(A) \subset B$ (so the composite map $H(X) := (G \circ F)(X)$ is defined for all $X \in A$). If $X_0 \in A$, let $Y_0 = F(X_0) \in B$. Then the composite map H is differentiable at X_0 and*

$$H'(X_0) = G'(Y_0) \circ F'(X_0).$$

REMARK: The multiplication $G' \circ F'$ is the multiplication of the linear operators G' and F'. If F and G are given in terms of coordinates, then the formula is just the product of two matrices G' and F'.

Before proving this theorem, we shall illustrate its meaning.

EXAMPLE: Let $F : \mathbb{R}^2 \to \mathbb{R}^2$ and $G : \mathbb{R}^2 \to \mathbb{R}^3$ be defined by $Y = F(X)$ and $Z = G(Y)$ as follows

$$\begin{cases} y_1 = x_1 - x_2^2 \\ y_2 = x_2 \sin \pi x_1 \end{cases} \quad \begin{cases} z_1 = y_1 y_2 \\ z_2 = 1 + y_1^2 + y_2 \\ z_3 = 5 - y_2^3. \end{cases}$$

Then

$$F'(X) = \begin{pmatrix} 1 & -2x_2 \\ \pi x_2 \cos \pi x_1 & \sin \pi x_1 \end{pmatrix}, \quad G'(X) = \begin{pmatrix} y_2 & y_1 \\ 2y_1 & 1 \\ 0 & -3y_2^2 \end{pmatrix}.$$

At $X_0 = (3, 2)$, we find $Y_0 = F(X_0) = (-1, 0)$. Thus

$$F'(X_0) = \begin{pmatrix} 1 & -4 \\ -2\pi & 0 \end{pmatrix}, \quad G'(Y_0) = \begin{pmatrix} 0 & -1 \\ -2 & 1 \\ 0 & 0 \end{pmatrix}.$$

If $H(X) = (G \circ F)(X) = G(F(X))$, then the derivative of H at X_0 is

$$H'(X_0) = G'(Y_0) \circ F'(X_0) = \begin{pmatrix} 0 & -1 \\ -2 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & -4 \\ -2\pi & 0 \end{pmatrix} = \begin{pmatrix} 2\pi & 0 \\ -2 - 2\pi & 8 \\ 0 & 0 \end{pmatrix}.$$

The derivative could also have been found in a longer way by explicitly finding $Z = H(X)$ from the formulas for F and G

$$\begin{aligned} z_1 &= y_1 y_2 = (x_1 - x_2^2)(x_2 \sin \pi x_1) \\ z_2 &= 1 + y_1^2 + y_2 = 1 + (x_1 - x_2^2)^2 + x_2 \sin \pi x_1 \\ z_3 &= 5 - y_2^3 = 5 - (x_2 \sin \pi x_1)^3 \end{aligned}$$

and now directly computing $H'(X_0)$.

Proof of Theorem. Since F is differentiable at $X_0 \in A \subset \mathbb{R}^n$ and G is differentiable at $Y_0 \in B \subset \mathbb{R}^r$, for all sufficiently small vectors $h \in \mathbb{R}^n$ and $k \in \mathbb{R}^m$, we can write

$$F(X_0 + h) = F(X_0) + F'(X_0)h + R_1(X_0, h)\|h\|$$

$$G(Y_0 + k) = G(Y_0) + G'(Y_0)k + R_2(Y_0, k)\|k\|$$

where

$$\lim_{\|h\| \to 0} \|R_1(X_0; h)\| = 0 \quad \text{and} \quad \lim_{\|k\| \to 0} \|R_2(Y_0, k)\| = 0.$$

Consequently, since $H(X) := (G \circ F)(X) = G(F(X))$,

$$\begin{aligned} H(X_0 + h) &= G(F(X_0 + h)) \\ &= G(F(X_0) + F'(X_0)h + R_1(X_0; h)\|h\| \\ &= G(F(X_0)) + G'(Y_0)F'(X_0)h + R_3(X_0, h)\|h\|, \end{aligned}$$

where

$$R_3(X_0; h) = G'(Y_0)R_1(X_0; h) + \frac{R_2(Y_0, k)\|k\|}{\|h\|},$$

and

$$k = F'(X_0)h + R_1(X_0; h)\|h\|.$$

Thus, for all sufficiently small h,

$$H(X_0 + h) = H(X_0) + G'(Y_0)F'(X_0)h + R_3(X_0', h)\|h\|.$$

Because $G'(Y_0)$ and $F'(X_0)$ are linear maps, so is their product. Therefore we are done if we prove $\lim_{\|h\| \to 0} \|R_3(X_0; h)\| = 0$.

By the triangle inequality

$$\|R_3(X_0; h)\| \le \|G'(Y_0)R_1(X_0; h)\| + \frac{\|R_2(Y_0, k)\| \, \|k\|}{\|h\|}.$$

Since for fixed X_0, the operators $F'(X_0)$ and $G'(Y_0)$ are constant operators, by Theorem 17, p. 373, there exist constants α and β such that for any vectors $\xi \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^m$,

$$\|F'(X_0)\xi\| \le \alpha\|\xi\| \quad \text{and} \quad \|G'(Y_0)\eta\| \le \beta\|\eta\|.$$

This means

$$\|k\| \le \|F'(X_0)h\| + \|R_1(X_0;h)\|\,\|h\| \le (\alpha + \|R_1(X_0;h)\|)\|h\|$$

and

$$\|G'(Y_0)R_1(X_0;h)\| \le \beta\|R_1(X_0;h)\|.$$

Thus,

$$\|R_3(X_0;h)\| \le \beta\|R_1(X_0;h)\| + (\alpha + \|R_1(X_0;h)\|)\|R_2(Y_0,k)\|$$

Now, as $\|h\| \to 0$, so does $\|k\| \le (\alpha + \|R_1(X_0;h)\|)\|h\|$. From the definition of R_1 and R_2, this implies $\|R_3(X_0;h)\| \to 0$ as $\|h\| \to 0$ and completes the proof.

Incidentally, if one writes $Y = F(X)$ and $Z = G(Y)$, then the chain rule can be written in the form

$$\frac{d}{dx}(G \circ F) = \frac{dG}{dY} \circ \frac{dY}{dX},$$

which could hardly be more simple to remember.

For the balance of this section, we shall work out a few more illustrations showing how the chain rule is applied in different concrete situations. We isolate the next example as an important

Corollary 9.7 . *Let $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ and the scalar valued function $g : \mathbb{R}^m \to \mathbb{R}^1$ both satisfy the hypotheses of Theorem 1. If we write $Y = F(X)$ in coordinates $F = (f_1, f_2, \dots, f_m)$, and let $h = g \circ F$, then*

$$\begin{aligned} \frac{\partial h}{\partial x_1} &= \frac{\partial g}{\partial y_1}\frac{\partial f_1}{\partial x_1} + \frac{\partial g}{\partial y_2}\frac{\partial f_2}{\partial x_1} + \cdots + \frac{\partial g}{\partial y_m}\frac{\partial f_m}{\partial x_1} \\ &\cdot \\ &\cdot \\ &\cdot \\ \frac{\partial h}{\partial x_n} &= \frac{\partial g}{\partial y_1}\frac{\partial f_1}{\partial x_n} + \frac{\partial g}{\partial y_2}\frac{\partial f_2}{\partial x_n} + \cdots + \frac{\partial g}{\partial y_m}\frac{\partial f_m}{\partial x_n} \end{aligned}$$

REMARK: This is the chain rule for scalar-valued functions.
PROOF: By Theorem 3,

$$\frac{dh}{dX} = \frac{dq}{dY}\frac{dF}{dX}$$

Since

$$\frac{dq}{dY} = (\frac{\partial g}{\partial y_1}, \cdots, \frac{\partial g}{\partial y_m})$$

and

$$\frac{dF}{dX} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix},$$

we find upon multiplying the matrices that

$$\frac{dh}{dX} = (\sum_{j=1}^m \frac{\partial g}{\partial y_j}\frac{\partial f_j}{\partial x_1}, \quad \sum_{j=1}^m \frac{\partial g}{\partial y_j}\frac{\partial f_j}{\partial x_2}, \cdots, \sum_{j=1}^m \frac{\partial g}{\partial y_j}\frac{\partial f_j}{\partial x_n}).$$

But we also know

$$\frac{dh}{dX} = (\frac{\partial h}{\partial x_1}, \frac{\partial h}{\partial x_2}, \cdots \frac{\partial h}{\partial x_n}).$$

Comparison of the last two formulas gives the stated result.

EXAMPLE. Let $F : \mathbb{R}^2 \to \mathbb{R}^2$ and $g : \mathbb{R}^2 \to \mathbb{R}^1$ be defined by

$$\begin{cases} f_1(x_1,x_2) = x_1 - e^{x_2}, \\ f_2(x_1,x_2) = e^{x_1} + x_2 \end{cases} \quad g(y_1,y_2) = y_1^2 + y_1y_2.$$

Then

$$F'(X) = \begin{pmatrix} 1 & -e^{x_2} \\ e^{x_1} & 1 \end{pmatrix}, \quad g'(Y) = (2y_1 + y_2, y_1).$$

If $h = g \circ F = g(F(x_1,x_2))$, then

$$\begin{aligned} \frac{dh}{dX} &= (2y_1 + y_2, y_1)\begin{pmatrix} 1 & -e^{x_2} \\ e^{x_1} & 1 \end{pmatrix} \\ &= (2y_1 + y_2 + y_1e^{x_1}, -(2y_1 + y_2)e^{x_2} + y_1). \end{aligned}$$

In particular, we find

$$\frac{\partial h}{\partial x_1} = 2y_1 + y_2 + y_1e^{x_1}$$

and

$$\frac{\partial h}{\partial x_2} = -(2y_1 + y_2)e^{x_2} + y_1.$$

These formulas could also have been found by directly applying the corollary, viz.

$$\frac{\partial h}{\partial x_1} = \frac{\partial g}{\partial y_1}\frac{\partial f_1}{\partial x_1} + \frac{\partial g}{\partial y_2}\frac{\partial f_2}{\partial x_1} = (2y_1 + y_2)1 + y_1(e^{x_1}),$$

and similarly for $\partial h/\partial x_2$.

Many applications of the chain rule are more complicated. Consider a real valued function $g(x_1,x_2,x_3,t)$, which depends on the point $\tilde{X} = (x_1,x_2,x_3)$ as well as t. The function g could be an expression of the temperature at a point $\tilde{X}$ at time t. If the point $\tilde{X}$ represents your position in the room, then since you move around the room, $\tilde{X}$ is itself a function of t. Thus, if your position is specified by $\tilde{X} = \tilde{F}(t)$,

$$x_1 = f_1(t), \quad x_2 = f_2(t), \quad x_3 = f_3(t),$$

the temperature where you stand is $h(t) = g(f_1(t), f_2(t), f_3(t), t)$. Since $\tilde{F} : \mathbb{R}^1 \to \mathbb{R}^3$ while $g : \mathbb{R}^4 \to \mathbb{R}^1$, the chain rule is not directly applicable because g is defined on $\mathbb{R}^4$, while the image of $\tilde{F}$ is in $\mathbb{R}^3$.

A simple - if artificial - device clears up the difficulty. Introduce another variable x_4 and let $X = (x_1, x_2, x_3, x_4)$. Then write $g(x_1, x_2, x_3, x_4)$, as well as $X = F(t)$, with

$$x_1 = f_1(t), \quad x_2 = f_2(t), \quad x_3 = f_3(t), \quad x_4 = f_4(t) \equiv t.$$

Now, as before, $h(t) = g(f_1(t), f_2(t), f_3(t), t)$, but $F : \mathbb{R}^1 \to \mathbb{R}^4$ and $g : \mathbb{R}^4 \to \mathbb{R}^1$. The chain rule is thus applicable and gives

$$\frac{dh}{dt} = \frac{dg}{dX}\frac{dF}{dt}$$

$$= (\frac{\partial g}{\partial x_1}, \frac{\partial g}{\partial x_2}, \frac{\partial g}{\partial x_3}, \frac{\partial g}{\partial x_4} \begin{pmatrix} \frac{df_1}{dt} \\ \frac{df_2}{dt} \\ \frac{df_3}{dt} \\ 1 \end{pmatrix},$$

so that

$$\frac{dh}{dt} = \frac{\partial g}{\partial x_1}\frac{\partial f_1}{\partial t} + \frac{\partial g}{\partial x_2}\frac{\partial f_2}{\partial t} + \frac{\partial g}{\partial x_3}\frac{\partial f_3}{\partial t} + \frac{\partial g}{\partial x_4}.$$

Since $x_4 \equiv t$, the last equation can also be written as

$$\frac{dh}{dt} = \frac{\partial g}{\partial x_1}\frac{df_1}{dt} + \frac{\partial g}{\partial x_2}\frac{\partial f_2}{\partial t} + \frac{\partial g}{\partial x_3}\frac{df_3}{dt} + \frac{\partial g}{dt}.$$

From a less formal viewpoint, this could have been obtained directly from the equation $h(t) = g(f_1(t), f_2(t), f_3(t), t)$ without dragging in the artificial auxiliary variable x_4. The variable x_4 has been introduced to show how the chain rule applies. Once the process is understood, the variable x_4 can (and should) be omitted.

