

ISP

Mathematics 281

Leonard Evens
Department of Mathematics
Northwestern University

© Leonard Evens 1992, 1993



This work is licensed under the Creative Commons By-Attribution Share-Alike 3.0 Unported license, giving you permission to share and modify this work provided you attribute appropriately and license any new work in a compatible way.

Preface

This text has been specially created for the first year mathematics course of the Integrated Science Program at Northwestern University. Some of what we cover will be used in other first year courses such as physics. Such topics will generally be introduced before they are needed in the other courses and in a manner which emphasizes their relation to the subject matter of those courses. For the most part, the rest of the subject matter is mathematics which is used in later courses in the program, but on rare occasions we shall discuss some points which are mainly of interest to mathematicians. Overall, the perspective is that of a mathematician, which sometimes looks a bit different from that of a physicist or chemist. Mathematicians will tend to emphasize care in formulation of concepts and the need to prove mathematical statements by rigorous arguments, while scientists may concentrate on the physical content of such statements. You should be aware, however, that the underlying concepts are often the same, and you should make sure you understand how ideas introduced in your different courses are related.

It is assumed that the student has mastered differential and integral calculus of one variable as taught, for example, in a typical high school advanced placement calculus course. A reasonable reference for this material is Edwards and Penney's *Calculus and Analytic Geometry*, 3rd edition or any similar calculus text.

How to learn from this text

You should try to work *all* the problems except those marked as optional. You may have some trouble with some of the problems, but ultimately after discussions with fellow students and asking questions of your professor or teaching assistant, you should understand how to do them. You should write up the solutions and keep them for review. Some sections are clearly marked as optional, and your professor will indicate some others which are not part of the course. Such sections may contain proofs or other special topics which may be of interest to you, so you should look at these sections to decide if you want to study them. Some of these sections include material which you may want to come back to in connection with more advanced courses.

Use of computers

You will be learning about computer programming in a separate course. On occasion, you will be expected to make use of what you learn there to help you understand some mathematical point. Also, you will have various computer resources available to help you visualize some of the subject matter of the course, e.g., to graph curves and surfaces. You should learn to make use of these resources.

A note on indefinite integrals

A brief comment on the treatment of indefinite integrals may be helpful. You probably spent considerable time in your previous calculus course learning how to calculate indefinite

integrals (also called antiderivatives). That experience is useful in later work in so far as it gives you some perspective on what is involved in integrating. However, for the most part, finding indefinite integrals is a relatively mechanical process on which one does not want to spend a lot of time. When you encounter an integral, if you don't remember how to do it right away, you should normally either look it up in a table of integrals or use a symbolic manipulation program (such as Maple or Mathematica) to find it. Of course, there are some occasions where that won't suffice or where you get the wrong answer, so your previous mastery of the subject will have to be brought to bear, but that will be unusual. In this text, we have tried to encourage you in this attitude by letting Mathematica provide indefinite integrals wherever possible. Some students object to this 'appeal to authority' for an answer you can derive yourself. Its justification is that time is short and best reserved for less routine tasks.

Acknowledgments

Many of the problems were inspired by problems found in other introductory texts. In particular, *Edwards and Penney* was used as a source for many of the calculus problems, and Braun's *Differential Equations and Their Applications*, 3rd edition was used as a source for many of the differential equations and linear algebra problems. The problems weren't literally copied from those texts, and these problems are typical of problems found in many such texts, but to the extent that original ideas of the above authors were adapted, they certainly deserve the credit. Most of the treatment of calculus and linear algebra is original (to the extent that one can call any treatment of such material original), but parts of the treatment of differential equations, particularly systems of differential equations were inspired by the approach of Braun.

I should like to thank Jason Jarzembowski who compiled most of the problems for Chapters I through V. Michael R. Stein, Integrated Science Program Director, ensured that the text would come to fruition by allocating needed resources to that end and by exhorting others to get it done. Finally, I should like to thank my teaching assistant, John Gately, who helped with the development of problem sets and timely printing of the text.

Leonard Evens, July, 1992

This document was originally typeset in $\mathcal{A}\mathcal{M}\mathcal{S}$ - $\text{\textsf{TEX}}$. It was re-typeset in $\text{\textsf{L}^A\text{\textsf{T}}_E\text{\textsf{X}}$ by Jason Siefken in 2015.

Contents

Part I

Vector Calculus

Chapter 1

Vectors

1.1 Introduction

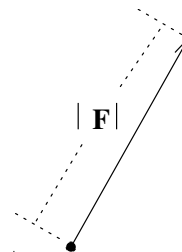
A *vector* is a quantity which is characterized by a *magnitude* and a *direction*. Many quantities are best described by vectors rather than numbers. For example, when driving a car, it may be sufficient to know your speed, which can be described by a single number, but the motion of an airplane must be described by a vector quantity called velocity which takes into account its direction as well as its speed.

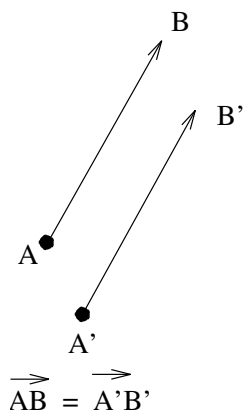
Ordinary numerical quantities are called *scalars* when we want to emphasize that they are not vectors.

In print, vectors are almost always denoted by bold face symbols, e.g., \mathbf{F} , but when written, one uses a variety of mechanisms for distinguishing vectors from scalars. One common such notation is \vec{F} . The magnitude of a vector \mathbf{F} is denoted $|\mathbf{F}|$. Indicating its direction symbolically is a bit more difficult, and we shall discuss that later.

Vectors are represented *geometrically* by *directed line segments*. The length of the line segment is the magnitude of the vector and its direction is the direction of the vector.

Note that parallel line segments of the *same length* and *same direction* represent the *same vector*. In drawing pictures, there is a tendency to identify a directed line segment with the vector it represents, but this can lead to confusion about how vectors are used. In general, there are infinitely many directed line segments which could be used to represent the same vector. You should always remember that *you are free to put the tail end of a vector at any point which might be convenient for your purposes*. We shall often use the notation \overrightarrow{AB} for the vector represented by





the directed line segment from A to B .

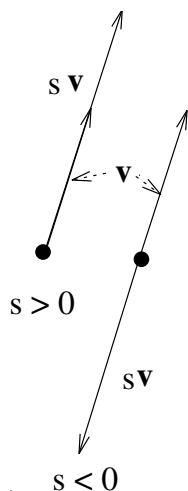
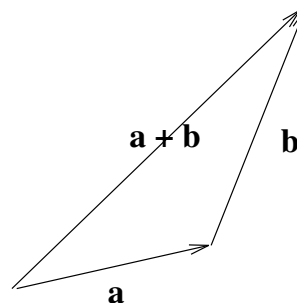
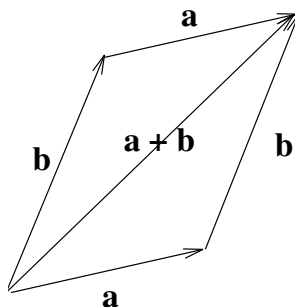
Many of the laws of physics relate vector quantities. For example, Newton's Second Law

$$\mathbf{F} = m\mathbf{a}$$

relates two vector quantities: force \mathbf{F} and acceleration \mathbf{a} . The constant of proportionality m is a scalar quantity called the mass.

Operations with vectors It is a familiar experience that when velocities are combined, their magnitudes do not add. For example, if an airplane flies North at 600 mph against an east wind of 100 mph, the resultant velocity will be somewhat west of north and its magnitude certainly won't be 700 mph. To calculate the correct velocity in this case, we need *vector addition*, which is defined geometrically as follows. Suppose \mathbf{a}, \mathbf{b} are vectors. To add them, choose directed line segments representing them, with tails at the same point, and consider the vector represented by the diagonal of the resulting parallelogram.

We call that vector $\mathbf{a} + \mathbf{b}$. This is called the parallelogram law of addition. There is also another way to do the same thing called the triangle law of addition. See the diagram.



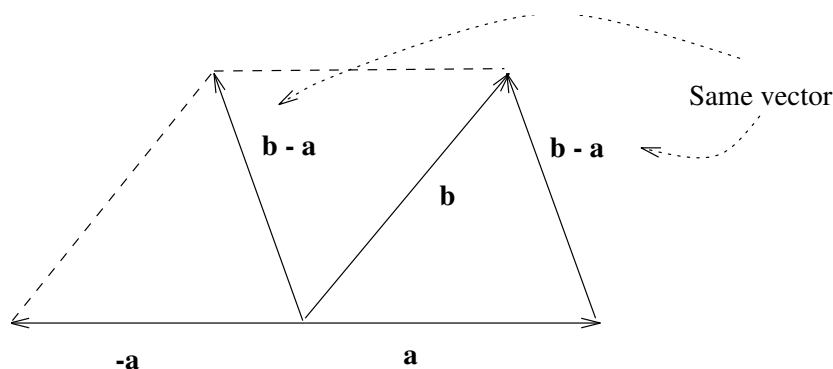
Note that for the triangle law, we place the tail of one directed line segment on the head of the other, taking advantage of freedom to place the tail where needed. If the vectors \mathbf{a} and \mathbf{b} have the same or opposite directions, then the diagrams become degenerate (collinear) figures. (You should draw several cases for yourself to make sure you understand.)

As we saw in Newton's Law, we sometimes want to *multiply* a vector by a scalar. This is defined as follows. If s is a scalar and \mathbf{v} is a vector, then $s\mathbf{v}$ is the vector with magnitude $|s||\mathbf{v}|$ and its direction is either the same as that of \mathbf{v} , if $s > 0$, or opposite to \mathbf{v} , if $s < 0$.

Note that the above definition omits the case that $s = 0$. In that case, we run into a problem, at least if we represent vectors by directed line segments. It does not make

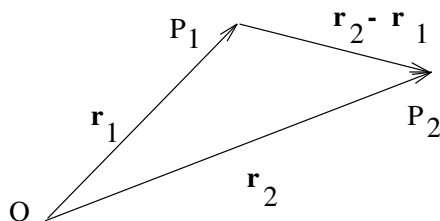
sense to talk of the direction of a directed line segment with length zero. For this reason, we introduce a special quantity we call the *zero vector* which has magnitude zero and no well defined direction. It is not a vector as previously defined, but we allow it as a degenerate case which is needed so operations with vectors will make sense in all cases. The zero vector is denoted by a bold face zero $\mathbf{0}$ or in writing by $\vec{0}$. With that notation, $s\mathbf{v} = \mathbf{0}$ if $s = 0$. Note also that a degenerate case of the parallelogram law yields $\mathbf{v} + \mathbf{0} = \mathbf{v}$ for any vector \mathbf{v} .

You can *subtract* a vector \mathbf{a} from a vector \mathbf{b} by adding the opposite vector $-\mathbf{a} = (-1)\mathbf{a}$. Study the accompanying diagram for some idea of how $\mathbf{b} - \mathbf{a}$ might be represented geometrically.



Note that if \mathbf{a} and \mathbf{b} are represented by directed line segments with the same tail, then the line segment from the end of \mathbf{a} to the end of \mathbf{b} represents $\mathbf{b} - \mathbf{a}$.

In the study of motion, which is of interest both in physics and in mathematics, it is common to use the so called *position vector* (also called *radius vector*). Thus, if a particle moves on a path, its position can be specified as the endpoint P of a directed line segment from a common *origin* O . The position vector is $\mathbf{r} = \overrightarrow{OP}$. In this context, we are often interested in comparing the position vectors \mathbf{r}_1 and \mathbf{r}_2 of the same particle at different times (or of two different particles). As the diagram indicates, the difference $\mathbf{r}_2 - \mathbf{r}_1$ is associated with the directed line segment from the first position to the second position. This is called the *displacement vector*, and it would be the same for any choice of common origin O .



The operations defined above satisfy the usual laws of algebra. Here are some of them

$$(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c}) \quad \text{Associative Law}$$

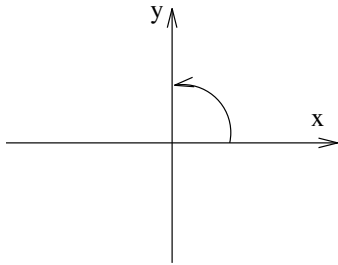
$$\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a} \quad \text{Commutative Law}$$

$$s(\mathbf{a} + \mathbf{b}) = s\mathbf{a} + s\mathbf{b}$$

$$(s + t)\mathbf{a} = s\mathbf{a} + t\mathbf{a}$$

$$(st)\mathbf{a} = s(t\mathbf{a})$$

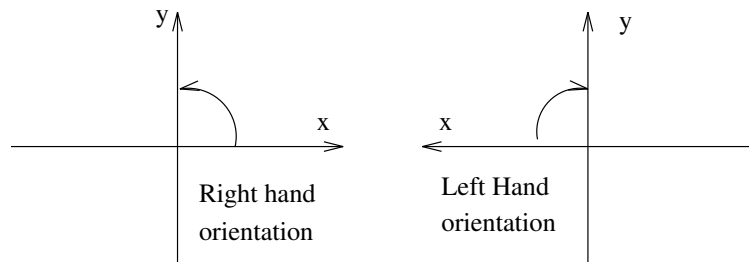
There are several others which you can probably invent for yourself. These rules are not too difficult to check from the definitions we gave, but we shall skip that here. (You might try doing it yourself by drawing the appropriate diagrams.)



Components The geometric definitions above result in nice pictures and nourish our intuition, but they are quite difficult to calculate with. To make that easier, it is necessary to introduce coordinate systems and thereby reduce geometry to algebra and arithmetic. We start with the case of vectors in the plane since it is easier to visualize, but of course to deal with the the real world, we shall have to also extend our notions to space.

Recall that a coordinate system in the plane is specified by choosing an origin O and then choosing two perpendicular axes meeting at the origin. These axes are chosen in some order so that we know which axis (usually the x -axis) comes first and which (usually the y -axis) second. Note that there are many different coordinate systems which could be used although we often draw pictures as if there were only one.

In physics, one often has to think carefully about the coordinate system because choosing it appropriately may greatly simplify the resulting analysis. Note that the axes are usually drawn with *right hand orientation* where the right angle from the positive x -axis to the positive y -axis is in the counter-clockwise direction. However, it would be equally valid to use the *left hand orientation* in which that angle is in the clockwise direction. One can easily switch the orientation of a coordinate system by reversing one of the axes. (The concept of orientation is quite fascinating and it arises in mathematics, physics, chemistry, and even biology in many interesting ways. Note that almost all of us base our intuitive concept of orientation on our inborn notion of “right” versus “left”.)

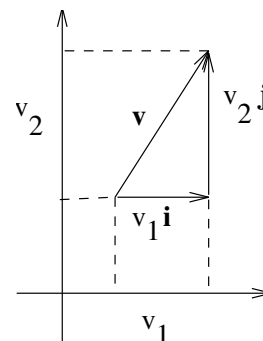


Given a vector \mathbf{v} , the projections—of any directed line segment representing it—onto the coordinate axes are called the *components* of the vector. The components of \mathbf{v} are often displayed as an ordered pair $\langle v_1, v_2 \rangle$ where each component is numbered according to the axis it is associated with. Another common notation is $\langle v_x, v_y \rangle$. We shall use both.

Notice that the Pythagorean Theorem tells us that

$$|\mathbf{v}| = \sqrt{v_1^2 + v_2^2}.$$

Our notation distinguishes between the coordinates (x, y) of a point P in the plane, and the components $\langle v_x, v_y \rangle$ of a vector. This is to emphasize the difference between a vector and a point. That distinction is a bit esoteric, particularly in the analytic context, since a pair of numbers is just that no matter what notation we use. Hence, you will find that many mathematics and physics books make no such distinction. (Review in your mind the distinction between a point P and the position vector \overrightarrow{OP} . What relation exists between the coordinates of P and the components of \overrightarrow{OP} ?)