EXAMPLE. Let $g(x_1, x_2, x_3, t) = x_1 t + 3x_2^2 - x_1 x_3 + \frac{4}{1+t^2}$, and let $x_1 = 3t - 1$, $x_2 = e^{t-1}$, $x_3 = t^2 - 1$. If $h(t) = g(x_1(t), x_2(t), x_3(t), t)$, we find

$$\frac{dh}{dt} = \frac{\partial g}{\partial x_1}\frac{\partial x_1}{dt} + \frac{\partial g}{\partial x_2}\frac{dx_2}{\partial t} + \frac{\partial g}{\partial x_3}\frac{dx_3}{dt} + \frac{\partial g}{\partial t}.$$

$$= (t - x_3)3 + (6x_2)e^{t-1} - (x_1)2t + x_1 - \frac{8t}{(1+t^2)^2}.$$

In particular, at $t = 1$, we have $x_1 = 2$, $x_2 = 1$, $x_3 = 0$ so that

$$\frac{dh(1)}{dt} = (1-0)3 + (6)1 - (2)2 + 2 - \frac{8}{4} = 5.$$

It is straightforward to compute the second derivative d^2h/dt^2 from the formula for the first derivative.

$$\frac{d^2h}{dt^2} = \frac{\partial}{\partial x_1}(\frac{dg}{dt})\frac{dx_1}{dt} + \frac{\partial}{\partial x_2}(\frac{dg}{dt})\frac{dx_2}{dt} + \frac{\partial}{\partial x_3}(\frac{dg}{dt})\frac{dx_3}{dt} + \frac{\partial}{\partial t}(\frac{dg}{dt}).$$

For this example, this gives

$$\frac{d^2h}{dt^2} = (-2t+1)3 + (6e^{t-1})e^{t-1} + (-3)2t +$$

$$(3 + 6x_2e^{t-1} - 2x_1 - 8\frac{1-3t^2}{(1+t^2)^3}).$$

At $t = 1$, we have

$$\frac{\partial^2 h}{\partial t^2}(1) = (-2+1)3 + 6 - 6 + (3 + 6 - 4 - 8\frac{-2}{8}) = 4.$$

The next example brings to the surface an ambiguity in the notation $\frac{\partial}{\partial x}$ for partial derivatives. This ambiguity is often a source of great confusion. Consider a scalar valued function $g(x_1, x_2, t, s)$. If $x_1 = f_1(t)$ and $x_2 = f_2(t)$, then

$$h(t, s) = g(f_1(t),\ f_2(t), t, s)$$

depends on the two variables t and s. In order to see how h changes with respect to t, we regard s as being held fixed and use the previous example to find

$$\frac{\partial h}{dt} = \frac{\partial g}{\partial x_1}\frac{\partial f_1}{dt} + \frac{\partial g}{\partial x_2}\frac{\partial f_2}{\partial t} + \frac{\partial g}{\partial t}.$$

We were careful and realized that the functions $g(x_1, x_2, t, s)$, a function with four independent variables, and $h(t, s) := g(f_1(t),\ f_2(t),\ t, s)$, a function with only two independent variables, were *different* functions. The usual (occasionally confusing) approach is to be less careful and write

$$\frac{\partial g}{dt} = \frac{\partial g}{\partial x_1}\frac{\partial f_1}{dt} + \frac{\partial g}{\partial x_2}\frac{\partial f_2}{\partial t} + \frac{\partial g}{\partial t}.$$

In the above equation, the term $\partial g/\partial t$ on the right is the partial derivative of $g(x_1, x_2, t, s)$ with respect to t while thinking of all four variables x_1, x_2, t and s as being independent. On the other hand, the term $\partial g/\partial t$ on the left is the partial derivative of $g(f_1(t), f_2(t), t, s)$ as a function of two variables. After being spelled out like this, the formula does have a clear meaning - but this is not at all obvious from a glance. One might even be *mistakenly* tempted to cancel the terms $\partial g/\partial t$ from both sides of the equation.

It is often awkward to introduce a new name, as $h(t,s)$, for $g(f_1(t), f_2(t), t, s)$. Another unambiguous procedure is available: use the numerical subscript notation for the partial derivatives. Then $g_{,1}$ always refers to the partial derivative of g with respect to its first variable, $g_{,2}$ with respect to the second variable, etc. Thus, for the above example of $g(x_1, x_2, t, s)$ where $x_1 = f_1(t)$ and $x_2 = f_2(t)$, we have

$$\frac{\partial g}{\partial t} = g_{,1}\frac{df_1}{dt} + g_{,2}\frac{df_2}{dt} + g_{,3}.$$

This clearly distinguishes the two time derivatives $g_{,3}$ and $\partial g/\partial t$.

The seemingly unnecessary comma in the notation is to take care of the possibility of vector valued functions $G(x_1, x_2, t, s)$ whose coordinate functions are indicated by subscripts. For example, if $G = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}$ is a map into $\mathbb{R}^2$, where the coordinate functions are $g_1(x_1, x_2, t, s)$ and $g_2(x_1, x_2, t, s)$, then if $x_1 = f_1(t)$ and $x_2 = f_2(t)$, we have

$$\frac{\partial G}{\partial t} = \begin{pmatrix} \frac{\partial g_1}{\partial t} \\ \frac{\partial g_2}{\partial t} \end{pmatrix} = \begin{pmatrix} g_{1,1}f_1' + g_{1,2}f_2' + g_{1,3} \\ g_{2,1}f_1' + g_{2,2}f_2' + g_{1,3} \end{pmatrix}.$$

Here $g_{1,1} = \partial g_1/\partial x_1$, etc. The notation f_1' for $df_1(t)/dt$ could also have been replaced by $f_{1,1}$—but this is unnecessary here since the f_j are functions of one variable.

In applications, one commonly meets a problem of the following type. Let $u(x,y)$ be a scalar valued function which satisfies the wave equation $u_{xx} - u_{yy} = 0$. If $F : \mathbb{R}^2 \to \mathbb{R}^2$ is defined by

$$x = f_1(\xi, \eta) = \frac{1}{2}(\xi, +\eta)$$

$$y = f_2(\xi, \eta) = \frac{1}{2}(\xi - \eta)$$

and if $h = u \circ F$, that is, $h(\xi, \eta) = u(f_1(\xi, \eta), f_2(\xi, \eta))$, what differential equation does h satisfy? First, we compute h_ξ and h_η

$$\frac{\partial h}{\partial \xi} = \frac{\partial u}{\partial x}\frac{\partial f_1}{\partial \xi} + \frac{\partial u}{\partial y}\frac{\partial f_2}{\partial \xi} = u_x(\frac{1}{2}) + u_y(\frac{1}{2}) = \frac{1}{2}(u_x + u_y)$$

$$\frac{\partial h}{\partial \eta} = \frac{\partial u}{\partial x}\frac{\partial f_1}{\partial \eta} + \frac{\partial u}{\partial y}\frac{\partial f_2}{\partial \eta} = u_x(\frac{1}{2}) + u_y(-\frac{1}{2}) = \frac{1}{2}(u_x - u_y)$$

In a similar way the second derivatives $h_{\xi\xi}, h_{\xi\eta}$ and $h_{\eta\eta}$ are found,

$$\frac{\partial^2 h}{\partial \xi^2} = \frac{\partial(h_\xi)}{\partial x}\frac{\partial f_1}{\partial \xi} + \frac{\partial(h_\xi)}{\partial y}\frac{\partial f_2}{\partial \xi}$$

$$= \frac{1}{2}\frac{\partial}{\partial x}(u_x + u_y)\frac{1}{2} + \frac{1}{2}\frac{\partial}{\partial y}(u_x + u_y)\cdot\frac{1}{2} = \frac{1}{4}[u_{xx} + 2u_{xy} + u_{yy}]$$

$$\frac{\partial^2 h}{\partial \xi \partial \eta} = \frac{\partial(h_\xi)}{\partial \eta} = \frac{\partial(h\xi)}{\partial x}\frac{\partial f_1}{\partial \eta} + \frac{\partial(h_\xi)}{\partial y}\frac{\partial f_2}{\partial \eta}$$

$$= \frac{1}{2}\frac{\partial}{\partial x}(u_x + u_y)\cdot\frac{1}{2} + \frac{1}{2}\frac{\partial}{\partial y}(u_x + u_y)\cdot\frac{-1}{2} = \frac{1}{4}[u_{xx} - u_{yy}]$$

$$\frac{\partial^2 h}{\partial \eta^2} = \frac{\partial(h_\eta)}{\partial x}\frac{\partial f_1}{\partial \eta} + \frac{\partial(h_\eta)}{\partial y}\frac{\partial f_2}{\partial \eta}$$

$$= \frac{1}{2}\frac{\partial}{\partial x}(u_x - u_y)\cdot\frac{1}{2} + \frac{1}{2}\frac{\partial}{\partial y}(u_x - u_y)(-\frac{1}{2}) = \frac{1}{4}[u_{xx} - 2u_{xy} + u_{yy}]$$

Since $h_{\xi\eta} = \frac{1}{4}[u_{xx} - u_{yy}]$, and u satisfies the wave equation, we see that h satisfies the equation

$$h_{\xi\eta} = 0,$$

so, in fact, the equations for $h_{xi\xi}$ and $h_{\eta\eta}$ are superfluous to obtain the desired result.

From this, it is easy to give another procedure for solving the wave equation, independent of Fourier series. Because $h_{\xi\eta} = 0$, we know that $h(\xi,\eta) = \varphi(\xi) + \psi(\eta)$, where the functions φ and ψ are any twice differentiable functions. However, $h(\xi,\eta) = u(\frac{\xi+\eta}{2}, \frac{\xi-\eta}{2})$. Since the equations $x = \frac{\xi+\eta}{2}$, $y = \frac{\xi-\eta}{2}$ may be solved for ξ and η in terms of x and y, viz. $\xi = x+y$ and $\eta = x-y$, we have $h(x+y, x-y) = u(x,y)$. But $h(\xi,\eta) = \varphi(\xi) + \psi(\eta)$. Consequently

$$u(x,y) = \varphi(x+y) + \psi(x-y).$$

This formula is the *general solution* of the one space dimensional wave equation. It expresses u in terms of two arbitrary functions φ and ψ.

These functions φ and ψ can be chosen so that the function $u(x,y)$, a solution of the wave equation, has any given initial position $u(x,0) = f(x)$ and initial velocity $u_y(x,0) = g(x)$. Let us do this.

From the initial conditions we find

$$f(x) = u(x,0) = \varphi(x) + \psi(x)$$

$$g(x) = u_y(x,0) = \varphi'(x) - \psi'(x).$$

After differentiating the first expression, one can solve for φ' and ψ',

$$\varphi'(x) = \frac{f'(x) + g(x)}{2}, \quad \psi'(x) = \frac{f'(x) - g(x)}{2}.$$

Integrate these:

$$\varphi(x) = \varphi(0) + \int_0^x \frac{f'(s) + g(s)}{2}\,ds = \varphi(0) + \frac{f(x) - f(0)}{2} + \frac{1}{2}\int_0^x g(s)\,ds.$$

$$\psi(x) = \psi(0) + \int_0^x \frac{f'(s) + g(s)}{2}\,ds = \psi(0) + \frac{f(x) - f(0)}{2} + \frac{1}{2}\int_0^x g(s)\,ds.$$

Thus,

$$u(x,y) = \varphi(x+y) + \psi(x-y) = \varphi(0) + \frac{f(x+y) - f(0)}{2} + \frac{1}{2}\int_0^{x+y} g(s)\,ds +$$

$$\psi(0) + \frac{f(x-y)-f(0)}{2} - \frac{1}{2}\, int_0^{x-y} g(s)\, ds.$$

Because $f(0) = \varphi(0) + \psi(0)$, this simplifies to

$$u(x,y) = \frac{f(x+y)-f(x-y)}{2s} + \frac{1}{2}\int_{x-y}^{x+y} g(s)\, ds,$$

the famous *d'Alembert formula* for the solution of the one space dimensional wave equation in terms of the initial position $f(x)$ and initial velocity $g(x)$. Unfortunately, simple formulas like this are exceedingly rare. That is why a different, more generally applicable, procedure was used earlier to solve the wave equation. As was seen in Exercise 6, p. 645, the d'Alembert formula is recoverable from the Fourier series.

Exercises

(1) For the following function g and f, compute $\frac{d}{dX}(g \circ F)$ and evaluate $\frac{\partial}{\partial x_1}(g \circ F)$ at the point $X_0 = (2,2)$.

(a) $g(y_1,y_2) = y_1y_2 - y_2e^{2y_1}$,

$$F: y_z = 2x_1 - x_1x_2, \quad y_2 = x_1^2 + x_2^2$$

(b) $g(y_1,y_2) = 7 + e^{y_1}\sin y_2$

$$F: y_1 = 2x_1x_2, \quad y_2 = x_1^2 - x_2^2$$

(c) $g(y_1,y_2,y_3) = y_1^2 - y_2^2 - 3y_1y_3 + y_2$

$$F: y_1 = 2x_1 - x_2, \quad y_2 = 2x_1 + x_2, \quad y_3 = x_1^2$$

(2) Let $\varphi(x_1,x_2,t) := x_2x_2 - te^{2x_1}$. If $X = F(t)$ is defined by $x_1 = 1-t^2, \quad x_2 = 3t+1$, find $\frac{d}{dt}(\varphi \circ F)$ at $t = 1$.

(3) Let $\varphi(x,s,t) := xs + xt + st$. If $x = f(t) = t^3 - 7$, compute $\frac{\partial}{\partial t}(\varphi \circ f)$ at $t = 3$. Also compute $\frac{\partial^2}{\partial t^2}(\varphi \circ f)$ at $t = 3$.

(4) If $u(x,y) = x^2 - y^2$, while $F := (f_1, f_x)$ is given by $x = f_1(r,\theta) = r\cos\theta, \quad y = f_2(r,\theta) = r\sin\theta$ find h_r and h_θ, where $h := u \circ F$. Also compute, $h_{rr}, \quad h_{r\theta}$ and $h_{\theta\theta}$.

(5) (a) Let $u(x,y)$ be a scalar valued function and $F: \mathbb{R}^2 \to \mathbb{R}^2$ be defined by the polar coordinate transformation

$$f_1(r,0) = r\cos\theta, \quad f_2(r\theta) = r\sin\theta,$$

Take $h := u \circ F$. Find $h_r, h_\theta, h_{rr}, h_{r\theta}$, and $h_{\theta,\theta}$. [Answer: $h_4 = u_x\cos\theta + u_y\sin\theta, \quad h_{rr} = -u_{xx}r\sin\theta + u_{yy}(r\cos\theta - r\sin\theta) + u_{yy}r\cos\theta - u_x\sin\theta + u_y\cos\theta$]

(b) Show that

$$u_{xx} + u_{yy} = h_{rr} + \frac{1}{r^2}h_{\theta\theta} + \frac{1}{r}h_r.$$

(6) The two space dimensional wave equation is

$$u_{tt} = u_{xx} + u_{yy}$$

(a) If the space variables x, y are changed to polar coordinates (ex. 5) while the time variable is not changed, the wave equation reads

$$h_{tt} = ?$$

where $h(r,\theta,t) = u(r\cos\theta, r\sin\theta, t)$.

(b) If a given wave form depends only on the distance r from the origin and time t, but not on the angle ∂, how does the wave equation for h simplify?

(c) Consider the equation you found in b. Use the method of separation of variables and seek a solution in the form $h(r,t) = R(r)T(t)$. What are the resulting ordinary differential equations? Compare the equation for $R(r)$ with Bessel's differential equation.

(7) If $w = f(x,y,s)$, while $x = \varphi(y,s,t)$ and $y = \psi(s,t)$, find expressions for the partial derivative of the composite function $g(\varphi(\psi,s,t),\psi s)$ with respect to s and t.

(8) (a) Let $u(x,y) = f(x-y)$. Show that u satisfies the partial differential equation

$$u_x + u_y = 0.$$

(b) Let $u(x,y) = f(xy)$. Show that u satisfies the equation $xu_x - yu_y = 0$.

(c) Let $u(x,y) = f(\frac{x}{y})$. Show that u satisfies the equation

$$xu_x + yu_y = 0.$$

(d) Let $u(x,y) = f(x^2+y^2)$, so u only depends on the distance from the origin. Show that u satisfies

$$yu_x - xu_y = 0.$$

(9) Let $u(x,y)$ satisfy the equation $xu_x + yu_y = 0$.

(a) Change the equation to polar coordinates [Answer: if $h(r,\theta) := u(r\cos\theta, r\sin\theta)$, then $rh_r = 0$].

(b) Solve the equation for $h(r,\theta)$ and use it to deduce that $u(x,y) = f(\frac{x}{y})$ for some function f. (cf. Ex. 8c)

(10) Assume $u(x,y)$ satisfies the equation

$$u_{xx} - 2u_{xy} - 3u_{yy} = 0.$$

(a) Choose the constants α, β, γ, and δ so that after the change of variables $x = \alpha\xi + \beta\eta$, $\quad y = \gamma\xi + \delta\eta$, the equation for $h(\xi, \eta) = u(\alpha\xi + \beta\eta, \gamma\xi + \delta\eta)$ is $h_{\xi\eta} = 0$.

(b) Use the result of part (a) to find the general solution of the equation for u. [Answer: $u(x, y) = \varphi(3x - y) + \psi(x + y)$].

(11) If $f(x, y)$ is a known scalar valued function, find both partial derivatives of the function $f(f(x, y), y)$.

(12) If $W = G(Y)$ and $Y = F(X)$ are defined by

$$G : \begin{cases} w_1 = e^{y_1 - y_2} \\ w_2 = e^{y_1 + y_2} \end{cases}, \quad F : \begin{cases} y_1 = x_1^2 - 3x_2 - x_3 \\ y_2 = x_1 + x_2^2 + 3x_3, \end{cases}$$

find $\frac{d}{dX}(G \circ F)$.

(13) Let $u(x, y)$ be a solution of the two dimensional Laplace's equation $u_{xx} + u_{yy} = 0$.

(a) If u depends only on the distance from the origin $u(x, y) = h(r)$, where $r = x^2 + y^2$, what ordinary differential equation does h satisfy? Compare your answer with that found in Exercise 5.

(b) Solve the resulting equation for h and deduce that all the solutions of the two dimensional Laplace equation which depend only on the distance from the origin are of the form

$$u(x, y) = A + B \log(x^2 + y^2),$$

where A and B are constants.

(c) Now do the same thing all over again for a solution $u(x_1, x_2, \ldots, x_n)$ of the n dimensional Laplace equation $u_{x_1x_1} + \ldots + u_{x_nx_n} = 0$, i.e. find the form of all solutions which only depend on $r = \sqrt{x_1^2 + \ldots + x_n^2}$, $u(x_1, \ldots, x_n) = h(r)$. [Answer: $u(x_1, \ldots, x_n) = A + \frac{B}{(x_1^2 + \ldots + x_n^2)}\frac{n-2}{2} = A + \frac{B}{r^{n-2}}, \quad n \geq 3$].

(14) If $f(t)$ is a differentiable scalar valued function with the property that $f(x + y) = f(x) + f(y)$ for all $x, y \in \mathbb{R}^1$, prove that $f(x) \equiv kx$ where $k = f(1)$.

(15) (a) Find the general solution of the partial differential equation $u_x - 2u_y = 0$. [Hint: Introduce new variables as in Ex. 10]

(b) What is the solution if one requires that $u(x, 0) = x^2$? [Answer: $u(x, y) = (x + \frac{1}{2}y)^2$].

Chapter 10

Miscellaneous Supplementary Problems

1. (a) S_n, $n = 1, 2, \ldots$, be a given sequence. Find another sequence a_n such that $S_N = \sum_{n=1}^{N} a_n$. In other words, given the partial sums S_n, find a series whose partial sums are S_n. To what extent are the a_n uniquely determined?

   (b) Apply part (a) to find an infinite series $\sum a_n$ whose nth partial sum S_n is given by

$$(i) \quad S_n = \frac{1}{n}, \qquad (ii) \quad S_n = e^{-n}$$

2. Let $S = \{x \in \mathbb{R} : x \in (-1, 1)\}$. Define addition on S by the formula $x \oplus y = \frac{x+y}{1+xy}$, $x, y \in S$, where the operations on the right are the usual ones of arithmetic. Show that the elements of S form a commutative group with the operation $\oplus$.

3. (a) If $a_n \to a$, prove that $\frac{a_1 + a_2 + \cdots + a_n}{n} \to a$ also.

   (b) Assume that f is continuous on the interval $[0, \infty]$ and $\lim_{x\to\infty} f(x) = A$. Define $H_N = \frac{1}{N}\int_0^N f(x)\, dx$. Prove that $\lim_{x\to\infty} H_N$ exists and find its value. [Hint: Interpret H_N as the average height of the function f].

4. (a) Suppose that *all* the zeroes of a polynomial $P(x)$ are real. Does this imply that all the zeroes of its derivative $P'(x)$ are also real? (Proof or counterexample). What can you say about higher derivatives $P^{(k)}(x)$?

   (b) Define the nth Laguerre polynomial by

$$L_n(x) = e^x \frac{d^n}{dx^n}[x^n e^{-x}].$$

Show that L_n is a polynomial of degree n. Prove that the zeroes of $L_n(x)$ are all positive real numbers, and that there are exactly n of them.

5. If $f(x)$ has a Taylor series: $f(x) = \displaystyle\sum_{n=0}^{\infty} a_n x^n$ (which converges to f for $|x| < \rho$ so the remainder does go to zero there) prove that $f(cx^k)$, where c is a constant and k a positive integer, has the Taylor series

$$f(cx^k) = \sum_{n=0}^{\infty} a_n c^n x^{nk}$$

which converges to $f(cx^k)$ for $|x| < (\frac{\rho}{|c|})^{1/k}$. You must show that i) the Taylor coefficients for $f(cx^k)$ are $a_n c^n$, that ii) the power series for $f(cx^k)$ converges for $|x| < (\frac{\rho}{|c|})^{1/k}$, and that iii) the remainder tends to zero. Apply the result to obtain the Taylor series for $\cos(2x^2)$ from that of $\cos x$.

6. Yet another proof of Taylor's Theorem. Beginning with equation 9 on p. 98, define the function $K(s)$ by

$$K(s) = f(s) - \sum_{n=0}^{N} \frac{f^{(n)}(x_0)}{n!}(s - x_0)^n - \frac{A(s - x_0)^{N+1}}{(N+1)!},$$

where A is picked so that $K(\hat{x}) = 0$.

   (a) Verify that $K(x_0) = K'(x_0) = \ldots K^{(N)}(x_0) = 0$.

   (b) Use Rolle's Theorem to prove that if a function $K(s)$ satisfies the properties of a), and if $K(\hat{x}) = 0$, then there is a ξ between $\hat{x}$ and x_0 such that $K^{(N+1)}(\xi) = 0$.

   (c) Apply parts a) and b) to prove Taylor's Theorem.

7. Assume $\sum a_n$ converges. You are to investigate the convergence of $\sum a_n^2$ and $\sum \sqrt{|a_n|}$ under various hypotheses.

   (a) a_n arbitrary complex number

   (b) $a_n \geq 0$.

   (c) $\displaystyle\lim_{n\to\infty} \left| \frac{a_n + 1}{a_n} \right| < 1$ (*not* $= 1$).

8. The harmonic series $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ has been said to diverge with "infuriating slowness". Find a number N such that $1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{N}$ is at least 100. Compare this with Avogadro's number $\sim 6 \times 10^{23}$.

9. Consider the series $\sum_{n=1}^{\infty} a_n$, where the a_n's are real.

   (a) Let $b_1, b_2, b_3, \ldots$ and $c_1, c_2, c_3, \ldots$ denote the positive and negative terms respectively from $a_1, a_2, \ldots$. If $\sum_{n=1}^{\infty} a_n$ converges conditionally but not absolutely, prove that both series $\sum_{n=1}^{\infty} b_n$ and $\sum_{n=1}^{\infty} c_n$ *diverge*.

   (b) Let $d_1, d_2, d_3, \ldots$, denote the terms $a_1, a_2, a_3, \ldots$ rearranged in any way. Prove Riemann's theorem, which states that if $\sum_{n=1}^{\infty} a_n$ converges conditionally but not absolutely, then by picking some suitable rearrangement, the series $\sum_{n=1}^{\infty} d_n$ can be made to converge to any real number, while using other rearrangements, it can be made to diverge to plus or minus infinity.

10. If A and B are subsets of a linear space V, a) show that $\text{span}\{\, A \cap B \,\} \subset \text{span}\{\, A \,\} \cap \text{span}\{\, B \,\}$. Give an example showing that $\text{span}\{\, A \cap B \,\}$ may be smaller than $\text{span}\{\, A \,\} \cap \text{span}\{\, B \,\}$.

    b). Show that if $A \subset B \subset \text{span}\{\, A \,\}$, then $\text{span}\{\, A \,\} \supset \text{span}\{\, B \,\}$.

11. Let $A = \{\, X_1, \ldots, X_k \,\}$ be a set of vectors in a linear space V. Denote by cs A (coset of A) the set

$$csA = \{\, X \in V \colon X = \sum_{j=1}^{k} a_j X_j, \quad \text{where} \quad \sum_{j=1}^{k} a_j = 1 \,\}.$$

Prove that cs A is a coset of V, in fact, the smallest coset of V which contains the vectors $X_1, \ldots, X_k$.

12. (a) Consider the set of real numbers of the form $a + b\sqrt{2}$, where a and b are rational numbers. Prove that this set is a vector space over the field of *rational* numbers. What is the dimension of this vector space?

   (b) Consider the set of numbers of the form $a + bi$, where a and b are real numbers and $i = \sqrt{-1}$. Prove that this set is a vector space over the field of *real* numbers and find its dimension.

13. If F_1 and F_2 are fields with $F_1 \subset F_2$, we call F_2 an *extension field* of F_1 – such as $\mathbb{R} \subset \mathbb{C}$. As such, we may think of F_2 as a vector space over the field F_1 (see exercise 11). In other words, take F_2 as an additive group and take the scalars from F_1. If this vector space is finite dimensional, the field F_2 is called a *finite extension* of F_1, and the dimension n of this vector space is called the *degree of the extension* and written $n = [F_2 : F_1]$.

   (a) Prove that every element $\xi \in F_2$ satisfies an equation

$$a_n \xi^n + a_{n-1} \xi^{n-1} + \cdots + a_0 = 0,$$

   where the $a_k \in F_1$ and $n = [F_2 : F_1]$. [Hint: look at the examples of exercise 11].

(b) If $F_1 \subset F_2 \subset F_3$ are fields with

$$[F_2 : F_1] = n < \infty \quad \text{and} \quad [F_3 : F_2] = m < \infty,$$

prove that $[F_3 : F_1] < \infty$, in fact, prove

$$[F_3 : F_1] = [F_3 : F_2]]F_2 : F_1] = nm.$$

(c) Let F_1 be the field of rationals, F_2 the field whose elements have the form $a + b\sqrt{3}$, where a and b are rational, and let F_3 be the field whose elements have the form $c + d\sqrt{5}$, where c and d are in F_2. Compute $[F_2 : F_1]$ and find the polynomial of part a) satisfied by $(1 - \sqrt{3}) \in F)_2$. Compute $[F_3 : F_2]$ and $[F_3 : F_1]$. Find a basis for F_3 as a vector space whose scalars are elements of F_1. [The ideas in this problem are basic to modern algebra, particularly Galois' theory of equations.]

14. Let $P_j = (\alpha_j, \beta_j)$, $j = 1, \ldots, n, \alpha_j \neq \alpha_k$ be any n distinct points in the plane $\mathbb{R}^2$. One often wants to find a polynomial $p(x) = a_0 + a_1 x + \cdots + a_N x^N$ which passes through these n points, $p(\alpha_j) = \beta_j$, $j = 1, \ldots, n$. Thus, $p(x)$ is an *interpolating polynomial*. Given any points $P_1, \ldots, P_n$, prove that a unique interpolating polynomial $p(x)$ degree $n - 1 (= N)$ can be found. (More about this is in Exercises 17-18 below).

15. Let L_1 and L_2 be linear operators mapping $V \to V$. Then they can be both multiplied and added (or subtracted). The *bracket product* or *commutator*

$$[L_1, L_2] \equiv L_1 L_2 - L_2 L_1$$

"measures the non-commutativity". It is important in mathematics and physics. [In quantum mechanics, the observables - like energy, momentum, and position - are represented by self-adjoint operators. Two observables can be measured at the same time if and only if their associated operators commute]. Prove the identities

(a) $[L_1, L_1] = 0$, $[L_1, I] = 0$
(b) $[L_1, L_2] = -[L_2, L_1]$
(c) $[aL_1, L_2] = a[L_1, L_2]$, a scalar
(d) $[L_1 + L_2, L_3] = [L_1, L_3] + [L_2, L_3]$
(e) $[L_1, L_2, L_3] = [L_1, L_2]L_3 + L_2[L_1, L_3]$
(f) $[L_1, [L_2, L_3]] + [L_2, [L_3, L_1]] + [L_3, [L_1, L_2]] = 0$

(Part f is the *Jacobi identity.* It has been said that everyone should verify it once in her lifetime.)

16. * Consider the normalized Legendre Polynomials,

$$e_n(x) = \sqrt{\frac{2}{2n+1}} \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n, \qquad n = 0, 1, 2, \ldots$$

which are an orthonormal set of polynomials in $L_2[-1, 1]$, e_n being of degree n. If $f \in C[-1, 1]$, prove that

$$P_N f = \sum_{n=0}^{N} \langle f, e_n \rangle e_n$$

converges to f in the norm of $L_2[-1, 1]$. [Hint: Use the form of the Weierstrass Approximation Theorem (p. 255) and the method of Theorem (p. 241)].

17. * We again take up the interpolation problem begun in Exercise 13 above. Let $P_j = (\alpha_j, \beta_j)$, $j = 1, 2, \ldots, n$ be n points in the plane, $\alpha_i \neq \alpha_j$. Although we proved there is a unique polynomial $p(x) = a_0 + a_1 x + \cdots + a_{n-1} x^{n-1}$ of degree $n - 1$ passing through the n points, the proof was entirely non-constructive. Here we (or you) explicitly construct the polynomial.