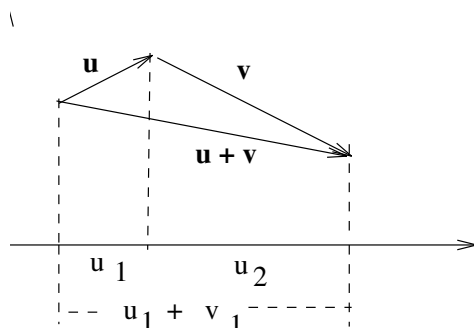
There is another common way to indicate components. Let \mathbf{i} denote a vector of length 1 pointing in the positive direction along the x -axis, and let \mathbf{j} denote the corresponding unit vector for the y -axis. (These are also commonly written $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$.) Then the diagram makes clear that

$$\mathbf{v} = v_1 \mathbf{i} + v_2 \mathbf{j}.$$

Vector operations are mirrored by corresponding operations for components. Thus, if \mathbf{u} has components $\langle u_1, u_2 \rangle$ and \mathbf{v} has components $\langle v_1, v_2 \rangle$, then

$$\begin{aligned} \mathbf{u} + \mathbf{v} &= u_1 \mathbf{i} + u_2 \mathbf{j} + v_1 \mathbf{i} + v_2 \mathbf{j} \\ &= (u_1 + v_1) \mathbf{i} + (u_2 + v_2) \mathbf{j} \end{aligned}$$

from which we conclude that the components of $\mathbf{u} + \mathbf{v}$ are $\langle u_1 + v_1, u_2 + v_2 \rangle$. (The diagram below exhibits a more geometric argument.



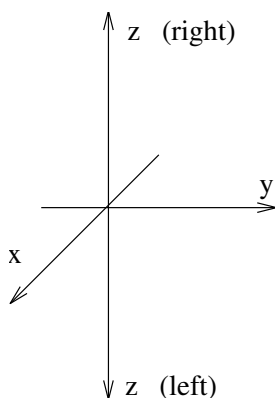
(Note that I only drew one of the many diagrams needed to handle all possible cases.) In words, we may restate the rule as follows

the components of the sum of two vectors are the sums of the components of the vectors.

A similar argument shows that the components of $s\mathbf{v}$ are $\langle sv_1, sv_2 \rangle$, i.e.,

the components of a scalar multiple of a vector are that multiple of the components of the vector.

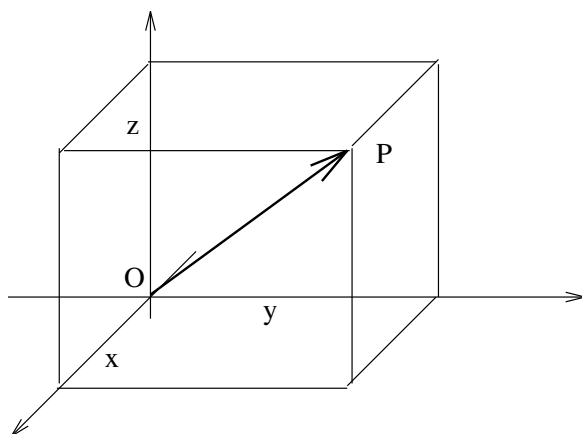
The same sort of considerations apply in space. To set up a coordinate system, we first choose an origin O , and then choose, in a some order, three mutually perpendicular axes through O , each with a specified positive direction. These are usually called the x -axis, the y -axis, and the z -axis. Since any two *intersecting* lines in space determine a plane, we can think of the first two axes generating an x, y -plane. Since the z -axis must be perpendicular to this plane, the line along which it lies is completely determined, but we have two possible choices for its positive direction. (See the diagram.)



A set of axes in space has the *right hand orientation* if when you point the fingers of your right hand from the positive x -axis to the positive y -axis, your upright thumb points in the direction of the positive z -axis. Otherwise, it has the *left hand orientation*. As in the plane case, reversing the direction of one axis reverses the orientation. Almost all authors today use the right hand orientation for coordinate axes.

Given a set of coordinate axes, a point P in space is assigned coordinates (x, y, z) as follows. Let the origin O and the point P be opposite vertices of a rectangular box.

The coordinates x, y , and z are the (signed) magnitudes of the sides of this box. Points with $x > 0$ are in front of the y, z -plane, and points with $x < 0$ are in back of that plane. You should think out the possibilities for the signs of y and z . A point has coordinate $x = 0$ if and only if it lies in the y, z -plane (in which case the “box” is degenerate). Similarly, $y = 0$ characterizes the x, z -plane and $z = 0$ characterizes the x, y -plane.



Our previous discussion of vectors generalizes in a more or less obvious way to space. The components $\langle v_1, v_2, v_3 \rangle$ (sometimes $\langle v_x, v_y, v_z \rangle$) of a vector \mathbf{v} are obtained by projecting a directed line segment representing the vector onto each of the coordinate axes.

As before, we have

$$|\mathbf{v}| = \sqrt{v_1^2 + v_2^2 + v_3^2}.$$

(Look at the diagram for an indication of why this is so. It requires two applications of the Pythagorean Theorem.) In addition, the same rules for components of sums and scalar multiples apply as before except, of course, there is one more component to worry about.

In space, in addition to \mathbf{i} and \mathbf{j} , we have a third unit vector \mathbf{k} pointing along the positive z -axis, and any vector can be resolved

$$\mathbf{v} = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$$

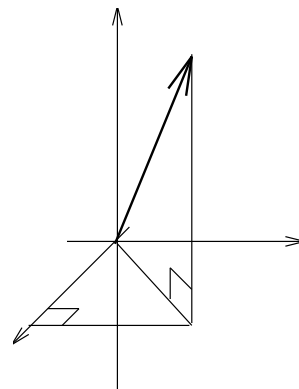
in terms of its components.

Often, to save writing, we shall not distinguish between a vector and its components, and we will write

$$\begin{aligned} \mathbf{v} &= \langle v_1, v_2 \rangle && \text{in the plane,} \\ \mathbf{v} &= \langle v_1, v_2, v_3 \rangle && \text{in space.} \end{aligned}$$

This is an ‘abuse of notation’ since a vector is not the same as its set of components, which generally *depend on the choice of coordinate axes*. But, it is a convenient notation, and you usually won’t get in trouble if you use it with discretion.

Higher dimensions and \mathbf{R}^n One can’t progress very far in the study of science and mathematics without encountering a need for higher dimensional “vectors”. For example, physicists have known since Einstein that the physical universe is best



thought of as a 4-dimensional entity called spacetime in which time plays a role close to that of the 3 spatial coordinates. Since, we don't have any way to deal intuitively with any higher dimensional geometries, we must proceed by analogy with two and three dimensions, and the easiest way to proceed is to generalize the analytic approach by adding additional coordinates. Thus, in general, we consider n -tuples

$$(x_1, x_2, \dots, x_n)$$

where n can be any positive integer. The collection of all such n -tuples is denoted \mathbf{R}^n , where the \mathbf{R} refers to the fact that the entries (coordinates) are supposed to be real numbers. From this perspective, it doesn't make a whole lot of sense to distinguish points from vectors, so the two terms are often used interchangeably. Vector operations in \mathbf{R}^n may be *defined* in terms of components.

$$\begin{aligned} |(x_1, x_2, \dots, x_n)| &= \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}, \\ (x_1, x_2, \dots, x_n) + (y_1, y_2, \dots, y_n) &= (x_1 + y_1, x_2 + y_2, \dots, x_n + y_n), \\ s(x_1, x_2, \dots, x_n) &= (sx_1, sx_2, \dots, sx_n). \end{aligned}$$

The case $n = 1$ yields "geometry" on a line, the cases $n = 2$ and $n = 3$ geometry in the plane and in space, and the case $n = 4$ yields the geometry of "4-vectors" which are used in the special theory of relativity. Larger values of n are used in a variety of contexts, some of which we shall encounter later in this course.

Exercises for 1.1.

- Find $|\mathbf{a}|$, $5\mathbf{a} - 2\mathbf{b}$, and $-3\mathbf{b}$ for each of the following vector pairs:
 - $\mathbf{a} = 2\mathbf{i} + 3\mathbf{j}$, $\mathbf{b} = 4\mathbf{i} - 9\mathbf{j}$
 - $\mathbf{a} = \langle 1, 2, -1 \rangle$, $\mathbf{b} = \langle 2, -1, 0 \rangle$
 - $\mathbf{a} = \mathbf{0}$, $\mathbf{b} = \langle 2, 3, 4 \rangle$
 - $\mathbf{a} = 3\mathbf{i} - 4\mathbf{j} + \mathbf{k}$, $\mathbf{b} = \mathbf{k}$
 - $\mathbf{a} = \langle \cos t, \sin t \rangle$, $\mathbf{b} = \langle -\sin t, \cos t \rangle$
- Find the vector of the form $a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$ that is represented by an arrow from the point $P(7, 2, 9)$ to the point $Q(-2, 1, 4)$.
- Find unit vectors with the same direction as the vectors (a) $\langle -4, -2 \rangle$, (b) $3\mathbf{i} + 5\mathbf{j}$, (c) $\langle 1, 3, -2 \rangle$.
- Show that if \mathbf{v} is any non-zero vector, then $\mathbf{u} = \mathbf{v}/|\mathbf{v}|$ is a unit vector. Hint: Use the following property for the magnitude of vectors, $|s\mathbf{v}| = |s||\mathbf{v}|$. (Note that $|s|$ means the absolute value of the scalar s , but $|\mathbf{v}|$ and $|s\mathbf{v}|$ mean the magnitudes of the indicated vectors).
- Show by direct calculation that the rule $\overrightarrow{AB} + \overrightarrow{BC} + \overrightarrow{CA} = \mathbf{0}$ holds for the three points $A(2, 1, 0)$, $B(-4, 1, 3)$, and $C(0, 12, 0)$. Can you prove the general rule for any three points in space?

6. Show that if a vector \mathbf{v} in the plane has components $\langle v_x, v_y \rangle$ then the scalar multiple $s\mathbf{v}$ has components $\langle sv_x, sv_y \rangle$.
7. Use Newton's Second Law, $\mathbf{F} = m\mathbf{a}$, to find the acceleration of a 10 kg box if a 120 N force is applied in the horizontal direction. Draw the vector diagram. Are \mathbf{F} and \mathbf{a} in the same direction? Is this always true? Why? What if the force were applied at a 30 angle to the horizontal?
8. If an airplane flies with apparent velocity \mathbf{v}_a relative to air, and the wind velocity is denoted \mathbf{w} , then the planes true velocity relative to the ground, is $\mathbf{v}_g = \mathbf{v}_a + \mathbf{w}$. Draw the diagram to assure yourself of this.
 - (a) A farmer wishes to fly his crop duster at 80 km/h north over his fields. If the weather vane atop the barn shows easterly winds at 10 km/h, what should his apparent velocity, \mathbf{v}_a be?
 - (b) What if the wind were northeasterly? southeasterly?
9. Suppose a right handed coordinate system has been set up in space. What happens to the orientation of the coordinate system if you make the following changes? (a) Change the direction of one axis. (b) Change the direction of two axes. (c) Change the direction of all three axes. (d) Interchange the x and y axes.
10. In body-centered crystals, a large atom (assumed spherical) is surrounded by eight smaller atoms. If the structure is placed in a "box" that just contains the large atom, the smaller atoms each occupy a corner. If the central atom has radius R , what is the greatest atomic radii the smaller atoms may have?
11. Prove that the diagonals of a parallelogram bisect each other. (Hint: show that the position vectors from the origin to the midpoints are equal).

1.2 Kinematics and Vector Functions

In physics, you will study the motion of particles, so we need to investigate here the associated mathematics.

Example 1 You probably learned in high school that the path of a projectile moving under the force of gravity near the earth is pretty close to a parabola.

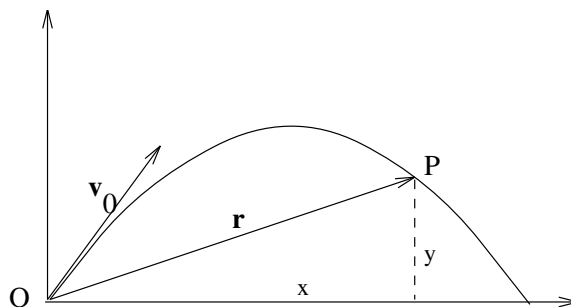
If we choose coordinates x (for horizontal displacement) and y (for vertical displacement), the path may be described by the equations

$$\begin{aligned}x &= v_{x0}t \\ y &= v_{y0}t - \frac{1}{2}gt^2,\end{aligned}$$

where t denotes time, v_{x0} and v_{y0} are the components of a vector \mathbf{v}_0 called the *initial velocity*, and g is a constant giving the acceleration of gravity at the surface

of the Earth. This can be simplified if we combine the two coordinate functions of t in a single vector quantity

$$\mathbf{r} = x\mathbf{i} + y\mathbf{j} = (v_{x0}t)\mathbf{i} + (v_{y0}t - \frac{1}{2}gt^2)\mathbf{j}.$$



We can think of this as expressing \mathbf{r} as a *single vector function* of time t : $\mathbf{r} = \mathbf{r}(t)$. (Note that \mathbf{r} is just the position vector connecting the origin to the position of projectile at time t .)

In general, we can think of any vector function $\mathbf{r} = \mathbf{r}(t)$ as describing the *path* of a particle as it moves through space. For such a function we will have

$$\mathbf{r} = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k} \quad (1)$$

so giving a vector function is equivalent to giving three scalar functions $x(t)$, $y(t)$, and $z(t)$. The single vector equation (1) can also be written as three scalar equations

$$x = x(t)$$

$$y = y(t)$$

$$z = z(t)$$

If, as in Example 1, the motion is restricted to a plane, then, with suitably a chosen coordinate system, we may omit one coordinate, e.g., we may write $\mathbf{r} = x(t)\mathbf{i} + y(t)\mathbf{j}$, or $x = x(t), y = y(t)$.

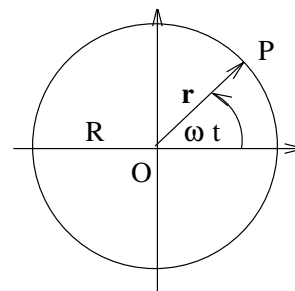
Example 2. Uniform Circular Motion Let

$$x = R \cos \omega t$$

$$y = R \sin \omega t.$$

The (end of) the position vector $\mathbf{r} = x\mathbf{i} + y\mathbf{j}$ traces out a circle of radius R centered at the origin.

To see this note that $x^2 + y^2 = R^2(\sin^2 \omega t + \cos^2 \omega t) = R^2$ so the path is certainly a subset of that circle. The exact part of the circle traced out depends on which values of t are prescribed (i.e., on the *domain* of the function.) If t extends over an interval of size $2\pi/\omega$ (i.e., ωt extends over an interval of size 2π), the circle will be traced exactly once, but for other domains, only part of the circle may be traced or the circle might be traced several times. The constant ω determines the rate at which the particle moves around the circle. You should try some representative values of t , say with $\omega = 1$ to convince yourself that the circle is traced counterclockwise if $\omega > 0$ and clockwise if $\omega < 0$. (What about $\omega = 0$?)



The vector equation for the circle would be $\mathbf{r} = (R \cos \omega t)\mathbf{i} + (R \sin \omega t)\mathbf{j}$ or, in component notation, $\mathbf{r} = \langle R \cos \omega t, R \sin \omega t \rangle$.

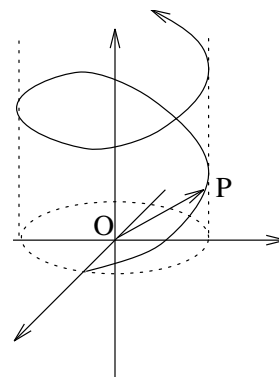
Example 3 Let

$$x = R \cos \omega t$$

$$y = R \sin \omega t$$

$$z = bt.$$

Then the position vector $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ traces out a path in space. The projection of this path in the x, y -plane (obtained by setting $z = 0$) is the circle described in Example 2. At the same time, the particle is rising (assuming $b > 0$) in the z -direction at a constant rate. The resulting path is called a *helix*. (What if $b < 0$ or $b = 0$?)