(a) Show that the polynomial of degree $n - 1$

$$\tilde{p}_j(x) = \Pi^n_{\substack{k=\ell \\ k \neq j}} (x - \alpha_k)$$

is zero if $x = \alpha_k$, $k \neq j$, but $\tilde{p}_j(\alpha_j) \neq 0$.

(b) Construct a polynomial $p_j(x)$ with the property $p_j(\alpha_k) = \delta_{jk}$.

(c) Show that

$$p(x) = \sum_{j=1}^{n} \beta_j p_j(x)$$

is the desired (unique by Ex. 13) interpolating polynomial.

(d) Let $P_1 = (1, 1)$, $P_2 = (2, 1)$, $P_3 = (4, -1)$, $P_4 = (-1, -2)$.

Find the interpolating polynomial using the above construction.

18. * If f is some complicated function, it is often useful to use an interpolating polynomial instead of the function. Then the polynomial $p(x)$ will pass through the points $P_j = (\alpha_m, f(\alpha_j))$, $\quad j = 1, \ldots, n$, so by Exercise 16,

$$p(x) = \sum_{j=1}^{n} f(\alpha_j) p_j(x).$$

] How much will p differ from f in an interval $[a, b]$ containing the α_j? You must estimate the remainder $R = f - p$.

(a) Assume $f \in C^n[a,b]$. Since $R(x) = f(x) - p(x)$ vanishes at $x = \alpha_j$, $j = 1, \ldots, n$, it is reasonable to write

$$R(x) = (x - \alpha_1) \cdots (x - \alpha_n) \cdot (?)$$

Fix $\hat{x}$ and define the constant A by

$$f(\hat{x}) - p(\hat{x}) = A(\hat{x} - \alpha_1) \cdots (\hat{x} - \alpha_n).$$

By a trick similar to that used in Taylor's Theorem (cf. P. 104j Ex. 12), prove that $A = f^{(n)}(\xi)/n!$ where ξ is some point in (a,b). Thus,

$$f(\hat{x}) = p(\hat{x}) + \frac{(\hat{x} - \alpha_1) \cdots (\hat{x} - \alpha_n)}{n!} f^{(n-1)}(\xi),\ \xi \in (a,b).$$

(b) Let $f(x) = 2^x$, and $\alpha_1 = -1$, $\alpha_2 = 0$, $\alpha_3 = 1$, $\alpha_4 = 2$.
Find the approximating polynomial and find an *upper bound* for the error in the interval $[-2, 2]$.

19. If x is irrational and a,b,c, and d are rational (with $ad - bc \neq 0$), prove that $\frac{ax+b}{cx+d}$ is irrational.

20. Prove by induction that

$$1 + 3 + 5 + \cdots + (2n-1) = n^2.$$

21. (a) If $x \geq 0$, use the mean value theorem to prove

$$e^x \geq 1 + x.$$

(b) If $a_k \geq 0$, prove that

$$\sum_{k=1}^{n} a_k \leq \Pi_{k=1}^{n}(1 + a_k) \leq e^{\sum_{k=1}^{n} a_k},$$

(where $\Pi_{k=1}^{n} b_k = b_1 b_2 \cdots b_n$).

(c) If $a_k \geq 0$, prove that the *infinite product* $\Pi_{k=1}^{\infty}(1 + a_k) := \lim_{n\to\infty} \Pi_{k=1}^{n}(1 + a_k)$ converges if and only if the infinite series $\sum_{k=1}^{\infty}$ converges.

22. Let $a_{n+1} = \frac{2}{1+a_n}$, where $a_1 > 1$. Prove that

(a) the sequence a_{2n+1} is monotone decreasing and bounded from below.

(b) the sequence a_{2n} is monotone increasing and bounded from above.

(c) does $\lim\limits_{n\to\infty} a_n$ exist?

23. Let a_k, $k = 1, \ldots, n+1$ be arbitrary real numbers which satisfy $a_1 + \frac{a_2}{2} + \cdots + \frac{a_n}{n} + \frac{a_{n+1}}{n+1} = 0$. Show that $P(x) = a_1 + a_2 x + \cdots + a_n x^{n-1}$ has at least one zero for $x \in (0,1)$.

24. Suppose $f \in C^2$ in some neighborhood of x_0. Prove that

$$\lim_{h\to 0} \frac{f(x_0 + h) - 2f(x_0) + f(x_0 - h)}{h^2} = f''(x_0).$$

25. Let $s(x)$ and $c(x)$ be continuously differentiable functions defined for all x, and having the properties

$$s'(x) = c(x), \qquad c'(x) = s(x)$$
$$s(0) = 0, \qquad c(0) = 1.$$

(a) Prove that $c^2(x) - s^2(x) = 1$.

(b) Show that $c(s)$ and $s(x)$ are uniquely determined by these properties.

26. Consider $\sum a_n$ and $\sum b_n$.

(a) If $\lim\limits_{n\to\infty} \left|\frac{b_n}{a_n}\right| = K$, $K \neq 0, \infty$, then the series both converge or diverge together.

(b) If $\sum a_n$ converges and $\lim\limits_{n\to\infty} \left|\frac{b_n}{a_n}\right| = 0$, then $\sum b_n$ converges.

(c) If $\sum a_n$ converges and $\lim\limits_{n\to\infty} \left|\frac{b_n}{a_n}\right| = \infty$, then the series $\sum b_n$ may converge or diverge (give examples).

(d) Apply these to:

(i) $\displaystyle\sum_{n=2}^{\infty} \frac{1}{n - \sqrt{n}}$

(ii) $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^3 - 2\sqrt{n}}$

(iii) $\displaystyle\sum_{n=1}^{\infty} (-1)^n \sin\frac{\pi}{n}$. (Hint: as $x \to 0$, $\frac{\sin x}{x} \to 1$).

27. The following (a weak form of *Stirling's formula*) is an improvement of the result on page 64, Ex. 6.

$$n \log n - (n-1) < \log n! < (n+1)\log(n+1) - 2\log 2 - (n-1),$$

from which one finds

$$n^n e^{-n+1} < n! < \tfrac{1}{4}(n+1)^{(n+1)} e^{-n+1}.$$

Prove these.

28. (a) Find the Taylor series expansion for $f(x) = e^{-x}$ about $x = 0$.
    (b) Show that the series found in (a) converges to e^{-x} for all x in the interval $[-r, r]$, where $r > 0$ is an arbitrary but fixed real number.

29. Consider the sequence
$$S_N = \int_2^N \frac{\sin \pi x}{x}\, dx.$$
Does $\lim_{N\to\infty} S_N$ exist? [Hint: observe that S_N can be written as
$$S_N = \sum_2^{N-1} a_n,$$
where
$$a_n = \int_n^{n+1} \frac{\sin \pi x}{x}\, dx.$$
Sketch a graph of $\frac{\sin \pi x}{x}, x \geq 2$, to deduce - by inspection - the needed properties of the a_n's. Please do not attempt to evaluate the integrals for a_n].

30. Let $A = \{\, p \in \mathcal{P}_9 \colon p(x) = p(-x)\,\}$.
    (a) Prove that A is a subspace of $\mathcal{P}_9$.
    (b) Compute the dimension of A.

31. Let X and Y be elements in a real linear space. Prove that $\|X\| = \|Y\|$ if and only if $(X+Y) \perp (X-Y)$.

32. In the space $\mathbb{R}^2$, introduce the new scalar product
$$< X, Y >= x_1y_1 + 4x_2y_2,$$
where $X = (x_1, x_2)$ and $Y = (y_1, y_2)$.
    (a) Verify that this indeed is a scalar product and define the associated norm $\|X\|$.
    (b) Let $X_1 = (0, 1)$ and $X_2 = (4, -2)$. Using *this* norm and scalar product, find an orthonormal set of vectors e_1 and e_2 such that e_1 is in the subspace spanned by X_1.

33. Let H be a scalar product space with X and Y in H. Find a scalar α which makes $\|X - \alpha Y\|$ a minimum. For this α, how are $X - \alpha Y$ and Y related? [Hint: Draw a picture in $\mathbb{R}^2$].

34. If $\sum_{n=1}^{\infty} a_n$ converges, where $a_n \geq 0$, does the series $\sum_1^{\infty} \frac{\sqrt{a}}{n^2}$ also converge? Proof or counterexample.

35. Use the Taylor series about $x_0 = 0$ to calculate $sin.2$ making an error less than $.005$. Justify your statements.

36. Let $A = \text{span}\{\,(1,1,1,1),\ (1,0,1,0)\,\}$ be a subspace of $\mathbb{R}^4$. Find the orthogonal complement, $A^{\perp}$, of A by giving a basis for $A^{\perp}$.

37. Prove that
    (a) $1 + \frac{1}{8} < \sum_{k=1}^{\infty} \frac{1}{k^3} < 1 + \frac{1}{2}$.
    (b) $1 + \frac{1}{2} \sum_{k=1}^{\infty} \frac{1}{k^2} < 1 + \frac{3}{4}$.

38. Let a_k be a sequence of positive numbers decreasing to zero, $a_k \to 0$, and let $S_N = a_1 + a_2 + \cdots + a_N$.
    (a) Prove that $S_N \geq N a_N$.
    (b) Use this to estimate the number, N, of terms needed to make
$$\sum_{k=1}^{N} k^{-1/4} > 1000.$$

39. Prove or give a counterexample:
    (a) If $\sum_{n=1}^{\infty} b_n$ converges, then $\sum_{n=1}^{\infty} b_{2n}$ must converge.
    (b) If $\sum_{n=1}^{\infty} |b_n|$ converges, then $\sum_{n=1}^{\infty} |b_{2n}|$ must converge.

40. Let X_1 and X_2 be elements of a scalar product space.
    (a) If $X_1 \perp X_2$, prove that $\|X_1 - aX_2\| \leq \|X_1\|$ for any real number a.
    (b) Prove the converse, that is, if $\|X_1 - aX_2\| \leq \|X_1\|$ for every real number a, then $X_1 \perp X_2$. [Hint: After your first approach has failed, try looking at the problem geometrically. How would you pick a to minimize the left side of the inequality?].

41. Let $S_n = a_1 + a_2 + \cdots + a_n$, where $a_n \to 0$ as $n \to \infty$. Prove that S_n converges if and only if $S_{2n} = a_1 + a_2 + \cdots + a_{2n-1} + a_{2n}$ converges (one could also use S_{3n} etc.).

42. Show that the error in approximating the series $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^n}$ by the first N terms is less than N^{-N-1}.

43. A sample "multiplication" for points $X = (x_1, x_2, x_3)$ and $Y = (y_1, y_2, y_3)$ in $\mathbb{R}^3$ is to define
$$X \odot Y \equiv (x_1 y_2, x_2 y_2, x_3 y_3).$$
Define a multiplicative identity by yourself. Using these definitions for the multiplicative structure and the usual rules for the additive structure, show that the resulting algebraic object is not a field.

44. (a) Assume $a_n \geq 0$ and $b_n \geq 0$. Prove that $\angle(a_n + b_n)$ converges if and only if the series $\angle a_n$ and $\angle b_n$ *both* converge.
(b) What if you allow the b_n's to be negative?

45. (a) Show that the vectors $e_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$, $e_2 = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$ form an orthonormal basis for $\mathbb{R}^2$.
(b) Write the vector $X = (7, -3)$ in the form $X = a_1 e_1 + a_2 e_2$, using the scalar product to find a_1 and a_2 (don't solve linear equations).

46. Consider the linear space $\mathcal{P}_2$ as a subspace of $L_2[0,1]$.
(a) If $p(x) = 1 - x^2$, compute $\|p\|$.
(b) Find the orthonormal basis for $A^{\perp}$, where $A = \text{span}\{\, 2 + x \,\}$.
(c) Find the polynomial $\varphi \in \mathcal{P}_2$ such that
$$\langle p,\, \varphi \rangle = p(1) \quad \text{for all} \quad p \in \mathcal{P}_2,$$
that is, the same φ should work for all p's.

47. Give formal proofs for the following (trivial) properties of a norm on a linear space. Only the axioms may be used.
(a) $\|-X\| = \|X\|$
(b) $\|X - Y\| = \|Y - X\|$
(c) $\|X + Y\| \geq \|X\| - \|Y\|$
(d) $\|X_1 + X_2 + \cdots + X_n\| \leq \|X_1\| + \|X_2\| + \cdots + \|X_n\|$ (I suggest induction here).

48. Consider $\mathbb{R}^2$ with the norms $\|\quad\|_1$, $\|\quad\|_2$, and $\|\quad\|_\infty$.

(a) Draw a sketch of $\mathbb{R}^2$ indicating the unit ball for each of these three norms. (The ball may not turn out to be "round").
(b) Which of these three linear spaces have the following property: "given any subspace M and a point X_0 not in M, then there is a *unique* point on M which is closest to M."

49. Are the following scalar products the set of functions continuous on $[a, b]$? Proof or counterexample.
(a) $[f, g] = (\int_a^b f(x)\, dx)(\int_a^b g(x)\, dx)$
(b) $[f, g] = (\int_a^b |f(x)|\, dx)(\int_a^b |g(x)|\, dx)$

50. (a) Let $\dim V = n$ and $\{\, X_1, \ldots, X_n \,\} \in V$. Prove that $\{\, X_1, \ldots, X_n \,\}$ are linearly independent if and only if they span V (so in either case, they form a basis for V).
(b) Let $\{\, e_1, \ldots, e_n \,\}$ be an orthonormal set of vectors for an inner product space H. Prove this set of vectors is a complete orthonormal set for H if and only if $n = \dim H$.
(c) Prove that $\dim V =$ largest possible number of linearly independent vectors in V.

51. (a) Let X and Y be any two elements in an inner product space. Prove that the parallelogram law holds
$$\|X + Y\|^2 + \|X - Y\|^2 = 2\|X\|^2 + 2\|Y\|^2$$
(cf. page 192, Ex. 9).
(b) Consider the set of continuous functions on $[0, 1]$ with the uniform norm, $\|f\|_\infty = \max_{0 \leq x \leq 1} |f(x)|$. Show that this norm cannot arise from an inner product, i.e. there is no inner product such that for all f, $\|f\|_\infty = \sqrt{\langle f,\, f \rangle}$. [Hint: If there were, the relationship of part a would hold between the norms of various elements. Show that relationship does not, in fact, hold for the function $f(x) = 1$ and $g(x) = x$].