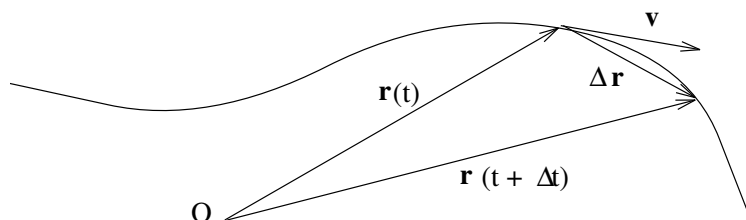


The vector equation for the helix would be $\mathbf{r} = (R \cos \omega t)\mathbf{i} + (R \sin \omega t)\mathbf{j} + bt\mathbf{k}$ or, in component notation, $\mathbf{r} = \langle R \cos \omega t, R \sin \omega t, bt \rangle$.

Velocity and Acceleration Suppose the path of a particle is described by a vector function $\mathbf{r} = \mathbf{r}(t)$. The derivative of such a function may be defined exactly as in the case of a scalar function of a real variable. Let t change by a small amount Δt , and put $\Delta \mathbf{r} = \mathbf{r}(t + \Delta t) - \mathbf{r}(t)$. Then define

$$\mathbf{r}'(t) = \frac{d\mathbf{r}}{dt} = \lim_{\Delta t \rightarrow 0} \frac{\Delta \mathbf{r}}{\Delta t}.$$

Although this looks formally exactly like the scalar case, the geometry is quite different. Study the accompanying diagram.



$\Delta \mathbf{r}$ is the displacement vector from the position of the particle at time t to its position at time $t + \Delta t$. It is represented by the directed *chord* in the diagram. As $\Delta t \rightarrow 0$, the direction of the chord approaches a limiting direction, which we call the *tangent direction* to the curve. The derivative $\mathbf{r}'(t)$ is also called *the instantaneous velocity*, and it is usually denoted \mathbf{v} (or $\mathbf{v}(t)$ if we want to emphasize its functional dependence on t .) It is, of course, a vector, and we usually picture it with its tail at the point with position vector $\mathbf{r}(t)$ and pointing (appropriately) along the tangent line to the curve.

Calculating the derivative (or velocity) is much easier if we use components. If $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$, then we may write $\Delta \mathbf{r} = \Delta x\mathbf{i} + \Delta y\mathbf{j} + \Delta z\mathbf{k}$, and

$$\begin{aligned}\mathbf{r}'(t) &= \lim_{\Delta t \rightarrow 0} \frac{\Delta \mathbf{r}}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \frac{\Delta x}{\Delta t} \mathbf{i} + \lim_{\Delta t \rightarrow 0} \frac{\Delta y}{\Delta t} \mathbf{j} + \lim_{\Delta t \rightarrow 0} \frac{\Delta z}{\Delta t} \mathbf{k} \\ &= \frac{dx}{dt} \mathbf{i} + \frac{dy}{dt} \mathbf{j} + \frac{dz}{dt} \mathbf{k}.\end{aligned}$$

In other words, *the components of the derivative of a vector function are just the derivatives of the component functions.*

Of course, in calculus, one need not stop with the first derivative. The second derivative

$$\frac{d^2 \mathbf{r}}{dt^2} = \mathbf{r}''(t) = \frac{d^2 x}{dt^2} \mathbf{i} + \frac{d^2 y}{dt^2} \mathbf{j} + \frac{d^2 z}{dt^2} \mathbf{k}$$

is called the *acceleration*, and it is often denoted \mathbf{a} . It may also be described as $\mathbf{a} = \frac{d\mathbf{v}}{dt}$.

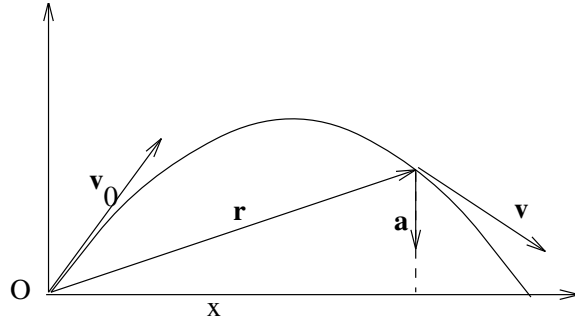
Example 1 Let $x(t) = v_{x0}t$, $y(t) = v_{y0}t - (1/2)gt^2$ as in the previous section. Then

$$\mathbf{v} = \mathbf{r}'(t) = v_{x0}\mathbf{i} + (v_{y0} - gt)\mathbf{j}$$

and, in particular, at $t = 0$, we have $\mathbf{r}'(0) = v_{x0}\mathbf{i} + v_{y0}\mathbf{j}$. In other words, the velocity vector at $t = 0$ is the vector \mathbf{v}_0 with components $\langle v_{x0}, v_{y0} \rangle$, as expected. (Where on the path does the velocity vector point horizontally?) Similarly, the acceleration

$$\mathbf{a} = \frac{d\mathbf{v}}{dt} = 0\mathbf{i} + (0 - g)\mathbf{j} = -g\mathbf{j}.$$

is directed vertically downward and has magnitude g . You are probably familiar with this from your previous work in physics.



Example 2 Let $\mathbf{r} = R \cos \omega t \mathbf{i} + R \sin \omega t \mathbf{j}$, (i.e., $x = R \cos \omega t, y = R \sin \omega t$.) Then

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = (-R\omega \sin \omega t) \mathbf{i} + (R\omega \cos \omega t) \mathbf{j}$$

and a little trigonometry will convince you that \mathbf{v} is perpendicular to \mathbf{r} . For, \mathbf{r} makes angle $\theta = \omega t$ with the positive x -axis while \mathbf{v} makes angle $\theta + \pi/2$. Hence, \mathbf{v} is tangent to the circle as expected. Also,

$$|\mathbf{v}| = \sqrt{R^2 \omega^2 \sin^2 \omega t + R^2 \omega^2 \cos^2 \omega t} = R\omega. \quad (2)$$

Hence, the *speed* $|\mathbf{v}|$ is constant.

The acceleration is given by

$$\mathbf{a} = \frac{d\mathbf{v}}{dt} = (-R\omega^2 \cos \omega t) \mathbf{i} + (-R\omega^2 \sin \omega t) \mathbf{j} = -\omega^2 \mathbf{r}.$$

Hence, the acceleration is directed opposite to the position vector and points from the position of the particle toward the origin. This is usually called *centripetal acceleration*. Also, $|\mathbf{a}| = \omega^2 |\mathbf{r}| = \omega^2 R$. By equation (2), $\omega = \frac{|\mathbf{v}|}{R}$, so

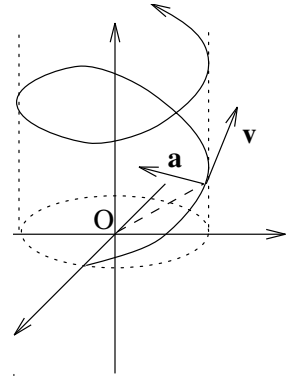
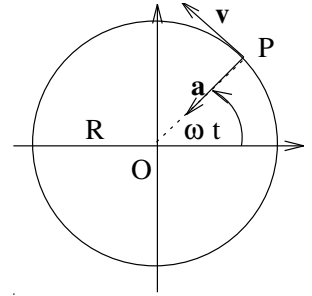
$$|\mathbf{a}| = \frac{|\mathbf{v}|^2}{R^2} R = \frac{|\mathbf{v}|^2}{R}.$$

Note that in this example, acceleration results entirely from changes in the direction of the velocity vector since its magnitude is constant. In general, both the direction and magnitude of the velocity vector will be changing.

Example 3 Let $\mathbf{r} = \langle R \cos \omega t, R \sin \omega t, bt \rangle$ represent a helix as above.

The velocity and acceleration are given by

$$\begin{aligned} \mathbf{v} &= \langle -R\omega \sin \omega t, R\omega \cos \omega t, b \rangle \\ \mathbf{a} &= \langle -R\omega^2 \cos \omega t, R\omega^2 \sin \omega t, 0 \rangle. \end{aligned}$$



Since its third component is zero, \mathbf{a} points parallel to the x, y -plane. Also, if you compare the first two components with what we obtained in the example of uniform circular motion, you will see that \mathbf{a} points from the position of the particle directly at the z -axis.

Example 4. Uniformly Accelerated Motion Suppose the acceleration vector \mathbf{a} is *constant*. Then, just as would be the case for scalar functions, we can *integrate* the equation

$$\frac{d\mathbf{v}}{dt} = \mathbf{a}$$

to obtain

$$\mathbf{v} = \mathbf{a}t + \mathbf{c}$$

where \mathbf{c} is a *vector constant*. In fact, putting $t = 0$ shows that $\mathbf{c} = \mathbf{v}(0)$, the value of $\mathbf{v}(t)$ at 0, and this is usually denoted \mathbf{v}_0 . Thus, we may write $\frac{d\mathbf{r}}{dt} = \mathbf{v} = t\mathbf{a} + \mathbf{v}_0$. (It is customary to put the scalar t in front of the vector, but it is not absolutely necessary.) If we integrate this, we obtain

$$\mathbf{r} = \frac{1}{2}t^2\mathbf{a} + t\mathbf{v}_0 + \mathbf{C},$$

but putting $t = 0$, as above, yields $\mathbf{C} = \mathbf{r}(0) = \mathbf{r}_0$, the value of the position vector at $t = 0$. Thus, we obtain

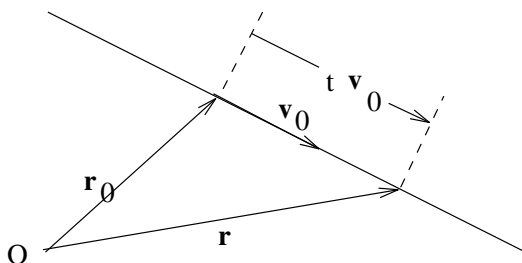
$$\mathbf{r} = \frac{1}{2}t^2\mathbf{a} + t\mathbf{v}_0 + \mathbf{r}_0.$$

The path of such a particle is a *parabola*. (Can you see why?)

The special case $\mathbf{a} = \mathbf{0}$ is of interest. This is called *uniform linear motion* and is described by

$$\mathbf{r} = t\mathbf{v}_0 + \mathbf{r}_0.$$

The path is a straight line. (See the diagram.) Moreover, $\mathbf{v} = d\mathbf{r}/dt = \mathbf{v}_0$, so the velocity vector is constant. Of course, it points along the line of motion. The particle moves along this line at constant speed.



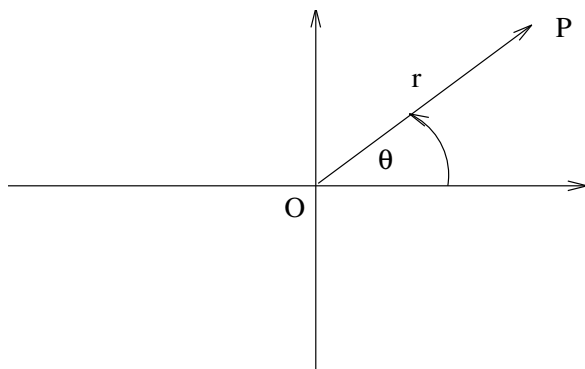
All our discussions have assumed implicitly that $\mathbf{v} \neq \mathbf{0}$. If \mathbf{v} vanishes for one particular value of t , there may not be a well defined tangent vector at the corresponding

point on the path. A simple example of this would be a particle which rises vertically in a straight line with decreasing velocity until it reaches its maximum height where it reverses direction and falls back along the same line. At the point of maximum height, the velocity would be zero, and it does not really make sense to talk about a tangent *vector* there since we can't assign it a well defined direction. If $\mathbf{v}(t)$ vanishes for all t in an interval, then the particle stays at one point without moving. It would make even less sense to talk about a tangent vector in that case.

Remark. In our discussion of derivatives (and integrals) of vector functions, we shall ignore for the present the issue of how *limits* are handled for such functions. This would be a matter of some importance for a completely rigorous treatment because those concepts are defined by limits. These issues are best postponed to a course in *real analysis* where there is time for such matters. It suffices for now to say that such limits behave exactly as you would expect. Also, one way to avoid worrying about the matter is to reduce all limits for vector functions of a single variable t to statements about limits for the scalar component functions.

Polar Coordinates For motion in the plane where there is some sort of circular symmetry, it is often more convenient to use polar coordinates. This would certainly be the case for circular motion, but it is also useful, for example, in celestial mechanics where we think of a planet as a particle moving in the gravitational field of a star. In that case, the gravitational force may be assumed to point directly toward the origin, and Newton, confirming what Kepler had demonstrated, showed that the motion would lie in a plane and follow a conic section.

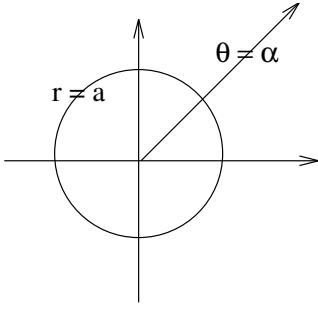
Recall how polar coordinates are defined in the plane. We choose an origin O , and a (right handed) cartesian coordinate system with that origin. The polar coordinates (r, θ) of a point P are the distance $r = |\mathbf{r}| = |\overrightarrow{OP}|$ to the origin and the angle θ which the position vector \mathbf{r} makes with the positive x -axis.



By definition $r \geq 0$ since it is a distance. Generally, θ is allowed to have any value, but so that the point P will uniquely determine its polar angle θ , it is common

Turns around





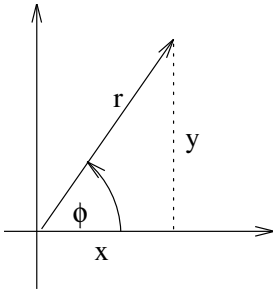
to restrict θ to some interval of length 2π . Common choices are $0 \leq \theta < 2\pi$ or $-\pi \leq \theta < \pi$, but many other choices are possible. Note that in any case, there is no way to define θ unambiguously at the origin where $r = 0$.

The set of points with a fixed value $r = a > 0$ constitute a circle with radius a . The set of points with a fixed value $\theta = \alpha$ constitute a *ray*, i.e., a half line, emanating from the origin, making angle α with the positive x -axis.

It is common in elementary mathematics books to interpret a point with polar coordinates (r, θ) where $r < 0$ as lying on the ray opposite to the ray for the given value of θ . This is convenient in some formulas, but it adds another degree of ambiguity since we can get to the opposite ray just as well by adding π to θ . Such practices are best avoided. In this course, you may generally assume $r \geq 0$, but each time you encounter polar coordinates in other contexts, you will have to check what the author intends.

r and θ are the most commonly used symbols for polar coordinates in mathematics and physics books, but there are others you may encounter. For example, you may sometimes see ρ in place of r or ϕ in place of θ . Later in this course, we shall introduce coordinates in space which are analogous to polar coordinates in the plane. For these, there is even more variation of symbols in common use. It is specially important when you see any of these coordinate systems used that you concentrate on the geometric and physical meaning of the quantities involved rather than on the particular letters used to represent them.

You may remember the following formulas relating rectangular and polar coordinates. In any case, they are clear by elementary trigonometry from the diagram.



$$x = r \cos \theta$$

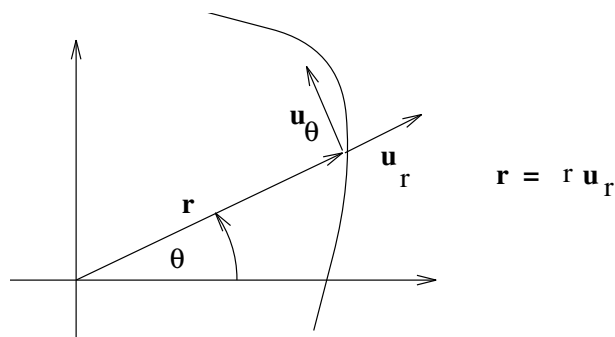
$$y = r \sin \theta$$

and

$$r = \sqrt{x^2 + y^2}$$

$$\tan \theta = \frac{y}{x}, \quad \text{if } x \neq 0.$$

The description of motion in polar coordinates is both simpler and more complicated than in rectangular coordinates. Instead of expressing vectors in terms of \mathbf{i} and \mathbf{j} , it is useful instead to use unit vectors associated with the polar directions. These depend on the value of θ at the point P under consideration and should be viewed as placed with their tails at that point. \mathbf{u}_r is chosen to point directly away from the origin, so it is parallel to the position vector $\mathbf{r} = \overrightarrow{OP}$. \mathbf{u}_θ is chosen perpendicular to \mathbf{u}_r and pointing in the counter-clockwise direction (positive θ), so it is tangent to a circle passing through P and centered at O .



(These are also commonly denoted $\hat{\mathbf{r}}$ and $\hat{\boldsymbol{\theta}}$). Because of the definition of \mathbf{u}_r , we have

$$\mathbf{r} = r\mathbf{u}_r$$

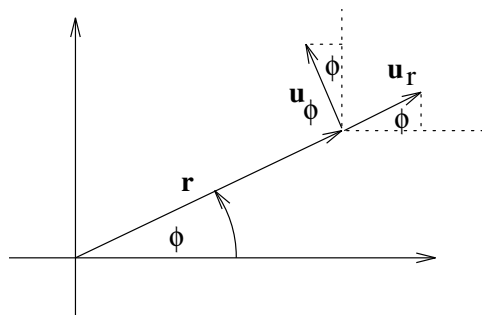
(This equation is a little confusing since the position vector \mathbf{r} is commonly thought of with its tail at O while \mathbf{u}_r was supposed to be thought of with its tail at P . However, if you remember that the tail of a vector can be placed wherever convenient without changing the vector, you won't have a problem.)

It follows that the velocity vector is given by

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \frac{d(r\mathbf{u}_r)}{dt} = \frac{dr}{dt}\mathbf{u}_r + r\frac{d\mathbf{u}_r}{dt},$$

where we have calculated the product using the *product rule* for derivatives. Since, \mathbf{u}_r changes its direction as we move along the path, it is necessary to keep the second term. (How does the case of rectangular coordinates differ?) To calculate further, we first express \mathbf{u}_r and \mathbf{u}_θ in rectangular coordinates. (See the diagram.)

$$\begin{aligned}\mathbf{u}_r &= \cos\theta\mathbf{i} + \sin\theta\mathbf{j}, \\ \mathbf{u}_\theta &= -\sin\theta\mathbf{i} + \cos\theta\mathbf{j}.\end{aligned}$$



Hence,

$$\frac{d\mathbf{u}_r}{dt} = -\sin\theta \frac{d\theta}{dt} \mathbf{i} + \cos\theta \frac{d\theta}{dt} \mathbf{j} = \frac{d\theta}{dt} \mathbf{u}_\theta. \quad (3)$$

A similar calculation shows that

$$\frac{d\mathbf{u}_\theta}{dt} = -\frac{d\theta}{dt} \mathbf{u}_r. \quad (4)$$

Putting (3) in the expression for \mathbf{v} yields

$$\mathbf{v} = \frac{dr}{dt} \mathbf{u}_r + r \frac{d\theta}{dt} \mathbf{u}_\theta.$$

This same process can be repeated to obtain the acceleration $\mathbf{a} = \frac{d\mathbf{v}}{dt}$. Using both (3) and (4) yields after considerable calculation

$$\mathbf{a} = \left(\frac{d^2r}{dt^2} - r \left(\frac{d\theta}{dt} \right)^2 \right) \mathbf{u}_r + \left(r \frac{d^2\theta}{dt^2} + 2 \frac{dr}{dt} \frac{d\theta}{dt} \right) \mathbf{u}_\theta. \quad (5)$$

I leave it as a challenge for your algebraic prowess to try to verify this formula.

Example. Uniform Circular Motion Suppose the particle moves in a circle of radius R centered at the origin, so that $r = R$, and $dr/dt = 0$. Suppose in addition that $d\theta/dt = \omega$ is constant. Then $d^2r/dt^2 = d^2\theta/dt^2 = 0$, and putting all this in the above expressions yield

$$\begin{aligned} \mathbf{v} &= R\omega \mathbf{u}_\theta \\ \mathbf{a} &= -R\omega^2 \mathbf{u}_r = -\frac{|\mathbf{v}|^2}{R} \mathbf{u}_r \end{aligned}$$

as we discovered previously.

Exercises for 1.2.

1. Suppose a projectile follows the path described in Example 1 of in this section. Show that the *range* (i.e., the x -coordinate of the point of impact) is given by $2 \frac{v_{x0}v_{y0}}{g}$. Hint: Find the time at which impact occurs.
2. Find the velocity and acceleration vectors at the indicated times if the position vector is given by:
 - (a) $\mathbf{r}(t) = 4\mathbf{i} + 5\mathbf{j} - 3\mathbf{k}$, $t = 2$;
 - (b) $\mathbf{r}(t) = 3\mathbf{i} \cos t - 4\mathbf{j} \sin t$, $t = 0$;
 - (c) $\mathbf{r}(t) = -2\mathbf{i}e^{3t} + \mathbf{j}t$, $t = 1$;
 - (d) $\mathbf{r}(t) = (2t - 5)\mathbf{i} + (3t + 1)\mathbf{j} + 2\mathbf{k}$, $t = -2$;
 - (e) $\mathbf{r}(t) = \langle \cos t, 0 \rangle$, t arbitrary.

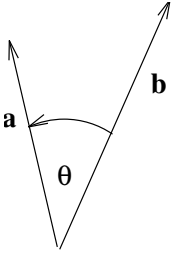
3. Calculate the following. (Integrate each component separately.)
 - (a) $\int_0^\pi ((1 + \cos t)\mathbf{i} - 2\mathbf{j} \sin t) dt$
 - (b) $\int_0^1 ((3 + 2x)\mathbf{i} + x^2\mathbf{j} - 5\mathbf{k}) dx$
 - (c) $\int_0^{2\pi} (3\mathbf{i} \sin \theta - 4\mathbf{j} \cos \theta) d\theta$.
4. Given the parametric equations $x = -2 \sin t$, $y = 7$, and $z = 4 \cos t$, find the particle's position, velocity, speed, and acceleration at $t = 2\pi$.
5. Using the method in Example 4 of this section, find the general velocity and position vectors if
 - (a) $\mathbf{a} = 2\mathbf{i}$, $\mathbf{r}_0 = \mathbf{i} + 2\mathbf{j}$, $\mathbf{v}_0 = 0$.
 - (b) $\mathbf{a} = \mathbf{0}$, $\mathbf{r}_0 = 3\mathbf{i} + 4\mathbf{j} + 5\mathbf{k}$, $\mathbf{v}_0 = 5\mathbf{i}$.
 - (c) $\mathbf{a} = \mathbf{i} \sin t - \mathbf{j} \cos t$, $\mathbf{r}_0 = \mathbf{i}$, $\mathbf{v}_0 = \mathbf{j}$
6. Show that the path described by the equation $\mathbf{r} = \frac{1}{2}t^2\mathbf{a} + t\mathbf{v}_0 + \mathbf{r}_0$ is a parabola. Hint: This is not too hard to see if you choose your coordinate system properly. Suppose first that the origin is the position of the particle at $t = 0$. Then $\mathbf{r}_0 = 0$. Suppose, moreover that the y -axis is chosen in the direction of \mathbf{a} . Then, $\mathbf{a} = a\mathbf{j}$. Finally, choose the z -axis so that it is perpendicular both to \mathbf{a} and to the initial velocity \mathbf{v}_0 . Then $\mathbf{v}_0 = v_{x0}\mathbf{i} + v_{y0}\mathbf{j}$. Write out the equations for x and y with these assumptions.
7. Consider the path in the plane described by the equation $\mathbf{r} = \langle a \cos \omega t, b \sin \omega t \rangle$.
 - (a) Show that the particle traces out the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$. How much time must elapse for the particle to trace the ellipse exactly once?
 - (b) Show that the acceleration is directed towards the origin.
8. Find $\mathbf{v}(t)$ and $\mathbf{r}(t)$ in terms of \mathbf{u}_r and \mathbf{u}_θ if
 - (a) $r = 2(\sin t)$ and $\theta = 2t$
 - (b) $r = a$ and $\theta = 2t$
 - (c) $r = t$ and $\theta = t$
9. Show that $\frac{d\mathbf{u}_\theta}{dt} = -\frac{d\theta}{dt}\mathbf{u}_r$. (This is equation (4) in this section.)
10. (Optional) Verify formula (5) in this section.
11. A bead is on a spoke of a wheel 20 inches in diameter. The wheel rotates at a rate of 2 revolutions per second. At the same time, the bead moves uniformly out from the center at 2 inches per second. Assume the bead starts at the center at $t = 0$. (a) Find expressions for r and θ as functions of t . (b) Find the velocity and acceleration vectors at each point between the center and the rim.

12. A billiard ball bounces off the side of a billiard table at an angle. What can you say about the velocity vector at the point of impact? Is there a well defined tangent direction? Note that since this question involves making some assumptions about the physics of the collision, it does not have a single mathematical answer.

1.3 The Dot Product

Let \mathbf{a} and \mathbf{b} be vectors. We assume they are placed so their tails coincide. Let θ denote the *smaller* of the two angles between them, so $0 \leq \theta \leq \pi$. Their *dot product* is defined to be

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos \theta.$$



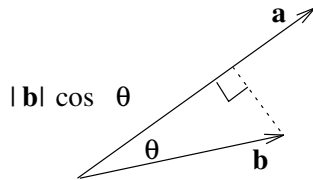
This is also sometimes called the *scalar product* because the result is a scalar. Note that $\mathbf{a} \cdot \mathbf{b} = 0$ when either \mathbf{a} or \mathbf{b} is zero or, more interestingly, if their directions are perpendicular. If the two vectors have parallel directions, $\mathbf{a} \cdot \mathbf{b}$ is the product of their magnitudes if they point the same way or the negative of that product if they point in opposite directions. (Make sure you understand why.)

The dot product is a useful concept when one needs to find the component of one vector in the direction of another. For example, in a typical inclined plane problem in elementary mechanics, one needs to resolve the vertical gravitational force into components, one parallel to the inclined plane, and one perpendicular to it. To see how the dot product enters into that, note that

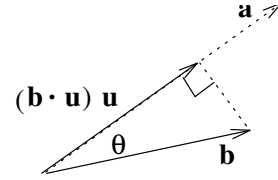
the quantity $|\mathbf{b}| \cos \theta$ is just the perpendicular projection of \mathbf{b} onto any line parallel to \mathbf{a} . (See the diagram.) Hence, we have

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| (\text{projection of } \mathbf{b} \text{ on } \mathbf{a}). \quad (6)$$

(Note that this description is symmetric; either vector could be projected onto a line parallel to the other.) In particular, if \mathbf{a} is a unit vector ($|\mathbf{a}| = 1$), $\mathbf{a} \cdot \mathbf{b}$ is just the value of the projection. (What does it mean if $\mathbf{a} \cdot \mathbf{b} < 0$?)



Scalar projection



Vector projection

The above projection is a scalar quantity, but one is often interested in the *vector* obtained by multiplying the scalar projection of \mathbf{b} on \mathbf{a} with a unit vector in the direction of \mathbf{a} . (See the accompanying diagram.) $\mathbf{u} = \frac{1}{|\mathbf{a}|}\mathbf{a}$ is such a unit vector, so this *vector projection* of \mathbf{b} on \mathbf{a} is given by

$$(\mathbf{u} \cdot \mathbf{b})\mathbf{u} = \left(\frac{1}{|\mathbf{a}|}\mathbf{a} \cdot \mathbf{b}\right)\frac{1}{|\mathbf{a}|}\mathbf{a} = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}|^2}\mathbf{a}. \quad (7)$$

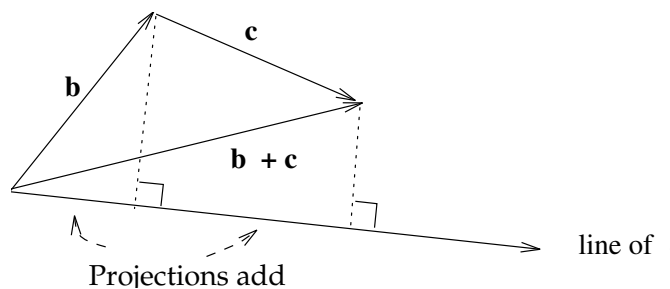
The dot product satisfies certain simple algebraic rules.

$$\begin{aligned} \mathbf{a} \cdot \mathbf{b} &= \mathbf{b} \cdot \mathbf{a} && \text{commutative law,} \\ \mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) &= \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c} && \text{distributive law,} \\ (s\mathbf{a}) \cdot \mathbf{b} &= \mathbf{a} \cdot (s\mathbf{b}) = s(\mathbf{a} \cdot \mathbf{b}), \end{aligned}$$

and

$$\mathbf{a} \cdot \mathbf{a} = |\mathbf{a}|^2.$$

These can be proved without too much difficulty from the geometric definition. See, for example, the accompanying diagram which illustrates the proof of the distributive law.



Theorem 1.1 Let the components of \mathbf{a} be $\langle a_1, a_2, a_3 \rangle$ and let those of \mathbf{b} be $\langle b_1, b_2, b_3 \rangle$. Then

$$\mathbf{a} \cdot \mathbf{b} = a_1b_1 + a_2b_2 + a_3b_3.$$

For plane vectors, the same principle applies without the third components.

Examples

For $\langle 1, 2, -1 \rangle$ and $\langle 2, -1, 3 \rangle$, the dot product is $2 - 2 - 3 = -3$. In particular, that means the angle between the two vectors is obtuse.

For $\langle 1, 1, 2 \rangle$ and $\langle -1, -1, 1 \rangle$, the dot product is $-1 - 1 + 2 = 0$. That means the two vectors are perpendicular.

Proof. We have

$$\begin{aligned}\mathbf{a} &= a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k} \\ \mathbf{b} &= b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k},\end{aligned}$$

so the aforementioned rules of algebra yield

$$\begin{aligned}\mathbf{a} \cdot \mathbf{b} &= a_1b_1\mathbf{i} \cdot \mathbf{i} + a_1b_2\mathbf{i} \cdot \mathbf{j} + a_1b_3\mathbf{i} \cdot \mathbf{k} \\ &\quad + a_2b_1\mathbf{j} \cdot \mathbf{i} + a_2b_2\mathbf{j} \cdot \mathbf{j} + a_2b_3\mathbf{j} \cdot \mathbf{k} \\ &\quad + a_3b_1\mathbf{k} \cdot \mathbf{i} + a_3b_2\mathbf{k} \cdot \mathbf{j} + a_3b_3\mathbf{k} \cdot \mathbf{k}.\end{aligned}$$

However, \mathbf{i} , \mathbf{j} , and \mathbf{k} are mutually perpendicular unit vectors, so the off-diagonal dot products (e.g., $\mathbf{i} \cdot \mathbf{j}$, $\mathbf{i} \cdot \mathbf{k}$, etc.) are all zero while the diagonal dot products (e.g., $\mathbf{i} \cdot \mathbf{i}$) are all one. Hence, only the three diagonal terms survive and we obtain $a_1b_1 + a_2b_2 + a_3b_3$ as claimed. \square \square

The theorem gives us a way to calculate the angle between two vectors *in case we know their components*. Indeed, the formula may be rewritten

$$\cos \theta = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}.$$

The right hand side may be computed using $\mathbf{a} \cdot \mathbf{b} = a_1b_1 + a_2b_2 + a_3b_3$, $|\mathbf{a}| = \sqrt{a_1^2 + a_2^2 + a_3^2}$, and similarly for $|\mathbf{b}|$, and from this we can determine θ .