52. (a) Let H be a finite dimensional inner product space and $\ell(X)$ a linear functional defined for all $X \in H$. Show that there is a fixed vector $X_0 \in H$ such that
$$\ell(X) = \langle X,\, X_0 \rangle \qquad \text{for all} \quad X \in H.$$
This shows that every linear functional can be represented simply as the result of taking the inner product with some vector X_0. [Hint: First pick a basis $\{\, e_1, \ldots, e_n \,\}$ for H and let $c_j = \ell(e_n)$. Now use the fact that the e_j's are a basis and that ℓ is linear].

(b) Consider the linear space $\mathcal{P}_2$ with the $L_2[0,1]$ inner product. This gives an inner product space H.

(i) Show that $\ell(p) = p(\frac{1}{3})$ is a linear functional.

(ii) Find a polynomial p_0 such that $\ell(p) = \langle p, p_0\rangle$ for all $p \in H$.

53. Consider the set S of pairs of real numbers $X = (x_1, x_2)$. Define

$$X + Y = (x_1 + y_1, x_2 + y_2), \quad aX = (ax_1, x_2).$$

Is S, with this definition of vector addition and multiplication by scalars, a vector space?

54. By inspection, place suitable restrictions on the contents $a, b, c, \cdots$ in order to make the following operator linear:

$$Tu = a[\frac{d^3u}{dx^3}]^2 + bx^2\frac{d^2u}{dx^2} + cu\frac{du}{dx} + eu + f\sin u + g.$$

55. Consider the operator $D = \frac{d}{dx}$ on the linear space $\mathcal{P}_n$ of all polynomials of degree less than or equal to n. Find $\mathcal{R}(D)$ and $\mathcal{N}(D)$ as well as $\dim\mathcal{R}(D)$ and $\dim\mathcal{N}(D)$.

56. Let

$$A = \begin{pmatrix} 1 & -2 \\ 2 & 0 \\ 3 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & 0 & -2 \\ 2 & 1 & 0 \end{pmatrix}, \quad \text{and} \quad C = \begin{pmatrix} 3 & 0 & 2 \\ 1 & 4 & -1 \\ 0 & -2 & 0 \end{pmatrix}.$$

Compute all of the following products which make sense:

$$AB,\ BA,\ AC,\ CA,\ BC,\ CB,\ A^2,\ B^2,\ C^2,\ ABC, CAB.$$

57. Consider the mapping $A : \mathbb{R}^4 \to \mathbb{R}^3$ which is defined by the matrix

$$A = \begin{pmatrix} 1 & -1 & 1 & 1 \\ 2 & 1 & 1 & 4 \\ 0 & -3 & 1 & -2 \end{pmatrix}$$

(a) Find bases for $\mathcal{N}(A)$ and $\mathcal{R}(A)$.

(b) Compute $\dim\mathcal{N}(A)$ and $\dim\mathcal{R}(A)$.

58. Let A be a square matrix. Consider the system of linear algebraic equations

$$AX = Y_0,$$

where Y_0 is a fixed vector. Assume these equations have two *distinct* solutions X_1 and X_2,

$$AX_1 = Y_0, \quad AX_2 = Y_0, \quad X_1 \neq X_2.$$

(a) Find a third solution X_3.

(b) Does there exist a vector Y_1 such that the equations

$$AX = Y_1$$

have *no* solutions? Why?

(c) $\det A =$?

59. Let Q be a parallelepiped in $\mathbb{R}^n$ whose vertices X_k are at points with *integer* coordinates,

$$X_k = (a_{1k}, a_{2k}, \cdots a_{nk}), \quad a_{ik} \quad \text{integers}.$$

Prove that the volume of Q is an integer.

60. Let A and B be self-adjoint matrices. Prove that their product AB is self-adjoint if and only if $AB = BA$.

61. Solve the following initial value problems.

(a) $u'' + 8u' + 16u = 0, \quad u(0) = \frac{1}{2},\ u'(0) = 0$

(b) $u'' + 10u' + 16u = 0, \quad u(0) = 1,\ u'(0) = 2$

(c) $u'' + 64u = 0, \quad u(0) = \frac{1}{4},\ u'(0) = 1$

(d) $u'' + 4u' + 5u = 0, \quad u(0) = 2,\ u'(0) = -1$

(e) $2u'' + 6u' + 5u = 0, \quad u(0) = 0,\ u'(0) = -2$

(f) $4u'' - 4u' + u = 0, \quad u(1) = -1,\ u'(1) = 0$

(g) $u'' + 8u' + 16u = 2, \quad u(0) = \frac{1}{2},\ u'(0) = 0$

(h) $u'' + 8u' + 16u = t, \quad u(0) = \frac{1}{2},\ u'(0) = 0$

(i) $u'' + 8u' + 16u = t - 2, \quad u(0) = 0,\ u'(0) = 0$

(j) $u'' + 8u' + 16u = t - 2, \quad u(0) = \frac{1}{2},\ u'(0) = 0$

(k) $u'' + 10u' + 16u = t, \quad u(0) = 1,\ u'(0) = 2$

(l) $u'' + 64u = 64, \quad u(0) = \frac{1}{4},\ u'(0) = 2$

(m) $u'' + 64u = t - 64, \quad u(0) = \frac{3}{4},\ u(0) = 0$

(n) $2u'' + 6u' + 5u = t^2, \quad u(0) = 0,\ u'(0) = -2$

62. (The complex numbers as matrices).

(a) Show that the set of matrices

$$\mathbb{C} = \{ \begin{pmatrix} a & -b \\ b & a \end{pmatrix} : a \,\text{and}\, b \,\text{are real numbers} \}$$

is a field.

(b) Find a map $\varphi : \mathbb{C} \to$ complex numbers such that φ is bijective and such that for all $A, B \in \mathbb{C}$

(i) $\varphi(A+B) = \varphi(A) + \varphi(B)$

(ii) $\varphi(AB) = \varphi(A)\varphi(B)$.

63. (Quaternions as matrices). A definition: A *division ring* is an algebraic object which satisfies all of the field axioms *except* commutativity of multiplication.

(a) Show that the set of matrices

$$Q = \mathbb{R}\begin{pmatrix} z & -\bar{w} \\ w & \bar{z} \end{pmatrix} : z, w \quad \text{are complex numbers}$$

form a division ring with the usual definitions of additions and multiplication for matrices.

(b) If we write $z = x + iy$, $w = u + iv$ where $i = \sqrt{-1}$ and x, y, u, and v are real numbers, then Q can be considered as a vector space over the reals with basis

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \mathbf{i} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \quad \mathbf{j} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \mathbf{k} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}.$$

Compute $\mathbf{i}^2, \mathbf{j}^2, \mathbf{k}^2, \mathbf{ij}, \mathbf{jk}, \mathbf{ki}, \mathbf{ji}, \mathbf{kj}$, and $\mathbf{ik}$. (The set Q is called the *quaternions*).

64. Let

$$A = \begin{pmatrix} 2 & -3 & 1 & 0 \\ 0 & 2 & -3 & 1 \\ 0 & 0 & 2 & -3 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$

(a) Find $\det A$.

(b) Find A^{-1}.

(c) Solve $AX = Y$, where $Y = \begin{pmatrix} 2 \\ 8 \\ 8 \\ -16 \end{pmatrix}$.

(d) Let $L : \mathcal{P}_3 \to \mathcal{P}_3$ be the linear operator defined by

$$Lp = p'' - 3p' + 2p, \quad (u' = \frac{du}{dx}).$$

Find the matrix $e^L e$ for L with respect to the following basis for $\mathcal{P}_3$

$$e_1(x) = 1, \quad e_2(x) = x, \quad e_3(x) = \frac{x^2}{2}, \quad e_4(x) = \frac{x^3}{3!}.$$

(e) Use the above results to find a solution of

$$Lu = 2 + 8x + 4x^2 - \frac{8}{3}x^3.$$

[Hint: Express the right side in the basis of part d.].

65. Let H be an inner product space, and suppose that A is a symmetric operator, $A* = A$, with the additional property that $A^2 = A$. Show that there exist two subspaces V_1 and V_2 of H with all of the following properties

(i) $V_1 \perp V_2$

(ii) If $X \in V_1$, then $AX = X$

(iii) If $Y \in V_2$, then $AY = 0$

(iv) If $Z \in H$, then Z can be written uniquely as $Z = X + Y$ where $X \in V_1$ and $Y \in V_2$.

66. (a) Find the inverse of the matrix

$$A = \begin{pmatrix} 2 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & -1 \end{pmatrix}.$$

(b) Use the result of a) to solve $AX = b$ for X where $b = (7, -3, 2)$.

67. Let A and B be 2×2 positive definite matrices with $\det A = \det B$. Prove that $\det(A - B) < 0$.

68. Let $L : V_1 \to V_2$ be a linear operator with $LX_1 = Y_1$ and $LX_2 = Y_2$. Give a proof or counterexample to each of the following assertions:

(a) If X_1 and X_2 are linearly independent, then Y_1 and Y_2 must be linearly independent.

(b) If Y_1 and Y_2 are linearly independent, then X_1 and X_2 must be linearly independent.

69. Let $p_0, p_1, p_2, , \ldots$ be an orthogonal set of polynomials in $[a, b]$ where p_n has degree n.

(a) Prove that p_n is orthogonal to $1, x, x^2, \ldots, x^{n-1}$.

(b) Prove that p_n is orthogonal to any polynomial q of degree less than n.

(c) Prove that p_n has exactly n distinct real zeros in (a,b). [Hint: Let $\alpha_1,\dots,\alpha_k$ be the places in (a,b) where $p_n(x)$ changes sign, so $p(x) = r(x)(x-\alpha_1)(x-\alpha_2)\dots(x-\alpha_k)$ where $r(x)$ is a polynomial of degree $n-k$ which does *not* change sign for x in (a,b), say $r(x)\ge 0$. Show that

$$\int_a^b p(x)(x-\alpha_1)\cdots(x-\alpha_k)\,dx > 0.$$

If $k<n$, show that this contradicts the result of part b).].

70. Consider the system of inhomogeneous equations

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b, \\ &\vdots \\ a_{k1}x_1 + \cdots + a_{kn}x_n &= b_n. \end{aligned}$$

Let $A = ((a_{ij}))$ and let A_b denote the *augmented matrix*

$$A_b = \begin{pmatrix} a_{11} & \cdots & a_{1n} & b_1 \\ \vdots & & & \\ a_{k1} & \cdots & a_{kn} & b_n \end{pmatrix}$$

formed by adding the b_j's as an extra column to A. Prove that the given system of equations has a solution if and only if $\dim\mathcal{R}(A) = \dim\mathcal{R}(A_b)$.

71. Let A be an $n\times n$ matrix.

(a) Show that you can not solve the equation

$$A^2 = -I$$

if n is odd.

(b) Find a 2×2 matrix A such that $A^2=-I$.

(c) If n is even, find an $n\times n$ matrix A such that $A^2=-I$.

72. Let A be an $n\times n$ matrix such that $A^2=I$. Prove that $\dim\mathcal{R}(A+I)+\dim\mathcal{R}(A-I) = n$.

73. Let $f(x,y) = (y-2x^2)(y-x^2)$. Show that the origin is a critical point. Then show that if you approach the origin along a straight line, the origin appears to be a minimum. On the other hand, show that if curved paths are also used, then the origin is a saddle point of f. [The point of this exercise is to illustrate the fact that the nature of a critical point cannot be determined by merely approaching it along straight lines].

74. (a) Let A be a diagonal matrix, no two of whose diagonal elements are the same. If B is another matrix and $AB=BA$, prove that B is also diagonal.

(b) Let A be a diagonal matrix, B a matrix with at least one zero-free column and with the further property that $AB=BA$. Prove that all of the diagonal elements of A are equal.

75. (a) If $\sum a_n$ converges, where $a_n\ge 0$, prove that $\sum \frac{\sqrt{a_n}}{n^p}$ converges if $p>\frac12$. [Hint: Schwarz].

(b) Find an example showing that the series may diverge if $p=\frac12$.

76. Let $[X,Y]$ be an inner product on $\mathbb{R}^3$ with basis vectors e_1,e_2,e_3, not necessarily orthonormal. Let $a_{ij}=[e_i,e_j]$. Prove that the quadratic form

$$Q(X) = \sum_{i=1}^3\sum_{j=1}^3 a_{ij}x_ix_j$$

is positive definite.

77. If A is self-adjoint and $AX=\lambda_1X$, $AY=\lambda_2Y$ with $\lambda_1\ne\lambda_2$, prove that $X\perp Y$.

78. Let S be a positive definite matrix. Prove that $\det S>0$. [Hint: Consider the matrix $A(t)\equiv tS+(1-t)I$, where $0\le t\le 1$. Show that $A(t)$ is positive definite, so $\det A(t)\ne 0$. Then use the fact that $A(0)=I$ and $A(1)=S$ to obtain the conclusion].

79. Consider the linear space of infinite sequences

$$X=(x_1,x_2,x_3,\cdots)$$

with the usual addition. Define the linear operator S (the *right shift* operator) by

$$SX=(0,x_1,x_2,x_3,\cdots)$$

(a) Does S have a left inverse? If so, what is it?

(b) Does S have a right inverse? If so, what is it?

80. Find a right inverse for the matrix

$$A=\begin{pmatrix}1&0&1\\0&1&0\end{pmatrix}.$$

Can A have a left inverse? Why?

81. Which of the following statements are true for *all* square matrices A? Proof or counterexample.

(a) If $A^2 = I$, then $\det A = I$.

(b) If $A^2 = A$, then $\det A = 1$.

(c) If $A^2 = 0$, then $\det A = 0$

(d) If $A^2 = I - A$, then $\det A^2 = 1 - \det A$.

82. Let L be a linear operator on an inner product space H with inner product $<,>$. Define
$$[X, Y] = \langle LX,\, LY\rangle.$$
Under what further condition(s) on L is $[X, Y]$ an inner product too?

83. Let $L : H \to H$ be an invertible transformation on the inner product space H. If L "preserves orthogonality" in the sense that $X \perp Y$ implies $LX \perp LY$, prove that there is a constant α such that $R \equiv \alpha L$ is an orthogonal transformation.

84. Let H be an inner product space. If the vectors X_1 and X_2 are at opposite ends of a diameter of the sphere of radius r about the origin, and if Y is any other point on that sphere, prove that $Y - X_1$ is perpendicular to $Y - X_2$, proving that an angle inscribed in a hemisphere is a right angle.