Example Suppose $\mathbf{a} = \mathbf{i} + 2\mathbf{j} - \mathbf{k}$, $\mathbf{b} = 2\mathbf{i} - \mathbf{j} + 3\mathbf{k}$. Then

$$\begin{aligned}\mathbf{a} \cdot \mathbf{b} &= (1)(2) + (2)(-1) + (-1)(3) = -3 \\ |\mathbf{a}| &= \sqrt{1 + 4 + 1} = \sqrt{6} \\ |\mathbf{b}| &= \sqrt{4 + 1 + 9} = \sqrt{14}\end{aligned}$$

so

$$\begin{aligned}\cos \theta &= \frac{-3}{\sqrt{84}} \\ \theta &= 1.90427 \text{ radians}\end{aligned}$$

The use of components simplifies other calculations also. For example, suppose we want to know the *projection* of the vector \mathbf{b} on a line parallel to \mathbf{a} . We know from (7) that this is

$$|\mathbf{b}| \cos \theta = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}|} = \frac{-3}{\sqrt{6}}.$$

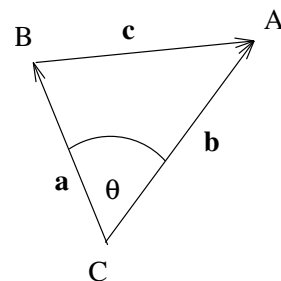
Similarly, from (7), we see that the vector projection of \mathbf{b} on \mathbf{a} is given by

$$\frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}|^2} \mathbf{a} = \frac{-3}{6} (\mathbf{i} + 2\mathbf{j} - \mathbf{k}) = -\frac{1}{2}\mathbf{i} - \mathbf{j} + \frac{1}{2}\mathbf{k}.$$

Note that in all these calculations we must rely on the use of components to make the formulas useful.

The Law of Cosines Let A, B , and C be the vertices of a triangle, and define vectors

$$\begin{aligned}\mathbf{a} &= \overrightarrow{CB} \\ \mathbf{b} &= \overrightarrow{CA} \\ \mathbf{c} &= \overrightarrow{BA} = \mathbf{b} - \mathbf{a}.\end{aligned}$$



Then we have

$$\begin{aligned}|\mathbf{c}|^2 &= \mathbf{c} \cdot \mathbf{c} = (\mathbf{b} - \mathbf{a}) \cdot (\mathbf{b} - \mathbf{a}) \\ &= \mathbf{b} \cdot \mathbf{b} - \mathbf{b} \cdot \mathbf{a} - \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{a} \\ &= |\mathbf{b}|^2 - 2\mathbf{a} \cdot \mathbf{b} + |\mathbf{a}|^2\end{aligned}$$

which may be rewritten

$$|\mathbf{c}|^2 = |\mathbf{a}|^2 + |\mathbf{b}|^2 - 2|\mathbf{a}||\mathbf{b}|\cos\theta.$$

You should recognize that as the *Law of Cosines* from Trigonometry.

Generalizations We have already mentioned higher dimensional vectors being characterized as n -tuples for values of $n \geq 4$. The dot product of n -tuples can be defined by analogy with the formula derived above. Thus for 4-vectors, $\mathbf{a} = \langle a_1, a_2, a_3, a_4 \rangle$, $\mathbf{b} = \langle b_1, b_2, b_3, b_4 \rangle$, we would define

$$\mathbf{a} \cdot \mathbf{b} = a_1b_1 + a_2b_2 + a_3b_3 + a_4b_4.$$

In the theory of special relativity, it turns out to be useful to consider instead a “dot product” of 4-vectors of the form

$$a_1b_1 + a_2b_2 + a_3b_3 - c^2a_4b_4$$

where c is the speed of light. (Some authors prefer the negative of this quantity.) This is a bit bizarre, because the “dot product” of a vector with itself can be zero without the vector being zero.

We shall consider some of these strange and wonderful ideas later in this course.

Differentiation of Dot Products The usual differentiation rules, e.g., the product rule, the chain rule, etc., apply to vector functions as well as to scalar functions. Indeed, we have already used some of these rules without making a point of it. The proofs are virtually the same as those in the scalar case, so we need not go through them again in this course. However, since there are so many different vector operations, it is worth exploring the consequences of these rules in interesting cases.

They are sometimes not what you would expect. For example, the *product rule* for the dot product of two vector functions $\mathbf{f}(t)$ and $\mathbf{g}(t)$ takes the form

$$\frac{d}{dt}\mathbf{f}(t) \cdot \mathbf{g}(t) = \frac{d\mathbf{f}(t)}{dt} \cdot \mathbf{g}(t) + \mathbf{f}(t) \cdot \frac{d\mathbf{g}(t)}{dt}.$$

Let us apply this to the case $\mathbf{f}(t) = \mathbf{g}(t) = \mathbf{r}(t)$, the position vector of a particle moving in the plane or in space. Assume the particle moves so that its distance to the origin is a constant R . Symbolically, this can be written $\mathbf{r} \cdot \mathbf{r} = |\mathbf{r}|^2 = R^2$. Then the product rule gives

$$0 = \frac{d(\mathbf{r} \cdot \mathbf{r})}{dt} = \frac{d\mathbf{r}}{dt} \cdot \mathbf{r} + \mathbf{r} \cdot \frac{d\mathbf{r}}{dt} = 2\mathbf{r} \cdot \mathbf{v}.$$

It follows from this that either $\mathbf{v} = 0$ or $\mathbf{v} \perp \mathbf{r}$. In the plane case, this is yet another confirmation that for motion in a circle, the velocity vector is perpendicular to the radius vector.

Similar reasoning shows that if $|\mathbf{v}|$ is constant, then the acceleration vector \mathbf{a} is perpendicular to \mathbf{v} . (You should write it out to convince yourself you understand the argument.)

Exercises for 1.3.

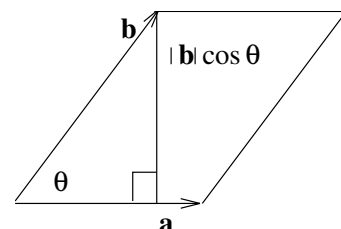
- In each case determine if the given pair of vectors is perpendicular.
 - $\mathbf{a} = 3\mathbf{i} + 4\mathbf{j}$, $\mathbf{b} = -4\mathbf{i} + 3\mathbf{j}$
 - $\mathbf{a} = \langle 4, -1, 2 \rangle$, $\mathbf{b} = \langle 3, 0, -6 \rangle$
 - $\mathbf{a} = 3\mathbf{i} - 2\mathbf{j}$, $\mathbf{b} = -2\mathbf{i} - 4\mathbf{k}$
- Assume $\mathbf{u} = \langle a, b \rangle$ is a non-zero plane vector. Show that $\mathbf{v} = \langle -b, a \rangle$ is perpendicular to \mathbf{u} . By examining all possible signs for a and b , convince yourself that the 90 degree angle between \mathbf{u} and \mathbf{v} is in the counter-clockwise direction.
- The methane molecule, CH_4 , has four Hydrogen atoms at the vertices of a regular tetrahedron and a Carbon atom at its center. Choose as the vertices of this tetrahedron the points $(0,0,0)$, $(1,1,0)$, $(1,0,1)$, and $(0,1,1)$. (a) Find the angle between two edges of the tetrahedron. (b) Find the bond angle between two Carbon–Hydrogen bonds.
- An inclined plane makes an angle of 30 degrees with the horizontal. Use vectors and the dot product to find the scalar and vector projections of the gravitational acceleration vector $-g\mathbf{j}$ along a unit vector pointing down the inclined plane.
- Show that if the velocity vector \mathbf{v} is perpendicular to the acceleration vector \mathbf{a} at every point of a path, then the speed $|\mathbf{v}|$ is constant.

6. Derive the formula

$$|\mathbf{a} + \mathbf{b}|^2 = |\mathbf{a}|^2 + |\mathbf{b}|^2 + 2|\mathbf{a}||\mathbf{b}|\cos\theta.$$

How is it related to the Law of Cosines? A picture might help.

7. Use the dot product to determine if the points $P(3, 1, 2)$, $Q(-1, 0, 2)$, and $R(11, 3, 2)$ are collinear.
8. If a relativity physicist claimed that the universe is orderly since the physical vectors $\langle 2.54, 9.8, 6.626 \times 10^{-34}, -c^2 \rangle$ and $\langle -9.8, -2.54, 1.509 \times 10^{33}, 0 \rangle$ are perpendicular in a four-space, would his mathematics be correct? Would it make any difference which ‘dot product’ he was using?



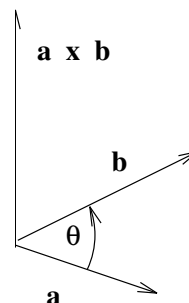
1.4 The Vector Product

Let \mathbf{a} and \mathbf{b} be two vectors in space placed so their tails coincide, and let θ be the smaller of the two angles between them (i.e., $0 \leq \theta \leq \pi$). Previously, we defined the dot or scalar product $\mathbf{a} \cdot \mathbf{b}$. There is a second product called the *vector product* or *cross product*. It is denoted $\mathbf{a} \times \mathbf{b}$, and as its name suggests, it is a *vector*. It is used extensively in mechanics for such notions as torque and angular momentum, and we shall use it shortly in studying solid analytic geometry.

$\mathbf{a} \times \mathbf{b}$ is defined as follows. Its magnitude is

$$|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}||\mathbf{b}|\sin\theta. \quad (9)$$

The quantity on the right has a very simple interpretation. $|\mathbf{b}|\sin\theta$ is the height of the parallelogram spanned by \mathbf{a} and \mathbf{b} , so $|\mathbf{a}||\mathbf{b}|\sin\theta$ is the *area of the parallelogram*. Note it is zero if the vectors point in the same ($\theta = 0$) or opposite ($\theta = \pi$) directions, but otherwise it is non-zero.



The direction of $\mathbf{a} \times \mathbf{b}$ (when it isn't zero) is a bit harder to describe. First, $\mathbf{a} \times \mathbf{b}$ is perpendicular to both \mathbf{a} and \mathbf{b} , and given that we know its magnitude that leaves precisely two possibilities. We specify that it has the *right hand* orientation. That is, if the fingers of your right hand point from \mathbf{a} to \mathbf{b} through the angle θ , $\mathbf{a} \times \mathbf{b}$ should point in the direction of your thumb.

The vector product has some surprising algebraic properties. First, as already noted,

$$\mathbf{a} \times \mathbf{b} = \mathbf{0}$$

when \mathbf{a} and \mathbf{b} point in the same or opposite directions. In particular, $\mathbf{a} \times \mathbf{a} = \mathbf{0}$ for any vector \mathbf{a} . Secondly, the commutative law fails, and we have instead

$$\mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a}.$$

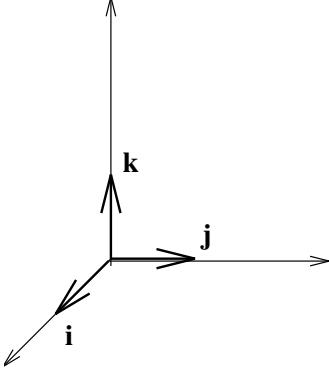
(Point from \mathbf{b} to \mathbf{a} , and your thumb reverses direction.) The vector product does satisfy other rules of algebra such as

$$\begin{aligned}\mathbf{a} \times (\mathbf{b} + \mathbf{c}) &= \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c} \\ (s\mathbf{a}) \times \mathbf{b} &= \mathbf{a} \times (s\mathbf{b}) = s(\mathbf{a} \times \mathbf{b}),\end{aligned}$$

but they are a bit tricky to verify from the geometric definition. (See below for another approach.)

The vector products of the basis vectors are easy to calculate from the definition. We have

$$\begin{aligned}\mathbf{i} \times \mathbf{j} &= -\mathbf{j} \times \mathbf{i} = \mathbf{k}, \\ \mathbf{j} \times \mathbf{k} &= -\mathbf{k} \times \mathbf{j} = \mathbf{i} \\ \mathbf{k} \times \mathbf{i} &= -\mathbf{i} \times \mathbf{k} = \mathbf{j}.\end{aligned}$$



To calculate vector products in general, we expand in terms of components.

$$\begin{aligned}\mathbf{a} \times \mathbf{b} &= (a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}) \times (b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k}) \\ &= a_1b_1\mathbf{i} \times \mathbf{i} + a_1b_2\mathbf{i} \times \mathbf{j} + a_1b_3\mathbf{i} \times \mathbf{k} \\ &\quad + a_2b_1\mathbf{j} \times \mathbf{i} + a_2b_2\mathbf{j} \times \mathbf{j} + a_2b_3\mathbf{j} \times \mathbf{k} \\ &\quad + a_3b_1\mathbf{k} \times \mathbf{i} + a_3b_2\mathbf{k} \times \mathbf{j} + a_3b_3\mathbf{k} \times \mathbf{k} \\ &= \mathbf{0} + a_1b_2\mathbf{k} - a_1b_3\mathbf{j} \\ &\quad - a_2b_1\mathbf{k} + \mathbf{0} + a_2b_3\mathbf{i} \\ &\quad + a_3b_1\mathbf{j} - a_3b_2\mathbf{i} + \mathbf{0}\end{aligned}$$

so

$$\mathbf{a} \times \mathbf{b} = (a_2b_3 - a_3b_2)\mathbf{i} - (a_1b_3 - a_3b_1)\mathbf{j} + (a_1b_2 - a_2b_1)\mathbf{k}.$$

(Note that this makes extensive use of the rules of algebra for vector products, so one should really prove those rules first.)

There is a simple way to remember the formula. Recall that a 2×2 determinant is defined by

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc.$$

Now consider the *matrix* or array

$$\begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{bmatrix}$$

formed from the components of \mathbf{a} and \mathbf{b} , and calculate the 2×2 determinants obtained by omitting successively each of the columns. For the first and third, use a (+) sign and for the second use a (−) sign.

Example 5 Let $\mathbf{a} = \langle 1, 2, -1 \rangle$ and $\mathbf{b} = \langle 1, -2, 3 \rangle$. Then

$$\mathbf{a} \times \mathbf{b} = \langle 6 - 2, -(3 + 1), -2 - 2 \rangle = \langle 4, -4, -4 \rangle.$$

We can check that this vector is indeed perpendicular to both \mathbf{a} and \mathbf{b} by calculating dot products.

$$\mathbf{a} \cdot (\mathbf{a} \times \mathbf{b}) = 4 - 8 + 4 = 0$$

$$\mathbf{b} \cdot (\mathbf{a} \times \mathbf{b}) = 4 + 8 - 12 = 0.$$

(Many texts suggest defining $\mathbf{a} \times \mathbf{b}$ by a 3×3 determinant

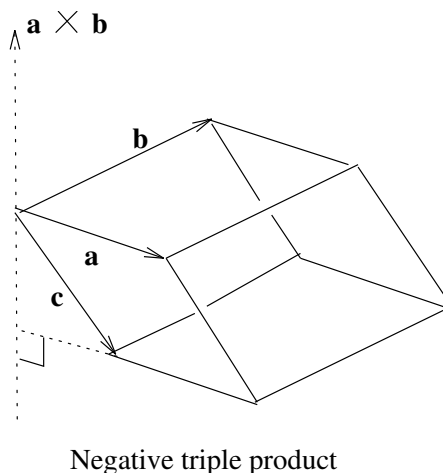
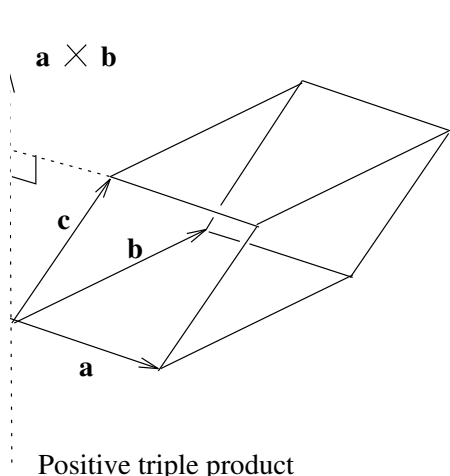
$$\mathbf{a} \times \mathbf{b} = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{bmatrix}. \quad (10)$$

If you are not familiar with 3×3 determinants, see the Exercises.)