85. If L is skew-adjoint, $L* = -L$, prove that
$$\langle X,\, LX\rangle = 0 \quad \text{for all } X.$$

86. Let D_n be a $n \times n$ matrix with x on the main diagonal and $1's$ on both the sub- and super-diagonals, so
$$D_2 = \begin{pmatrix} x & 1 \\ 1 & x \end{pmatrix}, \quad D_3 = \begin{pmatrix} x & 1 & 0 \\ 1 & x & 1 \\ 0 & 1 & x \end{pmatrix}, \quad D_4 = \begin{pmatrix} x & 1 & 0 & 0 \\ 1 & x & 1 & 0 \\ 0 & 1 & x & 1 \\ 0 & 0 & 1 & x \end{pmatrix}, \quad D_5 = \cdots .$$
If $x = 2\cos\theta$, prove that $\det D_n = \frac{\sin(n+1)\theta}{\sin\theta}$.

87. Let A and B be square matrices of the same size. If $I - AB$ is invertible, prove that $I - BA$ is also invertible by exhibiting a formula for its inverse.

88. Assume $\sum a_n$ converges, where $a_n \geq 0$. Does the series
$$\sum \sqrt{a_n a_{n+1}}$$
also converge? Proof or counterexample.

89. Let A be a square matrix.

(a) Prove that AA^* is self-adjoint.

(b) Is AA^* always equal to A^*A? Proof or counterexample.

90. Show that $C[0,1]$ is a direct sum of the space V_1 spanned by $e_1(x) = x$ and $e_2(x) = x^4$, and the subspace V_2 of all functions $\varphi(x)$ such that
$$0 = \int_0^1 x\varphi(x)\,dx, \quad 0 = \int_0^1 x^4\varphi(x)\,dx.$$
[Hint: Show that if $f \in [0,1]$, there are unique constants a and b such that $g(x) \equiv f(x) - [ax + bx^4]$ belongs to V_2].

91. Let V_1 be the linear space of all complex-valued analytic functions in the open unit disc, that is, V_1 consists of all complex-valued functions f of the complex variable z which have convergent power series expansions
$$f(z) = \sum_0^\infty a_n z^n$$
in the open disc, $|z| < 1$.

Let V_2 be the linear space of all sequences of complex numbers $(a_0, a_1, a_2, \cdots)$ with the natural definition of addition and multiplication by constants.

Define $L : V_1 \to V_2$ by the rule
$$Lf = (a_0, a_1, a_2, \cdots),$$
where the a_j's are the Taylor series coefficients of f. Answer the following questions with a proof or counterexample.

(a) Is L injective?

(b) Is L surjective?

(c) Is ℓ_2 contained in $\mathcal{R}(L)$? (Note: ℓ_2 is the subspace of V_2 such that
$$\sum_{k=0}^\infty |a_k|^2 < \infty).$$

92. Do the following series converge or diverge?
$$(a) \sum_{n=1}^\infty \sqrt{1 + 1/n}, \qquad (b) \sum_{n=1}^\infty (\sqrt{1 + 1/n^2} - 1).$$

93. Consider the set of four operators $\{T_1, T_2, T_3, T_4\}$ defined as follows on the set of square invertible matrices.

$$T_1A = A, \qquad T_2A = A^{-1}$$
$$T_3A = A^*, \qquad T_4A = (A^{-1})^*.$$

Show that this set of four operators forms a commutative group with the group operation being ordinary operator multiplication.

94. Let $S_n = a_1 - a_2 + a_3 - a_4 + a_5 - \cdots$. If $0 < a_k$ and the a_k's are increasing, prove that $|S_N| \le a_N$.

95. The *Monge-Ampere* equation is $u_{xx}u_{yy} - u_{xy}^2 = 0$. Show that it is satisfied by any $u(x,y) \in C^2$ of the form $u(x,y) = \varphi(ax+by)$, where a and b are constants.

96. (a) Consider the differential operator

$$Lu = u'' - 4u$$

(i) Find a basis for the nullspace of L.
(ii) Find a particular solution of $Lu = e^{2x+1}$.
(iii) Find the general solution of $Lu = e^{2x+1}$.

(b) Consider the differential operator

$$Lu = u'' + 4u$$

Repeat part (a), only here use $Lu = f$, where $f(x) = \sec 2x$.

97. Find the general solution for each of the following

(a) $2u'' + 5u' - 3u = 0$
(b) $u'' - 6u' + 9u = 0$
(c) $u'' - 4u' + 5u = 0$

98. Find the first four non-zero terms in the series solution of

$$4x^2u'' - 4xu' + (3 - 4x^2)u = 0$$

corresponding to the largest root of the indicial equation. Where does the series converge?

99. Find the complete solution of each of the following equations valid near $x = 0$ by using power series.

(a) $x^2u'' + xu' - (x^2 - \frac{1}{4})u = 0$
(b) $u'' + xu' - u = 0$ (only first five non-zero terms)

[Answers:

(a) $u(x) = Ax^{-1/2}\sum_{k=0}^{\infty}\frac{x^{2k}}{(2k)!} + Bx^{1/2}\sum_{k=0}^{\infty}\frac{x^{2k}}{(2k+1)!}$,
(b) $u(x) = Ax + B(1 + \frac{x^2}{2!} - \frac{x^4}{4!} + \frac{3x^6}{6!} - \frac{15x^8}{8!} + \cdots)$].

100. Consider the matrix

$$A = \begin{pmatrix} -1 & -4 & -12 & 0 \\ 1 & 3 & 6 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & -4 & -12 & 1 \end{pmatrix}.$$

(a) Compute $\det A$.
(b) Compute A^{-1}.
(c) Solve $AX = b$ where $b = (1, 2, 3, -1)$.

101. True or false. Justify your response if you believe the statement is false (a counterexample is adequate).

(a) The set $A = \{X \in \mathbb{R}^3 \colon x_1 = 2\}$ is a linear *subspace* of $\mathbb{R}^3$.
(b) The vectors $X_1 = (2,4)$ and $X_2 = (-2,4)$ *span* $\mathbb{R}^2$.
(c) The vectors $X_1 = (1,2,3)$, $X_2 = (-7,3,2)$, $X_3 = (2,-1,1)$, and $X_4 = (\pi, e, 5)$ are linearly *independent*.
(d) The set $A = \{u \in C[0,1] \colon u(x) = a_1x + a_2e^x\}$ is an *infinite dimensional* subspace of $C[0,1]$.
(e) The functions $f_1(x) = x$ and $f_2(x) = e^x$ are *linearly dependent* functions in $C[0,1]$.
(f) If $\{e_1, e_2, \ldots, e_n\}$ are an orthonormal set of vectors in $\mathbb{R}^8$, then $n \le 7$.
(g) The vector $Y = (1,2,3)$ is *orthogonal* to the subspace of $\mathbb{R}^3$ spanned by $e_1 = (0,3,-2)$ and $e_2 = (-1,-1,1)$.
(h) The elements of the set

$$A = \{u \in C^2[0,10] \colon u'' + xu' - 3u = 6x\}$$

can be represented as $u(x) = \tilde{u}(x) + x^3$, where

$$\tilde{u} \in S = \{u \in C^2[0,10] \colon u'' + xu - 3x = 0\}.$$

(i) The set of vectors $e_1 = (\frac{1}{3}, 0, \frac{2}{3}, -\frac{2}{3})$, $e_2 = (0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$, and $e_3 = (\frac{8}{9}, \frac{3}{9}, -\frac{2}{9}, \frac{2}{9})$ constitute a *complete* orthonormal basis for $\mathbb{R}^4$.

(j) In the vector space of bounded functions $f(x)$, $x \in [0,1]$, the functions

$$f_1(x) = 1, \quad f_2(x) = \begin{cases} 1, & 0 \le x \le \frac{1}{2}, \\ 0, & \frac{1}{2} < x \le 1 \end{cases} \quad f_3(x) = \begin{cases} 0, & 0 \le x \le \frac{1}{2} \\ 1, & \frac{1}{2} < x \le 1 \end{cases}$$

are linearly *independent*.

(k) The function $f(x) = |x|$ can be represented by a convergent Taylor series about the point $x_0 = 0$.

(l) The function $f(x) = x^2 - x^{73}$ can be represented by a convergent Taylor series about the point $x_0 = -1$.

(m) The function $f(x) = |x|$ can be represented by a convergent Taylor series about the point $x_0 = -1$.

(n) The plane of all points $(x_1, x_2, x_3, x_4) \in \mathbb{R}^4$ such that

$$2x_1 - 4x_2 + 6x_3 - 5x_4 = 7$$

is perpendicular to the vector $(2, -4, 6, -5)$.

(o) If $e_1 = (\frac{3}{5}, \frac{4}{5})$ and $e_2 = (\frac{4}{5}, -\frac{3}{5})$, then $X = (-1, 2)$ can be written as $X = 2e_1 - e_2$.

(p) The set of all integers (positive, negative, and zero) is a *field*.

(q) Consider the infinite series

$$\sum_{k=0}^{\infty} a_k.$$

If $\lim_{k \to 0} |a_k| = 0$, then the series *must converge*.

(r) Let $\{a_n\}$ be a sequence of rational numbers. If this sequence converges to a, then the limiting value, a, must be a *rational number* too.

(s) The equation $x^6 + 3 = 0$, where x is an element of an ordered field, *has no solutions*.

(t) It is possible to write $\sqrt{i}$ in the form $a + ib$, where a and b are real numbers. (Here $i = \sqrt{-1}$, of course).

(u) Let a_n be a *sequence* of complex numbers. If the sequence of absolute values, $|a_n|$, converges, then the sequence a_n *must converge*.

(v) If

$$\sum_{k=0}^{\infty} a_k z^k$$

converges at the point $z = 3$, then it *must converge* at $z = 1 + i$.

(w) The linear subspace $A = \{p \in \mathcal{P}_7 \colon p(x) = a_1 x + a_2 x^5\}$ is a *five dimensional* subspace of $\mathcal{P}_7$.

(x) The linear subspace $A = \{u \in C[-1,1] \colon u(x) = a_1 x + a_2 x^5\}$ is an *infinite dimensional* subspace of $C[-1,1]$.

(y) There is a number α such that the vectors $X = (1,1,1)$ and $Y = (1, \alpha, \alpha^2)$ form a *basis* for $\mathbb{R}^3$.

(z) The operator $T : C^2 \to C^1$ defined for $u \in C^2$ by $Tu = u' - 7u$ is a *linear* operator.

102. (a) The operator $T : C[0,1] \to \mathbb{R}$ defined for $u \in C[0,1]$ by

$$Tu = \int_0^1 |u(x)| \, dx$$

is a *linear* operator.

(b) The sequence $(1+i)^n$ converges to $\sqrt{2}$.

(c) The series

$$\sum_{k=1}^{\infty} \frac{k+1}{2k+1} = \frac{2}{3} + \frac{3}{5} + \frac{4}{7} + \frac{5}{9} + \cdots$$

converges.

(d) If t is real, then $\left|e^{it}\right| = 1$.

(e) Let V_1 and V_2 be linear spaces and let the operator T map V_1 into V_2. If $T0 = 0$, then T is a *linear* operator.

(f) The operator $T : C^{\infty}[-7,13] \to C^{\infty}[-7,13]$ defined by

$$Tu = u\frac{du}{dx}$$

is *linear*.

(g) The operator $T : C[0,13] \to C[0,13]$ defined by

$$(Tu)(x) = \int_0^x u(t) \sin t \, dt, \quad x \in [0,13]$$

is *linear*

(h) In the scalar product space $L_2[0,1]$, the functions f and g whose graphs are

A FIGURE GOES HERE

are *orthogonal*.

(i) Let L be a linear operator. IF $LX_1 = Y$ and $LX_2 = Y$, where $X_1 \neq X_2$, then the solution of the homogeneous equation $LX = 0$ is *not unique*.

(j) Let L be a linear operator. If X_1 and X_2 are solutions of $LX = 0$, then $3X_1 - 7X_2$ is *also* a solution of $LX = 0$.

(k) Let $e_1 = (1,1)$ and $e_2 = (0,1)$, and let the linear operator L which maps $\mathbb{R}^2$ into $\mathbb{R}^3$ satisfy

$$Le_1 = (1,2,3), \quad Le_2 = (1,-2,-1).$$

Then $L(2,3) = (1,1,1)$.

(l) In the space $L_2[0,1]$, if f is *orthogonal* to the function x^2, then either $f \equiv 0$ or *else* f must be *positive somewhere* in $[0,1]$.

(m) If $F'(X) = (2,3,4)$ for *all* $X \in \mathbb{R}^3$, then F is an *affine* mapping.

(n) If $f : \mathbb{R}^3 \to \mathbb{R}^1$ is such that $f : (1,0,0) \to 1$ and $f : (0,4,0) \to 2$, there is a point $Z \in \mathbb{R}^3$ such that $\|f'(Z)\| \geq \frac{1}{5}$.

(o) Let A and B be square matrices with $\det A = 7$ and $\det B = 3$. Then

$$\det AB = 10. \qquad \det(A+B) = 10.$$

(p) If $A : \mathbb{R}^3 \to \mathbb{R}^3$ is given by

$$A = \begin{pmatrix} 2 & 3 & 1 \\ 1 & 9 & 2 \end{pmatrix},$$

then $\dim \mathcal{N}(A) = 2$.

(q) The function $f(x,y,z) = 9 + 3x + 4y - 7z$ does not take on its maximum value.

(r) If the function $u(x)$ has two derivatives in some neighborhood of $x = 0$, and satisfies the differential equation

$$9x^2u'' - 28u = 0,$$

then $u(0) = 0$.

(s) There are *constants* a, b and c such that the function $u(x) = e^x + 2e^{2x} - e^{-x}$ is a solution of

$$au'' + bu' + cu = 0.$$

(t) The vector (xy, x) *is* the derivative of some real-valued function $f(x,y)$.

(u) The vector (y,x) is *not* the derivative of some real-valued function $f(x,y)$.

(v) Given any $q \times p$ matrix $A = ((a_{ij}(X)))$, where $X = (x_1, \cdots, x_p)$ and where the elements $a_{ij}(X)$ are sufficiently differentiable functions, then *there is a map* $F : \mathbb{R}^p \to \mathbb{R}^q$ such that $F'(X) = A$.

(w) If A is a square matrix and $A^2 = A$, then $A = I$.

(x) If A is a square matrix and $A^2 = 0$, then $A = 0$.

(y) If A is a square matrix and $\det A \neq 0$, then $A^2 = A$ if and only if $A = I$.

(z) If X, Y, and Z are three linearly independent vectors, then $X + Y$, $\;Y + Z$, and $X + Z$ are also linearly *independent*.

103. Define $L : \mathcal{P}_2 \to \mathcal{P}_2$ as follows: if $p \in \mathcal{P}_2$

$$Lp = (x+1)\frac{dp}{dx}$$

(a) Find the matrix ${}_eL_e$ representing the operator L with respect to the bases $e_1 = 1$, $e_2 = x_1$, $e_3 = x^2$ for $\mathcal{P}_2$.