The triple product Let \mathbf{a} , \mathbf{b} , and \mathbf{c} be three vectors in space and suppose they are placed so their tails coincide. Then the product

$$(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}$$

has a simple geometric interpretation. The vectors are three of the twelve sides of a *parallelepiped*. $|\mathbf{a} \times \mathbf{b}|$ is the area of the base of that parallelepiped. Moreover the projection of \mathbf{c} on $\mathbf{a} \times \mathbf{b}$ is either the *altitude* of the parallelepiped or the negative of that altitude depending on the relative orientation of the vectors. Hence, the dot product is, except for sign, the *volume* of the parallelepiped. It is positive if \mathbf{c} is on the same side of the plane determined by \mathbf{a} and \mathbf{b} as $\mathbf{a} \times \mathbf{b}$, and it is negative if they are on opposite sides.



Example 6 Let $\mathbf{a} = \langle 0, 1, 1 \rangle$, $\mathbf{b} = \langle 1, 1, 0 \rangle$, and $\mathbf{c} = \langle 1, 0, 1 \rangle$. Then $\mathbf{a} \times \mathbf{b} = \langle -1, 1, -1 \rangle$, so $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = -1 + 0 + (-1) = -2$. Hence, the volume is 2. You should study the diagram to make sure you understand why the answer is negative.

An immediate consequence of the geometric interpretation is the formula

$$(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}). \quad (11)$$

For, except for sign, both sides may be interpreted as the volume of the same parallelepiped, and careful inspection of all possible relative orientations shows that the signs will be the same.

Using 3×3 determinants—see the Exercises—you can check the formula

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \det \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix}. \quad (12)$$

It is worth noting that we can use (11) to verify the formula

$$\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}.$$

For, it would suffice to show that both sides have the same components. Consider the x -component. In general, the x -component of a vector \mathbf{v} is its projection on the x -axis, that is, it is $\mathbf{i} \cdot \mathbf{v}$. However,

$$\begin{aligned} \mathbf{i} \cdot (\mathbf{a} \times (\mathbf{b} + \mathbf{c})) &= (\mathbf{i} \times \mathbf{a}) \cdot (\mathbf{b} + \mathbf{c}) \\ &= (\mathbf{i} \times \mathbf{a}) \cdot \mathbf{b} + (\mathbf{i} \times \mathbf{a}) \cdot \mathbf{c} \\ &= \mathbf{i} \cdot (\mathbf{a} \times \mathbf{b}) + \mathbf{i} \cdot (\mathbf{a} \times \mathbf{c}) \\ &= \mathbf{i} \cdot (\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}). \end{aligned}$$

(Here we used the fact that the distributive law does hold for the dot product.) It follows that the two sides have the same x -component. Similar arguments work for the y and z -components, so the two sides are the same.

Product Rule Like the dot product, the cross product also satisfies the product rule. (Also, the proof is virtually identical to the proof for scalar functions.) Thus, if $\mathbf{f}(t)$ and $\mathbf{g}(t)$ are vector valued functions, then

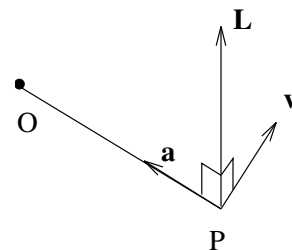
$$\frac{d}{dt}(\mathbf{f} \times \mathbf{g}) = \frac{d\mathbf{f}}{dt} \times \mathbf{g} + \mathbf{f} \times \frac{d\mathbf{g}}{dt}. \quad (13)$$

This formula has many useful consequences. We shall illustrate one such by showing that if a particle moves in space so that the acceleration \mathbf{a} is parallel to the position vector \mathbf{r} , then the motion must be constrained to a plane. To see this, consider the quantity $\mathbf{L} = \mathbf{r} \times \mathbf{v}$. We have

$$\begin{aligned} \frac{d\mathbf{L}}{dt} &= \frac{d\mathbf{r}}{dt} \times \mathbf{v} + \mathbf{r} \times \frac{d\mathbf{v}}{dt} \\ &= \mathbf{v} \times \mathbf{v} + \mathbf{r} \times \mathbf{a} = \mathbf{r} \times \mathbf{a}. \end{aligned}$$

However, since \mathbf{r} and \mathbf{a} are parallel, $\mathbf{r} \times \mathbf{a} = \mathbf{0}$, and it follows that $d\mathbf{L}/dt = 0$, i.e., \mathbf{L} is constant. In particular, the direction of $\mathbf{L} = \mathbf{r} \times \mathbf{v}$ does not change, so since $\mathbf{r} \perp \mathbf{L}$, it follows that the position vector is constrained to the plane perpendicular to \mathbf{L} and containing the origin.

The quantity $\mathbf{r} \times \mathbf{v}$ seems to have been pulled from a hat, but you will learn in physics how the related quantity $\mathbf{r} \times (m\mathbf{v})$, which is called *angular momentum*, is used to make sense of certain aspects of motion.



Exercises for 1.4.

- Find $\mathbf{a} \times \mathbf{b}$ for the following vector pairs:
 - $\mathbf{a} = \langle 4, -2, 0 \rangle$, $\mathbf{b} = \langle 2, 1, -1 \rangle$
 - $\mathbf{a} = \langle 3, 3, 3 \rangle$, $\mathbf{b} = \langle 4, -3, 2 \rangle$
 - $\mathbf{a} = 2\mathbf{i} + 3\mathbf{j} + 4\mathbf{k}$, $\mathbf{b} = \mathbf{i} - 3\mathbf{j} + 4\mathbf{k}$
- Use the vector product to find the areas of the following figures.
 - The parallelogram with vertices $(0, 0, 0)$, $(1, 1, 0)$, $(1, 2, 1)$ and $(0, 1, 1)$. (Perhaps you should first check that this is a parallelogram.)
 - The triangle with vertices $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$.
- Prove that the vector product is not associative by calculating $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$ and $(\mathbf{a} \times \mathbf{b}) \times \mathbf{c}$ for $\mathbf{a} = \mathbf{i}$, $\mathbf{b} = \mathbf{c} = \mathbf{j}$.
- Show that if \mathbf{a} , \mathbf{b} , and \mathbf{c} are mutually perpendicular, then $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{0}$. What is the analogous geometric situation?
- 3×3 determinants are defined by the formula

$$\det \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix} = a_1 b_2 c_3 + a_2 b_3 c_1 + a_3 b_1 c_2 - a_3 b_2 c_1 - a_2 b_1 c_3 - a_1 b_3 c_2.$$

(Can you see the pattern?) Using this rule, verify formulas (10) and (12) in this section.

- Find the volume of the parallelepiped spanned by the vectors $\langle 1, 1, 0 \rangle$, $\langle 0, 2, -2 \rangle$, and $\langle 1, 0, 3 \rangle$.
- A triangular kite is situated with its vertices at the points $(0, 0, 10)$, $(2, 1, 10)$, and $(0, 3, 12)$. A wind with velocity vector $20\mathbf{i} + 6\mathbf{j} + 4\mathbf{k}$ displaces the kite by blowing for $1/2$ second. Find the volume of the solid bounded by the initial and final positions of the kite. (Assume all the distance units are in feet.)

8. Prove the formula

$$(\mathbf{a} \times \mathbf{b}) \times \mathbf{c} = -(\mathbf{b} \cdot \mathbf{c})\mathbf{a} + (\mathbf{a} \cdot \mathbf{c})\mathbf{b}$$

as follows. Since the quantities on both sides of the equation are defined geometrically, you can use whatever coordinate system you find convenient. Choose the coordinate system so that \mathbf{a} points along the x -axis and \mathbf{b} lies in the x, y -plane. Under these assumptions, \mathbf{a} has components $\langle a_1, 0, 0 \rangle$, \mathbf{b} has components $\langle b_1, b_2, 0 \rangle$, and \mathbf{c} has components $\langle c_1, c_2, c_3 \rangle$. Show that both sides of the equation have components

$$\langle -a_1 b_2 c_2, a_1 b_2 c_1, 0 \rangle.$$

9. Verify the formula.

$$|\mathbf{a} \times \mathbf{b}|^2 = |\mathbf{a}|^2 |\mathbf{b}|^2 - (\mathbf{a} \cdot \mathbf{b})^2.$$

Hint: Use the definitions in terms of the sine and cosine of the included angle θ .

10. Suppose a particle moves in such a way that $\mathbf{r} \times \mathbf{v}$ is a constant vector. Show that at each point of the path either \mathbf{r} or \mathbf{a} is zero or they are parallel.
11. One may try to generalize the vector product to higher dimensions. One plausible approach would be to consider all 2×2 determinants (with appropriate signs) which one can extract from the array

$$\begin{bmatrix} a_1 & a_2 & \dots & a_n \\ b_1 & b_2 & \dots & b_n \end{bmatrix}.$$

You learned in high school that there are $\frac{n(n-1)}{2}$ ways to choose 2 things from a set of n things, so that is the number of components. Show that $\frac{n(n-1)}{2} = n$ only in the case $n = 3$. (The moral is that we may be able to define a ‘cross product’ for dimensions other than 3, but only in that case will we get something of the same dimension.)

1.5 Geometry of Lines and Planes

We depart for the moment from the study of concepts of immediate use in dynamics to discuss some analytic geometry in space. The notions we shall introduce have a variety of applications, but, more important, they will help you develop your spatial intuition and teach you something about expressing such intuition analytically.

Lines We saw before that the motion of a particle moving with constant speed

on a line is described by a vector equation of the form

$$\mathbf{r} = t\mathbf{v} + \mathbf{r}_0$$

where \mathbf{v} is the constant velocity vector pointing along the line in the direction of motion, and \mathbf{r}_0 is its position at $t = 0$. (Previously \mathbf{v} was denoted \mathbf{v}_0 , but since the velocity is constant, we can use either, and we drop the subscript to save writing.)

It is often more convenient to use a slightly different form of this equation

$$\mathbf{r} = (t - t_0)\mathbf{v} + \mathbf{r}_0.$$

Here, ‘ t ’ is replaced by ‘ $t - t_0$ ’ which denotes the time ‘elapsed’ since some initial time t_0 , and \mathbf{r}_0 denotes the position at $t = t_0$. If you multiply this out, you will see it is really the same as the previous equation with \mathbf{r}_0 being replaced by the constant vector $-t_0\mathbf{v} + \mathbf{r}_0$. (Why is ‘elapsed’ in quotes? Think about the case $t < t_0$.)

There are a couple of points that should be stressed at this point. First, the above equations apply in the plane as well as in space. You should think about how the ordinary analytic geometry of lines in the plane can be related to this approach. Secondly, while we have been thinking of t as representing “time”, it is not absolutely necessary to do so. It is better in some circumstances to think of it as just another variable which ranges over the *domain* of a vector function described by the equation

$$\mathbf{r} = \mathbf{r}(t) = t\mathbf{v} + \mathbf{r}_0,$$

and the line is the *image* of this function, i.e., the set of all $\mathbf{r}(t)$ so obtained. Such a variable is called a *parameter*, and the associated description of the line is called a *parametric representation*. This is a more static point of view since we need not think of the line as traced out by a moving particle. Also, there is no need to use the symbol ‘ t ’ for the parameter. Other letters, such as ‘ s ’ or ‘ x ’ are perfectly acceptable, and in some circumstances may be preferable because of geometric connotations in the problem.

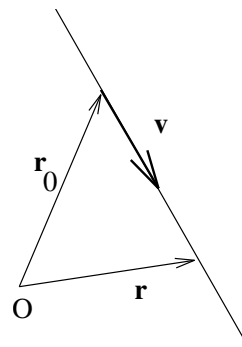
Example 7 Consider the line through the points P_0 and P_1 with respective coordinates $(1, 1, 0)$ and $(0, 2, 2)$.

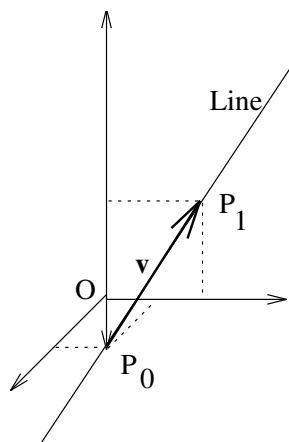
Choose $\mathbf{v} = \overrightarrow{P_0P_1}$. This is certainly a vector pointing along the line, and it does not matter what its magnitude is. (Thinking in kinematic imagery, we don’t care how fast the line is traced out.) Its components are $\langle 0 - 1, 2 - 1, 2 - 0 \rangle = \langle -1, 1, 2 \rangle$. Since the line passes through P_0 , we may choose $\mathbf{r}_0 = \overrightarrow{OP_0}$ which has components $\langle 1, 1, 0 \rangle$. With some abuse of notation, we may write the parametric equation of the line

$$\mathbf{r} = t\langle -1, 1, 2 \rangle + \langle 1, 1, 0 \rangle = \langle -t + 1, t + 1, 2t \rangle$$

or

$$\mathbf{r} = (1 - t)\mathbf{i} + (1 + t)\mathbf{j} + (2t)\mathbf{k}.$$





This can also be written as 3 component equations

$$\begin{aligned}x &= 1 - t, \\y &= 1 + t, \\z &= 2t.\end{aligned}$$

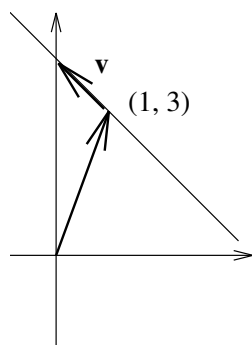
Note that there are other ways we could have approached this problem. We could have reversed the roles of P_0 and P_1 , we could have used $\frac{1}{2}\mathbf{v}$ instead of \mathbf{v} , etc. Each of these would have produced a valid parametric equation for the line, but they would all have looked different.

Example 8 Consider the line in the plane with parametric equation

$$\mathbf{r} = t\langle -1, 1 \rangle + \langle 1, 3 \rangle$$

or

$$\begin{aligned}x &= -t + 1, \\y &= t + 3.\end{aligned}$$



We may eliminate t algebraically by writing

$$t = 1 - x = y - 3$$

whence we obtain $y = -x + 4$. This is the usual way of representing a line in the plane by an equation of the form $y = mx + b$, where m is the slope, and b is the y -intercept.

Eliminating the parameter makes sense for lines in space, but unfortunately it does not lead to a single equation. The vector equation $\mathbf{r} = t\mathbf{v} + \mathbf{r}_0$ can be written as 3 scalar equations

$$\begin{aligned}x &= ta + x_0, \\y &= tb + y_0, \\z &= tc + z_0\end{aligned}$$

where $\mathbf{v} = \langle a, b, c \rangle$. If none of the components of \mathbf{v} are zero, we can solve these equations for t to obtain $t = \frac{x - x_0}{a}$, $t = \frac{y - y_0}{b}$, and $t = \frac{z - z_0}{c}$. Eliminating t yields so called *symmetric equations* of the line

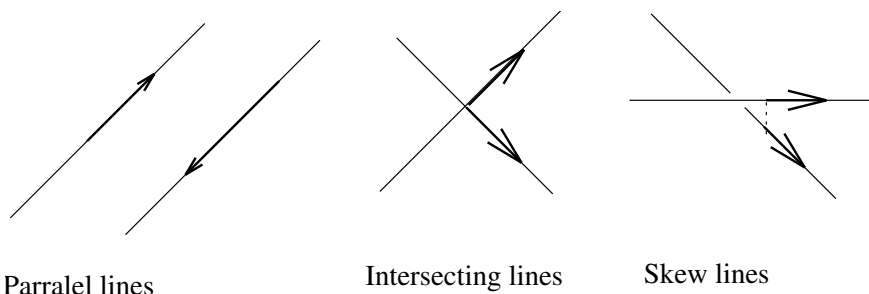
$$\frac{x - x_0}{a} = \frac{y - y_0}{b} = \frac{z - z_0}{c}.$$

These have the advantage that the components of \mathbf{v} and the coordinates of P_0 are clearly displayed. However, symmetric equations are not directly applicable if, as in Example 7, one or more of the components of \mathbf{v} are zero.

Example 7 revisited We found $\mathbf{v} = \langle -1, 1, 2 \rangle$ and $\mathbf{r}_0 = \langle 1, 1, 0 \rangle$ so the symmetric equations are

$$\frac{x-1}{-1} = \frac{y-1}{1} = \frac{z}{2}.$$

The geometry of lines in space is a bit more complicated than that of lines in the plane. Lines in the plane either intersect or are parallel. In space, we have to be a bit more careful about what we mean by ‘parallel lines’, since lines with entirely different directions can still fail to intersect.



Example 9 Consider the lines described by

$$\begin{aligned}\mathbf{r} &= t\langle 1, 3, -2 \rangle + \langle 1, 2, 1 \rangle \\ \mathbf{r} &= t\langle -2, -6, 4 \rangle + \langle 3, 1, 0 \rangle.\end{aligned}$$

They have parallel directions since $\langle -2, -6, 4 \rangle = -2\langle 1, 3, -2 \rangle$. Hence, in this case we say the lines are *parallel*. (How can we be sure the lines are not the same?)

Example 10 Consider the lines

$$\begin{aligned}\mathbf{r} &= t\langle 1, 3, -2 \rangle + \langle 1, 2, 1 \rangle \\ \mathbf{r} &= t\langle 0, 2, 3 \rangle + \langle 0, 3, 9 \rangle.\end{aligned}$$

They are not parallel because neither of the vectors \mathbf{v} is a multiple of the other. They may or may not intersect. (If they don't, we say the lines are *skew*.) How can we find out? One method, is to set them equal and see if we can solve for the point of intersection. There is one tricky point here. If we think of the parameter t as time, even if the lines do intersect, there is no guarantee that particles moving on these lines would arrive at the point of intersection *at the same instant*. Hence, the way to proceed is to introduce a second parameter, call it s , for one of the lines, and then try to solve for the point of intersection. Thus, we want

$$\mathbf{r} = t\langle 1, 3, -2 \rangle + \langle 1, 2, 1 \rangle = s\langle 0, 2, 3 \rangle + \langle 0, 3, 9 \rangle,$$

which after collecting terms yields

$$\langle t+1, 3t+2, -2t+1 \rangle = \langle 0, 2s+3, 3s+9 \rangle.$$

Picking out the components yields three equations

$$\begin{aligned}t + 1 &= 0 \\3t + 2 &= 2s + 3 \\-2t + 1 &= 3s + 9\end{aligned}$$

in 2 unknowns s and t . This is an *overdetermined* system, and it may or may not have a consistent solution. In this case, the first two equations yield $t = -1$ and $s = -2$. Putting these values in the last equation yields $(-2)(-1) + 1 = 3(-2) + 9$ which checks. Hence, the equations are consistent, and the lines do intersect. To find the point of intersection, put $t = -1$ in the equation for the first line (or $s = -2$ in that for the second) to obtain $\langle 0, -1, 3 \rangle$.

Example 11 Consider the lines

$$\begin{aligned}\mathbf{r} &= t\langle 1, 3, -2 \rangle + \langle 1, 2, 1 \rangle \\ \mathbf{r} &= s\langle 0, 2, 3 \rangle + \langle 0, 3, 8 \rangle.\end{aligned}$$

We argue exactly as above, except in this case, we obtain component equations

$$\begin{aligned}t + 1 &= 0 \\3t + 2 &= 2s + 3 \\-2t + 1 &= 3s + 8\end{aligned}$$

Again, the first two equations yield $t = -1, s = -2$, but these values are not consistent with the third equation. Hence, the lines are skew, they are not parallel and they don't intersect.

.

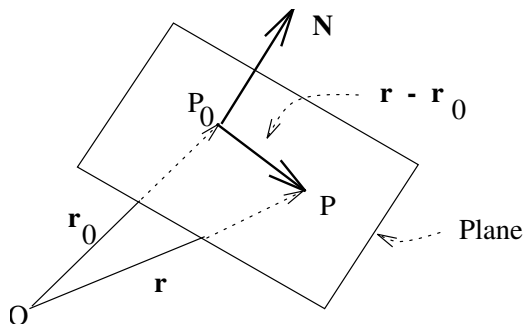
Planes A plane in space may be characterized geometrically in several different ways. Here are some of the most common characterizations.

1. Any three non-collinear points P_1, P_2 , and P_3 (points not on a common line) determine a plane.
2. There is a unique plane passing through a given point P_0 and perpendicular to a given line l . (A plane is perpendicular to a line l if each line in the plane through the point of intersection with l is perpendicular to l .)
3. Any two distinct lines l_1 and l_2 which intersect determine a plane.
4. A line l and a point P_0 not on l determine a plane.

We want to describe planes analytically. For this, it is best to start with the second characterization. Let P_0 with coordinates (x_0, y_0, z_0) be the given point. Clearly, we will get the same plane if we replace the perpendicular line l with any line parallel

to l , so we may as well assume that l passes through P_0 . Choose a vector \mathbf{N} pointing in the direction of l . If P is any point in the plane, then the displacement vector $\overrightarrow{P_0P}$ is perpendicular to \mathbf{N} . However, $\overrightarrow{P_0P} = \mathbf{r} - \mathbf{r}_0$, so the perpendicularity may be described algebraically by

$$\mathbf{N} \cdot (\mathbf{r} - \mathbf{r}_0) = 0. \quad (14)$$



Suppose \mathbf{N} has components $\langle a, b, c \rangle$. Since $\mathbf{r} - \mathbf{r}_0$ has components $\langle x - x_0, y - y_0, z - z_0 \rangle$, this equation may be rewritten

$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0. \quad (15)$$

This is called a *normal form* of an equation of a plane.

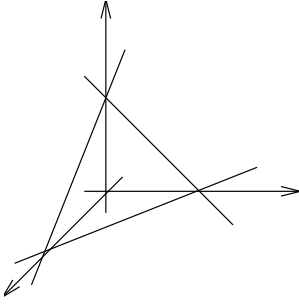
Example 12 We shall find an equation for the plane through the points with coordinates $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. By symmetry, the angles which a normal vector makes with the coordinate axes (called its *direction angles*), should be equal, so the vector \mathbf{N} with components $\langle 1, 1, 1 \rangle$ should be perpendicular to this plane. For P_0 , we could use any of the three points; we choose it to be the point with coordinates $(1, 0, 0)$. With these choices, we get the normal form of the equation of the plane

$$1(x - 1) + 1(y - 0) + 1(z - 0) = 0$$

which can be rewritten more simply

$$\begin{aligned} x + y + z - 1 &= 0 \\ \text{or } x + y + z &= 1. \end{aligned}$$

Note that we chose the normal direction by being clever, but there is a quite straightforward way to do it. If the three points are denoted P_0, P_1 and P_2 , then the directed line segments $\overrightarrow{P_0P_1}$ and $\overrightarrow{P_0P_2}$ lie in the plane, so the vector product $\overrightarrow{P_0P_1} \times \overrightarrow{P_0P_2}$ is perpendicular to the plane. In this case, $\overrightarrow{P_0P_1}$ has components $\langle 0 - 1, 1 - 0, 0 - 0 \rangle =$



$\langle -1, 1, 0 \rangle$, and $\overrightarrow{P_0P_2}$ has components $\langle 0 - 1, 0 - 0, 1 - 0 \rangle = \langle -1, 0, 1 \rangle$. Hence, the cross product has components $\langle -1, -1, -1 \rangle$. This yields the normal form

$$\begin{aligned} -1(x - 1) + (-1)(y - 0) + (-1)(z - 0) &= 0 \\ \text{or} \quad -x - y - z &= -1. \end{aligned}$$

As the above example illustrates, there is nothing unique about the equation of a plane. For example, if \mathbf{N} with components $\langle a, b, c \rangle$ is a normal vector, so is $s\mathbf{N}$ with components $\langle sa, sb, sc \rangle$ for any non-zero scalar s . Replacing \mathbf{N} by $s\mathbf{N}$ just multiplies the normal form by the factor s , and clearly this does not change the locus of the equation. (The locus of an equation is the set of all points whose coordinates satisfy the equation.)

In general, the normal form may be rewritten

$$\begin{aligned} ax + by + cz - ax_0 - by_0 - cz_0 &= 0 \\ \text{or} \quad ax + by + cz &= d \end{aligned}$$

where $d = ax_0 + by_0 + cz_0$. Conversely, the locus of any such linear equation, $ax + by + cz = d$, in which not all of the coefficients a, b , and c are zero, is a plane. This is not hard to prove in general, but we illustrate it instead in an example.

Example 13 Consider the locus of the equation

$$2x - 3y + z = 6.$$

The choice of a normal vector is clear, $\mathbf{N} = \langle 2, -3, 1 \rangle$. Hence, to express the above equation in normal form, it suffices to find one point P_0 in the plane. This could be done in many different ways, but one method that works in this case would be to set $x = y = 0$ and to solve for z . In this case, we get $z = 6$, so the point with coordinates $(0, 0, 6)$ lies in the plane. Write

$$\begin{aligned} 2x - 3y + z &= 6 && \text{for a general point} \\ 2(0) - 3(0) + 6 &= 6 && \text{for the specific point} \end{aligned}$$

and then subtract to obtain

$$2(x - 0) - 3(y - 0) + 1(z - 6) = 0.$$

That is an equation for the same plane in normal form. It may also be written

$$\langle 2, -3, 1 \rangle \cdot \langle x - 0, y - 0, z - 6 \rangle = 0$$

to express the perpendicularity of $\overrightarrow{P_0P}$ to \mathbf{N} .

You should convince yourself that this method works in complete generality. You should also ask what you might do for an equation of the form $2x + 3y = 12$ where you can't set $x = y = 0$ and solve for z .

Various special cases where one or more of the coefficients in the linear equation $ax+by+cz = d$ vanish represent special orientations of the plane with respect to the coordinate axes. For example, $x = d$ has as locus a plane parallel to the y, z -plane (i.e., perpendicular to the x -axis), and passing through the point $(d, 0, 0)$. $ax+by = d$ would have as locus a plane *perpendicular* to the x, y -plane and intersecting it in the line in that plane with equation $ax + by = c$.

It is important to note that the equation $ax+by = d$ does not by itself specify a locus. You must say whether you are considering a locus in the plane or in space. You should consider other variations on these themes and draw representative diagrams to make sure you understand their geometric significance.

We saw that the coefficients a, b , and c in the linear equation may be chosen to be the components of a normal vector \mathbf{N} . The significance of the constant d is a bit trickier. We have

$$d = ax_0 + by_0 + cz_0 = \mathbf{N} \cdot \mathbf{r}_0$$

It is easiest to understand this if \mathbf{N} is a unit vector, i.e., $a^2 + b^2 + c^2 = 1$, so we suppose that to be the case. Then $\mathbf{N} \cdot \mathbf{r}_0$ is the *projection* of the position vector \mathbf{r}_0 of the reference point P_0 on the normal direction. Since the vector \mathbf{N} can be placed wherever convenient, we move it so its tail is at the origin. Then it is clear that this projection is just the *signed* distance of the plane to the origin. The sign depends on whether the origin is on the side of the plane pointed at by the normal or the other side. (How?)

This reasoning can be extended further as follows. Let P_1 be any point, not necessarily the origin, and let (x_1, y_1, z_1) be its coordinates. Then the projection of $\overrightarrow{P_0 P_1} = \mathbf{r}_1 - \mathbf{r}_0$ on the vector \mathbf{N} gives the signed perpendicular distance from the point P_1 to the plane. Thus, if \mathbf{N} is a unit vector, the *distance* is given by

$$D = |\mathbf{N} \cdot (\mathbf{r}_1 - \mathbf{r}_0)|.$$

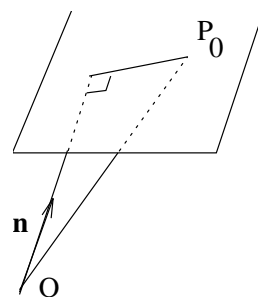
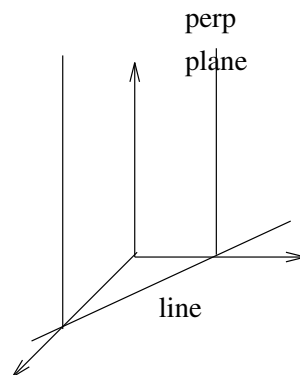
If \mathbf{N} is not a unit vector, we may replace it by $\mathbf{n} = \frac{1}{|\mathbf{N}|}\mathbf{N}$. This yields the formula for that distance

$$\begin{aligned} D &= \frac{1}{|\mathbf{N}|} |\mathbf{N} \cdot \mathbf{r}_1 - \mathbf{N} \cdot \mathbf{r}_0| \\ &= \frac{1}{\sqrt{a^2 + b^2 + c^2}} |ax_1 + by_1 + cz_1 - d|. \end{aligned} \quad (16)$$

Example 14 The distance of the point $(-1, 1, 0)$ to the plane with equation $x + y + z = 1$ is

$$\frac{1}{\sqrt{3}} |1(-1) + 1(1) + 1(0) - 1| = \frac{1}{\sqrt{3}}.$$

Note that the sign inside the absolute values is negative which reflects the fact that $(-1, 1, 0)$ is on the side of the plane opposite to that pointed at by $\langle 1, 1, 1 \rangle$.



Intersection of Planes In general, two distinct planes in space are either parallel or they intersect in a line. In the first case, the normal vectors are parallel.

Example 15 The loci of

$$2x - 3y + 2z = 12$$

$$\text{and} \quad -6x + 9y - 6z = 12$$

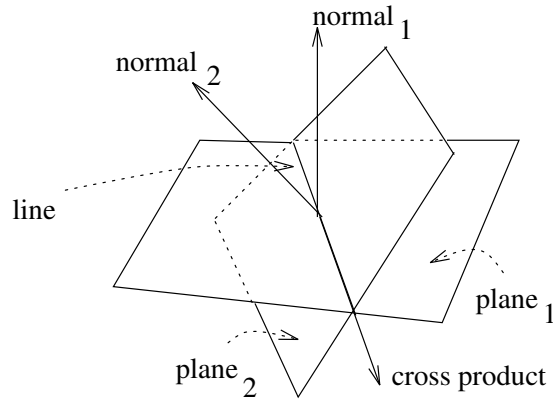
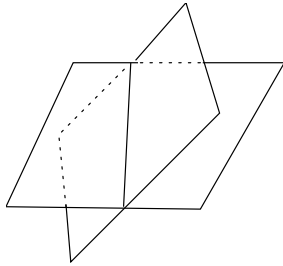
are parallel. Indeed the indicated normal vectors are $\langle 2, -3, 2 \rangle$ and $\langle -6, 9, -6 \rangle$ which are multiples of one another, so they are parallel. (How do you know the planes are not identical?)

In the second case, there is a simple way to find an equation for the line of intersection of the two planes.

Example 16 We find the line of intersection of the planes which are loci of the linear equations

$$6x + 2y - z = 2$$

$$x - 2y + 3z = 5. \tag{17}$$



The line of intersection of the two planes is perpendicular to normals to both planes. Thus it is perpendicular to $\langle 6, 2, -1 \rangle$ and also to $\langle 1, -2, 3 \rangle$. Thus, the cross product

$$\mathbf{v} = \langle 6, 2, -1 \rangle \times \langle 1, -2, 3 \rangle = \langle 4, -19, -14 \rangle$$

is parallel to the desired line. To find a parametric representation of the line, it suffices to find one point P_0 in the line. We do this as follows. In the system of equations (17), choose some particular value of z , say $z = 0$. this yields the system

$$6x + 2y = 2$$

$$x - 2y = 5$$

which may be solved by the usual methods of high school algebra to obtain $x = 1, y = -2$. That tells us that the point P_0 with coordinates $(1, -2, 0)$ is a point on the line. Hence,

$$\mathbf{r} = t\mathbf{v} + \mathbf{r}_0 = t\langle 4, -19, -14 \rangle + \langle 1, -2, 0 \rangle = \langle 4t + 1, -19t - 2, -14t \rangle$$

is a vector parametric representation of the line. (What would you have done if x did not appear in either equation, i.e., appeared with 0 coefficients? In that case, you would not be able to set $z = 0$ and solve for x .)