(b) Is L an invertible operator? Why?

(c) Find $\dim \mathcal{R}(L)$ and $\dim \mathcal{N}(L)$.

104. Let

$$A = \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}, \qquad B = \begin{pmatrix} 5 & \sqrt{3} \\ \sqrt{3} & 3 \end{pmatrix}.$$

(a) Compute AA^*, ABA^*, and $(ABA^*)^{100}$.

(b) How could you use the result of part (a) to compute B^{100}?

105. Consider the following system of three equations as a linear map $L : \mathbb{R}^2 \to \mathbb{R}^3$

$$x_1 + x_2 = y_1$$

$$4x_1 + x_2 = y_2$$

$$x_1 - 2x_2 = y_3$$

(a) Find a basis for $\mathcal{N}(L^*)$.

(b) Use the result of part a) to determine the value(s) of α such that $Y = (1,2,\alpha)$ is in $\mathcal{R}(L)$.

106. Find the unique solution to each of the following initial value problems.

(a) $u'' + u' - 2u = 0, \quad u(0) = 3, \quad u'(0) = 0$

(b) $u'' + 4u' + 4u = 0, \quad u(0) = 1 \quad u'(0) = -1$

(c) $u'' - 2u' + 5u = 0, \quad u(0) = 2, \quad u'(0) = 2$

107. Consider the special second order inhomogeneous constant coefficient O.D.E. $Lu = f$, where

$$Lu \equiv u'' - 4u,$$

and where f is assumed to be a suitably differentiable function which is periodic with period 2π, $f(x + 2\pi) = f(x)$.

(a) Expand f in its Fourier series and seek a candidate, u, for a solution of $Lu = f$ as a Fourier series, showing how the Fourier coefficients of u are determined by the Fourier coefficients of f.

(b) Apply the above procedure to the trivial example where

$$f(x) = \sin 3x - 4\cos 17x + 3\sin 36x.$$

108. (a) Find the directional derivative of the function

$$f(x, y) = 2 - x + xy$$

at the point $(0, 6)$ in the direction $(3, -4)$ by using the definition of the directional derivative as a limit. Check your answer by using the short method.

(b) Repeat part (a) for $f(x, y) = 1 - 3y + xy$.

109. Find and classify the critical points of the following functions.

(a) $f(x, y) = x^3 + y^2 - 3x - 2y + 2$
(b) $f(x, y) = x^2 - 4x + y^2 - 2y + 6$
(c) $f(x, y) = (x^2 + y^2)^2 - 8y^2$
(d) $f(x, y) = (x^2 - y^2)^2 - 8y^2$
(e) $f(x, y) = (x^2 - y^2)^2$
(f) $f(x, y) = x^2 - 2xy + \frac{1}{3}y^3 - 3y$

110. Consider the function $x^3 + y^2 - 3x - 2y + 2$. At the point $(2, 1)$ find the direction in which the directional derivative is greatest. Find the direction where it is least.

111. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a suitably differentiable function and let $X(t)$ be the equation of a smooth curve C in $\mathbb{R}^2$ on which f is identically constant, say, $f(X(t)) \equiv 4$. Show that on this curve, f' is perpendicular to the velocity vector $X'(t)$. [Hint: Do something to $\varphi(t) = f(X(t))$. The proof takes but one line.].

112. Consider the following statements concerning a function $f : \mathbb{R}^n \to \mathbb{R}$.

(A) f is continuous.
(B) f has a total derivative everywhere.
(C) f has first order partial derivatives everywhere.
(D) f has a total derivative everywhere which is continuous everywhere.
(E) f has first order partial derivatives everywhere and they are continuous functions everywhere.
(F) f is an affine function.
(G) $f' \equiv 0$.

(a) Which of these statements always imply which others. A sample (possibly incorrect) answer might look like

$$(A) \Rightarrow B, F, \cdots$$

$$(B) \Rightarrow A, \cdots$$

(b) Find examples illustrating each case where a given statement does not imply another (the Exercises, pp. 588-95, contain the required examples).

113. Solve the following ordinary differential equations subject to the given auxiliary conditions

(a) $u'' - u' - 6u = 0, \quad u(0) = 0, \quad u'(0) = 5$
(b) $xu' + u = e^{x-1}, \quad u(1) = 2$
(c) $u'' - 6u' + 10u = 0$, general solution.

114. (a) If $u(x, y, t) = xe^{xy} + t^2$, while $x = 1 - t^3$ and $y = \log t^2$, then let $w(t) = u(x(t), y(t), t)$. Find $\frac{dw}{dt}$ at $t = 1$.

(b) If $F : \mathbb{R}^3 \to \mathbb{R}^2$ and $G : \mathbb{R}^2 \to \mathbb{R}^2$ are defined by

$$F(X) = \begin{pmatrix} 2x_1 - x_2^2 + x_2x_3 + 1 \\ x_1^2 - x_3^2 + x_2 \end{pmatrix}, \quad G(Y) = \begin{pmatrix} y_1 + y_2 \sin y_1 \\ -3y_1y_2 + y_2^2 \end{pmatrix},$$

(i) Why doesn't $F \circ G$ make sense?
(ii) Compute $[G \circ F]'$ at the point $X_0 = (0, 1, 0)$.

115. Let $F = \mathbb{R}^2 \to \mathbb{R}^2$ and $G : \mathbb{R}^3 \to \mathbb{R}^2$ be defined by

$$F(w, z) = \begin{pmatrix} e^{w+z^2} \\ e^{z+w^2} \end{pmatrix} \quad G(r, s, t) = \begin{pmatrix} r + s^2 + t^3 \\ s + t^2 + r^3 \end{pmatrix}.$$

(a) Find F' and G'.
(b) Which of $F \circ G$ or $G \circ F$ makes sense?
(c) If $G \circ F$ makes sense, compute $(G \circ F)'$ at $(-1, -1)$.
(d) If $F \circ G$ makes sense, compute $(F \circ G)'$ at $(-1, 0, 0)$.

116. Let $F : X \to Y$ and $G : Y \to Z$ be defined by

$$F : \begin{cases} y_1 = x_2 - e^{x_1 + 2x_2} \\ y_2 = x_1 x_2 \end{cases}, \quad G : \begin{cases} w_1 = y_2 + y_2 \sin y_1 \\ w_2 = (y_1 + y_2)^2 \end{cases}$$

(a) Compute F' at $X_0 = (-2, 1)$ and G' at $Y_0 = F(X_0)$.

(b) Let $H = G \circ F$. Compute H' at $X_0 = (-2, 1)$.

117. Consider the map $F : \mathbb{R}^2 \to \mathbb{R}^3$ defined by

$$F : \begin{cases} f_1(x, y) = y + e^{x-y} \\ f_2(x, y) = \sin(x - 2y + 1) \\ f_3(x, y) = x - 3x^2 + y^2 \end{cases}$$

(a) Find the tangent map at the point $X_0 = (1, 1)$.

(b) Use the result of part (a) to evaluate approximately F at $X_1 = (1.1, .9)$.

118. Consider the system of O.D.E.'s

$$\begin{aligned} u' &= \alpha u \\ v' &= \alpha u - \beta v, \end{aligned}$$

where α and β are constants. If $u(0) = A$ and $v(0) = B$,

(a) Find $u(t)$.

(b) Find $v(t)$ (remember to consider the case $\alpha = \beta$ separately).

119. (a) Consider the homogeneous equation

$$u'' + a(t)u = 0,$$

where $a(t)$ is continuous and periodic with period P, so $a(t + P) = a(t)$.

(i) If $a(t) \equiv 1$, show that there is no non-trivial periodic solution by merely solving the equation.

(ii) If $a(t) = \cos t$, show (again by solving the equation) that there is a periodic solution $u(t)$ with period 2π.

(iii) In general, if $u(t)$ is a solution, not necessarily periodic, show that $v(t) \equiv u(t + P)$ is also a solution.

(iv) Show that the homogeneous equation has a non-trivial periodic solution of period P *if and only if*

$$\int_0^P a(t)\,dt = 0$$

(b) Consider the inhomogeneous equation

$$u + a(t)u = f(t),$$

where both $a(t)$ and $f(t)$ are continuous and periodic with period P.

(i) If $\int_0^P a(t)\,dt = K \neq 0$, show that the inhomogeneous equation has one and only one periodic solution with period P.

(ii) If $\int_0^P a(t)\,dt = 0$, find a necessary condition on f that the inhomogeneous equation have a periodic solution with period P.

120. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and denote the directional derivative in the direction of the unit vector e by $D_e f$. Prove that $D_{-e} f = -D_e f$.

121. Let $f : \mathbb{R}^n \to \mathbb{R}$ be of the form $f(a_1x_1 + \ldots + a_nx_n)$. Write $\alpha = (a_1, \cdots, a_n)$ and $\beta = (b_1, \cdots, b_n)$. If β is perpendicular to α, prove that $\beta \perp f'$.

122. Let R denote the rectangle $0 \leq x_1 < 2\pi$, $\;0 \leq x_2 < 2\pi$, and define the map $f : R \to \mathbb{R}^1$ by

$$f(x_1, x_2) = (3 + 2\cos x_2)\sin x_1$$

Find and classify the critical points of f. (This function is the height function of a torus with major radius 3 and minor radius 2).

123. Consider the constant coefficient differential operator

$$Lu \equiv au'' + bu' + cu, \quad (a, b, c \quad \text{real}, \quad a \neq 0.)$$

Let λ_1 and λ_2 denote the roots of the characteristic polynomial $p(\lambda) = a\lambda^2 + b\lambda + c$.

(a) If $\lambda_1 \neq \lambda_2$, find a formula for a particular solution of $Lu = f$.
[Answer: $u_p(x) = \dfrac{1}{\lambda_1 - \lambda_2} \displaystyle\int^x [e^{\lambda_1(x-t)} - e^{-\lambda_2(x-t)}] f(t)\,dt$.

(b) If λ_1 is complex, say, $\lambda_1 = \alpha + i\beta$, then $\lambda_2 = \bar{\lambda}_1 = \alpha - i\beta$. Show that in this case, the above formula simplifies to

$$u_p(x) = \frac{1}{\beta} \int^x e^{\alpha(x-t)} \sin\beta(x - t) f(t)\,dt.$$

(c) If $\lambda_1 = \lambda_2$, find a formula for a particular solution of $Lu = f$.
[Answer: $u_p(x) = \displaystyle\int^x (x - t)e^{\lambda_1(x-t)} f(t)\,dt$].

124. Consider $\displaystyle\iint_D f\,dA$ where D is the triangle with vertices at $(-1, 1)$, $(0, 0)$, and $(3, 1)$.

(a) Set up the iterated integrals in two ways.

(b) Evaluate one of the integrals in (a) for the integrand

$$f(x, y) = (x + y)^2.$$

125. When a double integral was set up for the mass M of a certain plate with density $f(x,y)$, the following sum of iterated integrals was obtained

$$M = \int_1^2 (\int_x^{x^3} f(x,y)\,dy)\,dx + \int_2^8 (\int_x^8 f(x,y)\,dy)\,dx.$$

(a) Sketch the domain of integration and express M as an iterated integral in which the order of integration is reversed.

(b) Evaluate M if

$$f(x,y) = \sqrt{\frac{x}{y}}.$$

126. Evaluate $\displaystyle\int_0^1 \int_0^1 x^y\,dx\,dy$.

127. It is difficult to evaluate the integral $I = \displaystyle\iint_D f\,dA$, where $f(x,y) = \frac{1}{1+x+y^2}$ and D is the indicated rectangle. However, you can show that (trivially)

$$\tfrac{1}{3} < I < \tfrac{3}{2},$$

and, with a bit more effort but the same method, that

$$\tfrac{1}{2} < I < \tfrac{3}{2}.$$

Please do so.

128. Consider the integral $I = \iint_D f\,dA$, where

$$f(x,h) = \frac{3}{8+\sqrt{x^4+y^4}}$$

and D is the domain inside the curve $x^4+y^4 = 16$. Show that

$$2\sqrt{2} < I < 6.$$

[Hint: Show that $\frac{1}{4} < f < \frac{3}{8}$ in D. Then approximate the area of D by an inscribed and circumscribed square. For the record, it turns out that $I = \frac{3A}{4}\ln(\frac{3}{2})$, where A is the

$$\text{area} = \frac{2}{\pi}\Gamma(\frac{1}{4})^2].$$

129. (a) Find the derivative matrix for the following mappings $Y = F(X)$ at the given point X_0.

(i) $F: \begin{cases} y_1 = x_1^2 + \sin x_1 x_2 \\ y_2 = x_2^2 + \cos x_1 x_2 \end{cases}$ at $X_0 = (0,0)$

(ii) $F: \begin{cases} y_1 = x_1^2 + x_3 e^{x_2} - x_2^3 \\ y_2 = x_1 - 3x_2 + x_1 \log x_3 \\ y_3 = x_2 + x_3 \\ y_4 = f x_1 x_2 x_3 \end{cases}$ at $X_0 = (2,0,1)$

(b) Find the equation of the tangent plane to the above surfaces at the given point.

130. Consider the following map F from $\mathbb{R}^2 \to \mathbb{R}^2$, the familiar change of variables from polar to rectangular coordinates.

$$F: \begin{cases} y_1 = x_1 \cos x_2 \\ y_2 = x_1 \sin x^2 \end{cases}$$

(a) Find the images of

(i) the semi-infinite strip $1 \le x_1 < \infty, \quad 0 \le x_2 \le \frac{\pi}{2}$.

(ii) the semi-infinite strip $0 \le x_1 < \infty, \quad 0 \le x_2 \le \frac{3\pi}{2}$.

(b) Compute F' and $\det F'$.

131. Given that $u_p(x) = e^{3x} + e^{-2x} - 2e^{x/2}$ is a solution of

$$au'' + bu' + cu = e^3 x,$$

find the constants a, b, and c.

132. Evaluate the determinants of the following matrices.

$$\text{(a)} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \end{pmatrix} \quad \text{(b)} \begin{pmatrix} 1 & 1 & y & z & t \\ 2 & x & z & t & y \\ w^2 & x^2 & 0 & 0 & 0 \\ w^3 & x^3 & 0 & 0 & 0 \\ w^4 & x^4 & 0 & 0 & 0 \end{pmatrix}.$$

133. For what value(s) of x is the following matrix invertible?

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2^2 & 2^3 \\ 1 & 3 & 3^2 & 3^3 \\ 1 & x & x^2 & x^3 \end{pmatrix}$$

(Hint: Observe that the determinant is a cubic polynomial all of whose roots are obvious).

134. Let $f(x) = \displaystyle\sum_{k=1}^n a_k \frac{\sin kx}{\sqrt{\pi}}$ and $g(x) = \displaystyle\sum_{r=1}^n b \frac{\sin r\,x}{\sqrt{\pi}}$.

By *direct integration* prove that

$$\int_{-\pi}^{\pi} f(x)g(x)\,dx = \sum_{j=1}^n a_j b_j.$$

After you are done, compare with Theorem 15, page 206-7 and its proof.

135. Let

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

(a) Find $\det A$.

(b) Find A^{-1}.

(c) Solve $AX = Y$, where $Y = \begin{pmatrix} 2 \\ 2 \\ 1 \\ 3 \end{pmatrix}$.

(d) Let $S = \{u\colon u(x) = ae^x + be^x + c\sin x + d\cos x\}$, where a, b, c, and d are any real numbers, and define a linear operator $L : S \to S$ by the rule

$$Lu \equiv u'' - u' + u.$$

Find the matrix ${}_eL_e$ for L with respect to the following basis for S:

$$e_1(x) = xe^x,\ e_2(x) = x,\ e_3(x) = \sin x,\ e_4(x) = \cos x.$$

(e) Use the above results to find a solution of

$$Lu = 2xe^x + 2e^x + \sin x + 3\cos x.$$

136. Let u_1 and u_2 be solutions of the homogeneous equation

$$Lu \equiv a_2(x)u'' + a_1(x)u' + a_0(x)u = 0.$$

(a) Show that $W(x) \equiv W(u_1, u_2)(x)$, the Wronskian of u_1 and u_2 satisfies the differential equation

$$W' = -\frac{a_1(x)}{a_2(x)}W.$$

(b) Find the equation of (a) for the particular operator

$$Lu \equiv x^2u'' - 2xu' + 2u$$

and solve it for W under the condition that $W(1) = 1$.

(c) Given that $u_1(x) = x$ is a solution of $Lu = 0$ for the operator of part (b), use the result of (b) to show that if u_2 is another solution of $Lu = 0$, then u_2 satisfies the equation

$$u_2' - \frac{1}{x}u_2 = x,$$

provided that $W(x, u_2)(1) = 1$.

(d) Solve the equation of part (c) under the assumption that $u_2(1) = 1$, and thus find a second independent solution of the equation $Lu = 0$ for the operator of part (b).

(e) Generalize the idea of parts (c) - (d) by stating and proving some theorem.

137. Here are some linear transformations defined in terms of matrices. In each case, describe geometrically what the transformation does, by computing the images of the three parallelograms

Q_1 : with vertices at $(0,0), (2,0), (3,1), (1,1)$.

Q_2 : with vertices at $(1,2), (3,2), (4,3), (2,3)$.

Q_3 : with vertices at $(1,0), (0,2), (-1,0), (0,-2)$.

(a) *Diagonal Maps* (Stretchings)

$$L_1 = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}, \quad L_2 = \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix}, \quad L_3 = \begin{pmatrix} -4 & 0 \\ 0 & 6 \end{pmatrix},$$

$$L_4 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad L_5 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad L_6 = \begin{pmatrix} -2 & 0 \\ 0 & 0 \end{pmatrix},$$

$$L_7 = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}, \quad L_8 = \begin{pmatrix} 1 & 0 \\ 0 & b \end{pmatrix}, \quad L_9 = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix},$$

(Remember to consider negative values of a and b).

(b) *Maps with 0 on the diagonal.*

$$L_1 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad L_2 = \begin{pmatrix} 0 & a \\ 0 & 0 \end{pmatrix}, \quad L_3 = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix},$$

$$L_4 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad L_5 = \begin{pmatrix} 0 & a \\ 1 & 0 \end{pmatrix}, \quad L_6 = \begin{pmatrix} 0 & a \\ b & 0 \end{pmatrix}.$$

(c) *Upper Triangular Matrices.*

$$L_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad L_2 = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}, \quad L_3 = \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix},$$

$$L_4 = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}, \quad L_5 = \begin{pmatrix} -1 & -1 \\ 0 & 0 \end{pmatrix}, \quad L_6 = \begin{pmatrix} a & 1 \\ 0 & b \end{pmatrix}.$$

(d) *Orthogonal Matrices* (Rotations and Reflections).

$$L_1 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad L_3 = \begin{pmatrix} \frac{3}{5} & \frac{4}{5} & \\ L_2 & = \frac{4}{5} & -\frac{3}{5} \end{pmatrix} \quad L_4 = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

138. Let a and b be real numbers such that $a^2 + b^2 = 1$. Let

$$S = \begin{pmatrix} a^2 - b^2 & 2ab \\ 2ab & b^2 - a^2 \end{pmatrix}, \quad P = \begin{pmatrix} a^2 & ab \\ ab & b^2 \end{pmatrix},$$

and let $e_1 = (a, b), \quad e_2 = (-b, a)$, so $e_1 \perp e_2$. Show that

(a) $Se_1 = e_2, \qquad Pe_1 = e_1$

(b) $Se_2 = -e_2, \qquad Pe_2 = 0$

(c) $S^2 = I, \qquad P^2 = P$

(d) Show that S can be interpreted as the reflection which leaves the line through e_1 fixed, and that P can be interpreted as the projection onto the line through e_1 parallel to e_2.

139. (a) Consider the following relation defined on the set of *all* integers: $n\mathcal{R}m$ if n and m are both even integers. Verify that this relation is symmetric and transitive - but not reflexive (since, for example, $1 \not\mathcal{R} 1$).

(b) Let $\mathcal{R}$ be a symmetric and transitive relation defined on a set A. If, given any element x in A, there is some element y related to it, $x\mathcal{R}y$, prove that the relation $\mathcal{R}$ is also reflexive. (The example in part (a) shows that the assertion will be false if some element is related to no others).

140. Let a_n be a decreasing sequence of positive real numbers which satisfy $a_{n-1}a_{n+1} \le a_n^2$. If $\sum a_n^{1/n}$ converges, prove that $\sum \frac{a_n}{a_{n-1}}$ converges too. [Hint: Show that $(a_n/a_{n-1})^{1/n} \le a_n$].

141. (a) Prove that the series $\sum a_n z^n$ and $\sum a_n^2 z^n$ have the same radii of convergence.

(b) Prove that the series $\sum a_n z^n$ and $\sum (a_n)^{k_z n}$, where $k > 0$, have the same radii of convergence.

142. Let V be a linear space and L an invertible linear map, $L : V \to V$. If $\{e_1, \dots, e_n\}$ is a basis for V, prove that its image $\{Le_1, Le_2, \dots, Le_n\}$ is also a basis for V.

143. Let H be an inner product space and R an orthogonal transformation, $R : H \to H$. If $\{e_1, \dots, e_n\}$ is a complete orthonormal set for H, prove that its image $\{Re_1, \dots, Re_n\}$ is also a complete orthonormal set for H.

144. (a) Let R be an orthogonal matrix and let ρ_1 and ρ_2 be any two of its column vectors. Prove that $\rho_1 \perp \rho_2$. Prove that any two rows of an orthogonal matrix are also orthogonal to each other.

(b) Conversely, let A be a square matrix whose column vectors are orthogonal. Must A be an orthogonal matrix? Proof or counterexample.

145. Let H be an inner product space and A the subspace of H spanned by the vectors $X_1, \dots, X_n$. The *Gram determinant* of those vectors is defined as

$$G(X_1, \dots, X_n) = \begin{vmatrix} \langle X_1, X_1 \rangle & \cdots & \langle X_n, X_1 \rangle \\ \langle X_1, X_2 \rangle & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \langle X_1, X_n \rangle & \cdots & \langle X_n, X_n \rangle \end{vmatrix}$$

(a) Prove that $X_1, \cdots, X_n$ are linearly dependent if and only if $G(X_1, \cdots, X_n) = 0$. [Suggestion: If $Z \in A$, then $Z = a_1 X_1 + \cdots a_n X_n$, where the scalars $a_1, \cdots, a_n$ are to be found. This can be done in two ways, by Theorem 31, page 428, or by solving the n equations

$$\begin{aligned} \langle Z, X_1 \rangle &= a_1 \langle X_1, X_1 \rangle + \cdots + a_n \langle X_n, X_1 \rangle \\ &\cdot \\ &\cdot \\ &\cdot \\ &\cdot \\ \langle Z, X_n \rangle &= a_1 \langle X_1, X_n \rangle + \cdots + a_n \langle X_n, X_n \rangle \end{aligned}$$

which are obtained from $\langle Z, X_j \rangle = \langle a_i X_1 + \cdots + a_n X_n, X_j \rangle$. Couple both methods to prove the result].

(b) If $X_1, \cdots, X_n$ are an orthogonal set of vectors, compute $G(X_1, \cdots, X_n)$.

(c) If $Y \in H$, prove that the distance of Y from the subspace A, $\quad \|Y - P_A Y\| = \delta$, is given by the formula

$$\delta^2 = \|Y - P_A Y\|^2 = \frac{G(Y, X_1, \dots, X_n)}{G(X_1, \dots, X_n)}.$$

[Suggestion: Observe that $\delta^2 = \|Y - P_A Y\|^2 = \langle Y - P_A Y, Y \rangle$ and that $\langle P_A Y, Y \rangle =$ $a_n \langle X_1, Y \rangle + \cdots + a_n \langle X_n, Y \rangle$. Now write $P_A Y$ as Z, use the n equations in a) and the one equation $\delta^2 = \langle Y, Y \rangle - a_1 \langle X_1, Y \rangle - \cdots - a_n \langle X_n, Y \rangle$ to solve for δ^2 by using Cramer's rule].

(d) Use the fact that $G(X_1) = \langle X_1, X_1 \rangle$ to prove the Gram determinant of linearly independent vectors is always *positive*. In particular, deduce the Cauchy - Schwarz inequality from $G(X_1, X_2) \ge 0$.

(e) In $L_2[0,1]$, let $X_1 = 1 + x$, and $X_2 = x^3$. Compute $G(X_1, X_2)$. Let $Y = 2 - x^4$ and compute $\|Y - P_A Y\|$, where A is the subspace spanned by X_1 and X_2.

(f) (Muntz) In $L_2[0,1]$, let $A_n = \text{span}\{x^{j_1}, x^{j_2}, \cdots, x^{j_n}\}$ where $j_1, \cdots, j_n$ are distinct positive integers. Let $Y = x^k$, where k is a positive integer by not one of the j's. Prove that $\lim_{n\to\infty} \|Y - P_{A_n} Y\| = 0$ if and only if $\sum \frac{1}{j_n}$ diverges.

146. (a) Use Theorem 17, page 217 to find linear polynomials P and Q such that, respectively,

   (i) $\int_{-1}^{1}[x^2 - P(x)]^2\,dx$ is minimized,

   (ii) $\int_{0}^{1}[x^2 - Q(x)]^2\,dx$ is minimized.

   (b) Write $P(x) = a + bx$ and use calculus to again find the values of a and b such that
$$\int_{-1}^{1}[x^2 - P(x)]^2\,dx$$
is minimized.

147. Let $Z = (1,1,1,1,1) \in \mathbb{R}^5$ and let A be the subspace of $\mathbb{R}^5$ spanned by $X_1 = (1,0,1,0,0)$, $X_2 = (1,0,0,-1,0)$, and $X_3 = (0,1,0,0,1)$. Find $\|Z - P_A Z\|$.

148. Let Γ_0 be a closed planar curve which encloses a convex region, and let Γ_r be the "parallel" curve obtained by moving out a distance of r along the outer normal.

   (a) Discover a formula relating the arc length of Γ_r to that of Γ_0. [Advise: Examine the special cases of a circle, rectangle, and convex polygon].

   (b) Prove the result you conjectured in part a).

149. The *hypergeometric function* $F(a,b;c;x)$ is defined by the power series
$$F(a,b;c;x) = 1 + \frac{a\cdot b}{1\cdot c}x + \frac{a(a+1)b(b+1)}{1\cdot 2\cdot c(c+1)}x^2 + \frac{a(a+1)(a+2)b(b+1)(b+2)}{1\cdot 2\cdot 3c(c+1)(c+2)}x^2 + \cdots$$

   (a) Show that the series converges for all $|x| < 1$.

   (b) Show that $\frac{d}{dx}F(a,b;c;x) = \frac{ab}{c}F(a+1,b+1;c+1;x)$.

   (c) Show that

   (i) $(1-x)^n = F(-n,b;b;x)$

   (ii) $(1+x)^n = F(-n,b;b;-x)$

   (iii) $\log(1-x) = -xF(1,1;2,x)$

   (iv) $\log(\frac{1+x}{1-x}) = 2xF(\frac{1}{2},1;\frac{3}{2};x^2)$

   (v) $e^x = \lim_{b\to\infty} F(1,b;1;x/b)$

   (vi) $\cos x = F(\frac{1}{2},-\frac{1}{2};\frac{1}{2},\sin^2 x)$

   (vii) $\sin^{-1} x = xF(\frac{1}{2},\frac{1}{2};\frac{3}{2};x^2)$

   (viii) $\tan^{-1} x = xF(\frac{1}{2},1;\frac{3}{2};-x^2)$

   (d) Show that F satisfies the hypergeometric differential equation
$$x(1-x)\frac{d^2F}{dx^2} + [c-(a+b+1)x]\frac{dF}{dx} - ab\,F = 0.$$
[This equation is essentially the most general one with three regular singular points - in this case located at $0, 1$, and ∞].

150. Let $\{e_1,\cdots,e_n\}$ be a complete orthonormal set of $\mathbb{R}^n$ and let $\{X_1,\cdots,X_n\}$ be a set of vectors which are close to the e_j's in the sense that
$$\sum_{j=1}^{n}\|X_j - e_j\|^2 < 1.$$
Prove that the X_j's are linearly independent. Give an example in $\mathbb{R}^3$ of linearly dependent vectors $\{X_1, X_2, X_3\}$ which satisfy
$$\sum_{j=1}^{n}\|X_j - e_j\|^2 = 1.$$
[In fact, one can prove that
$$\dim A^{\perp} \le \sum_{j=1}^{n}\|X_j - e_j\|^2,$$
] where $A = \text{span}\{X_1,\cdots,X_n\}$].

151. (a) Show that the function $f(z) = e^z$, $z \in C$, is never zero.

   (b) Scrutinize the proof of the Fundamental Theorem of Algebra (pp. 544-548) and find where it breaks down if one attempts to extend it to prove that e^z has at least one zero.