Note that the symmetric equations of a line

$$\frac{x - x_0}{a} = \frac{y - y_0}{b} = \frac{z - z_0}{c}$$

may be interpreted in this light as asserting the intersection of two planes. The first equality could be rewritten

$$bx - ay = bx_0 - ay_0$$

which defines a plane perpendicular to the x, y -plane. Similarly, the second equality could be rewritten

$$cy - bz = cy_0 - bz_0$$

which defines a plane perpendicular to the y, z -plane. The line is the intersection of these two planes.

In general, the intersection of three planes in space is a point, but there are many special cases where it is not. For example, the planes could be parallel, or the line of intersection of two of the planes might be parallel to the third, in which case they do not intersect at all. Similarly, all three planes could intersect in a common line. This geometry is reflected in the algebra. If the planes intersect in a single point, their defining equations provide a system of 3 equations in 3 unknowns which have a unique solution. In the second case, the equations are inconsistent, and there is no solution. In the third case there are infinitely many solutions. We shall return to a study of such issues when we study linear algebra and the solution of systems of linear equations.

Elaborations The analytic geometry of lines and planes in space is quite interesting and presents us with many challenging problems. We shall not go into this in detail in this course, but you might profit from trying some of the exercises. One interesting problem, is to determine the perpendicular distance between skew lines in space. If the lines are given parametrically by equations

$$\mathbf{r} = t\mathbf{v}_1 + \mathbf{r}_1$$

$$\mathbf{r} = t\mathbf{v}_2 + \mathbf{r}_2,$$

then the distance between the lines is given by

$$\frac{|(\mathbf{r}_2 - \mathbf{r}_1) \cdot (\mathbf{v}_2 \times \mathbf{v}_1)|}{|\mathbf{v}_2 \times \mathbf{v}_1|}. \quad (18)$$

See if you can derive this formula! (There are some hints in the exercises.)

Exercises for 1.5.

- Find a vector (parametric) equation for the line that
 - passes through $(0, 0, 0)$ and is parallel to $\mathbf{v} = 3\mathbf{i} + 4\mathbf{j} + 5\mathbf{k}$,
 - passes through $(1, 2, 3)$ and $(4, -1, 2)$,
 - passes through $(1, 1)$ and is perpendicular to $\mathbf{v} = \langle 3, 1 \rangle$,
 - passes through $(9, -2, 3)$ and $(1, 2, 3)$.
- Determine if the lines with the following vector equations intersect: $\mathbf{r} = \langle 1, -1, 2 \rangle + t\langle 2, 1, 1 \rangle$, $\mathbf{r} = \langle 0, 1, 1 \rangle + t\langle 1, 0, -1 \rangle$.
- Find an equation for the plane with the given normal vector and containing the given point: (a) $\mathbf{N} = \langle 2, -1, 3 \rangle$, $P(1, 2, 0)$,
(b) $\mathbf{N} = \langle 1, 0, 3 \rangle$, $P(2, 4, 5)$.
- Write parametric and symmetric equations for the line which
 - passes through $(0, 1, 2)$ and is perpendicular to the yz -plane,
 - passes through $(5, 2, -1)$ and is perpendicular to the plane with equation $3x + 4y - z = 2$,
 - passes through $(1, 3, 0)$ and is parallel to the line with parametric equations $x = t$, $y = t - 1$, $z = 2t + 3$.
- Write an equation for the plane that
 - passes through $(1, 4, 3)$ and is perpendicular to the line with equation $\mathbf{r} = \langle 1 + t, 2 + 4t, t \rangle$.
 - passes through the origin and is parallel to the plane with equation $3x + 4y - 5z = -1$,
 - passes through $(0, 0, 0)$, $(1, -2, 8)$, $(-2, -1, 3)$.
- Find a vector (parametric) equation for the line of intersection of the planes with equations $2x + 3y - z = 1$ and $x - y - z = 0$.
- Find the angle between the normals to the following planes:
 - the planes with equations $x + 2y - z = 2$ and $2x - y + 3z = 1$,
 - the plane with equation $2x + 3y - 5z = 0$ and the plane containing the points $(1, 3, -2)$, $(5, 1, 3)$, and $(1, 0, 1)$.
- Use formula (3) for the perpendicular distance from the point $P_0(x_0, y_0, z_0)$ to the plane $ax + by + cz = d$ to find the distance from
 - the origin to the plane $x - 3z = -2$,
 - the point $(-1, 2, 1)$ to the plane $3x + 4y - 5z = -2$.

9. Show that the distance between the lines given parametrically by equations

$$\mathbf{r} = t\mathbf{v}_1 + \mathbf{r}_1$$

$$\mathbf{r} = t\mathbf{v}_2 + \mathbf{r}_2,$$

is

$$\frac{|(\mathbf{r}_2 - \mathbf{r}_1) \cdot (\mathbf{v}_2 \times \mathbf{v}_1)|}{|\mathbf{v}_2 \times \mathbf{v}_1|}.$$

Hint: Consider the line segment Q_1Q_2 from one line to the other which is perpendicular to both. Note that $\frac{1}{|\mathbf{v}_1 \times \mathbf{v}_2|} \mathbf{v}_1 \times \mathbf{v}_2$ is a unit vector parallel to Q_1Q_2 . Convince yourself that Q_1Q_2 is the projection of the line segment connecting the endpoints of \mathbf{r}_1 and \mathbf{r}_2 onto this unit vector.

10. A projectile follows the path

$$\mathbf{r} = 1000t\mathbf{i} + (1000t - 16t^2)\mathbf{j}.$$

At $t_0 = 1$, an anti-gravity ray zaps the projectile, making it impervious to gravitation, so it proceeds off on the tangent line at the point. In general, the equation for the tangential motion would be

$$\mathbf{r} = (t - t_0)\mathbf{v}_0 + \mathbf{r}_0$$

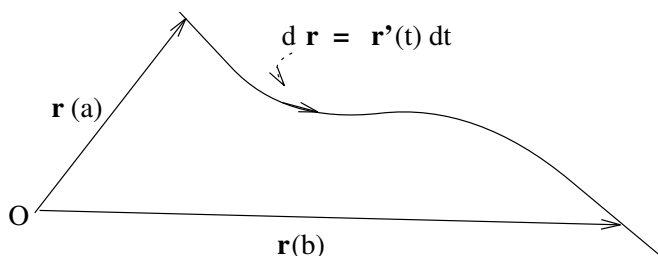
where \mathbf{r}_0 is the position vector at t_0 and \mathbf{v}_0 is the instantaneous velocity vector there. Where will the projectile be after one additional second?

1.6 Integrating on Curves

Arc Length Suppose a particle follows a path in space (or in the plane) described by $\mathbf{r} = \mathbf{r}(t)$. Suppose moreover it starts for $t = a$ at $\mathbf{r}(a)$ and ends for $t = b$ at $\mathbf{r}(b)$. We want to calculate the total distance it travels on the curve. The correct formula for this distance is

$$L = \int_a^b |\mathbf{r}'(t)| dt \quad (19)$$

where $\mathbf{r}'(t) = \frac{d\mathbf{r}}{dt}$ is the velocity vector at time t (when the particle has position vector $\mathbf{r}(t)$).



This makes sense, because $|\mathbf{r}'(t)|$ should be the *speed* of the particle at time t , so $|\mathbf{r}'(t)|dt$ should be a good approximation to the distance traveled in a small time interval dt . (Integration amounts to adding up the incremental distances.)

Example 17 Let $\mathbf{r} = R \cos t \mathbf{i} + R \sin t \mathbf{j}$ with $0 \leq t \leq 2\pi$. As we saw earlier, this describes one circuit of a circle of radius R centered at the origin. We have

$$\frac{d\mathbf{r}}{dt} = -R \sin t \mathbf{i} + R \cos t \mathbf{j},$$

so $|\mathbf{r}'(t)| = \sqrt{R^2 \cos^2 t + R^2 \sin^2 t} = R$. Hence,

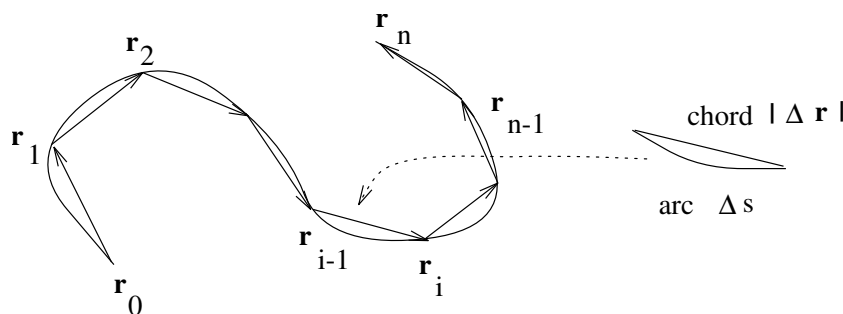
$$L = \int_0^{2\pi} R dt = R t \Big|_0^{2\pi} = 2\pi R$$

just as we would expect.

Formula (19) may be taken as the *definition* of the arc length of the curve, but a little more discussion will clarify its ramifications. Choose $n+1$ points $\mathbf{r}_0, \mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$ on the curve (as indicated in the diagram) by choosing a partition of the time interval

$$a = t_0 < t_1 < t_2 < \dots < t_{n-1} < t_n = b,$$

and letting the i th position vector $\mathbf{r}_i = \mathbf{r}(t_i)$.



For two neighboring points \mathbf{r}_{i-1} and \mathbf{r}_i on the curve, the displacement vector $\Delta \mathbf{r}_i = \mathbf{r}_i - \mathbf{r}_{i-1}$ is represented in the diagram by the *chord* connecting the points. Hence, for any plausible definition of length, the length of arc Δs_i on the curve from \mathbf{r}_{i-1} to \mathbf{r}_i is approximated quite well by the length of the chord, at least if the points are closely spaced. Thus

$$\Delta s_i \approx |\Delta \mathbf{r}_i|.$$

(‘ \approx ’ means ‘is approximately equal to’.) Hence, if we add everything up, we get

$$L = \sum_{i=1}^n \Delta s_i \approx \sum_{i=1}^n |\Delta \mathbf{r}_i|,$$

that is, *the length of the curve is approximated by the length of the polygonal path made up of the chords*. To relate the latter length to the integral, note that if we put $\Delta t_i = t_i - t_{i-1}$, then for closely spaced points

$$\Delta \mathbf{r}_i \approx \mathbf{r}'(t_{i-1}) \Delta t_i.$$

Putting this in the prior formula for L yields

$$L \approx \sum_{i=1}^n |\mathbf{r}'(t_{i-1})| \Delta t_i,$$

and the sum on the right is one of the approximating (Riemann) sums which approach the integral $\int_a^b |\mathbf{r}'(t)| dt$ as $n \rightarrow \infty$.

In our discussions, we have denoted the independent variable by ' t ' and thought of it as time. Although this is helpful in kinematics, from a mathematical point of view this is not necessary. Any vector valued function $\mathbf{r} = \mathbf{r}(u)$ of a variable u can be thought of as yielding a curve as u ranges through the set of values in the domain of the function. As mentioned elsewhere, the variable u is usually called a *parameter*, and often it will have some geometric or other significance. For example, in Example 17, it might have made more sense to call the parameter θ (instead of t) and to think of it as the angle the position vector makes with the positive real axis.

Example 18 We shall find the length of the parabola with equation $y = x^2$ on the interval $-1 \leq x \leq 1$. One parametric representation of the parabola would be $\mathbf{r} = x\mathbf{i} + y\mathbf{j} = t\mathbf{i} + t^2\mathbf{j}$ where $-1 \leq t \leq 1$. However, there is no real need to introduce a new variable; we might just as well use x , and write instead

$$\mathbf{r} = x\mathbf{i} + x^2\mathbf{j}, \quad -1 \leq x \leq 1.$$

The formula (19) still applies with t replaced by x . Then,

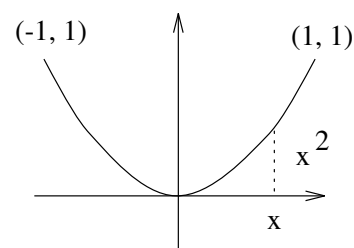
$$\frac{d\mathbf{r}}{dx} = \mathbf{i} + (2x)\mathbf{j},$$

so $|\mathbf{r}'(x)| = \sqrt{1 + 4x^2}$. Hence,

$$L = \int_{-1}^1 \sqrt{1 + 4x^2} dx = \sqrt{5} + \frac{1}{4} \ln \frac{\sqrt{5} + 2}{\sqrt{5} - 2}.$$

You should ponder what we just did, and convince yourself there is nothing peculiar about the use of x as parameter.

As the previous example shows, calculation of lengths often results in difficult integrals because of the square root. Recourse to integral tables or appropriate computer software is advised.



There is one very tricky point in using formula (19) to *define* length of arc on a curve. The same curve might be given by two different parametric representations $\mathbf{r} = \mathbf{r}_1(u)$, $a \leq u \leq b$ and $\mathbf{r} = \mathbf{r}_2(v)$, $c \leq v \leq d$. How can we be sure we will get the *same* answer for the length if we compute

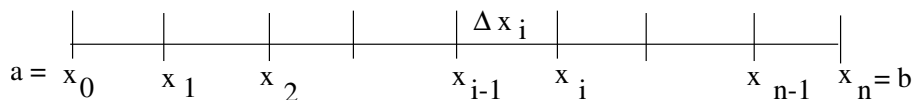
$$\int_a^b |\mathbf{r}'_1(u)| du \quad \text{and} \quad \int_c^d |\mathbf{r}'_2(v)| dv?$$

The superficial answer is that in general we won't always get the same answer. For example, a circle might be represented parametrically in two different ways, one which traverses the circle once, and one which traverses it twice, and the answers for the length would be different. The curve *as a point set* does not by itself determine the answer if part or all of it is covered more than once in a given parametric representation. This issue, while something to worry about in specific examples, is really a red herring. If we are careful to choose parametric representations which trace each part of the curve *exactly once*, then we expect always to get the same answer. (Such representations are called *one-to-one*.) Our reason for believing this is that we think of length as a physical quantity which can be measured by tapes (or more sophisticated means), so we expect the mathematical theory to fall in line with reality. Unfortunately, it is not logically or philosophically legitimate to identify the world of mathematical discourse with physical reality, so it is incumbent on mathematicians to *prove* that all one-to-one parametric representations yield the same answer. We won't actually do this in this course, but some of the relevant issues are discussed in the Exercises.

Line Integrals In elementary physics, the work W done by a constant force F exerted over a distance Δ is defined to be the product $F\Delta$.

In more complicated situations, physicists need to deal with forces which may vary or which may not point along the direction of motion. In addition, they have to worry about motion which is not constrained to a line. We shall investigate the mathematical concepts needed to generalize the concept of work in those situations. First suppose that the work is done along a line, but that the force varies with position on the line. If x denotes a coordinate describing that position, and the force is given by $F = F(x)$, then the work done going from $x = a$ to $x = b$ is

$$W = \int_a^b F(x) dx.$$



The justification for that formula is that if we think of the interval $[a, b]$ being partitioned into small segments by division points

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b,$$

then the work done going from x_{i-1} to x_i should be given approximately by $\Delta W_i \approx F(x_{i-1})\Delta x_i$ where $\Delta x_i = x_i - x_{i-1}$. Adding up and taking a limit yields the integral.

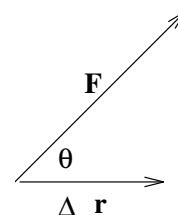
Example 19 Suppose $F(x) = -kx, 0 \leq x \leq D$. (This is the restoring force of a spring, where the constant k is the *spring constant*.) Then

$$W = \int_0^D (-kx)dx = -k \frac{x^2}{2} \Big|_0^D = -\frac{kD^2}{2}.$$

More generally, the force may not point in the direction of the displacement. If that is the case, we simply use the component of the force in the direction of the displacement. Thus, if the force \mathbf{F} is a constant vector, and the displacement is given by a vector $\Delta \mathbf{r}$, then the component in the direction of $\Delta \mathbf{r}$ is $|\mathbf{F}| \cos \theta$ and the work is $(|\mathbf{F}| \cos \theta)|\Delta \mathbf{r}|$, which you should recognize as the dot product

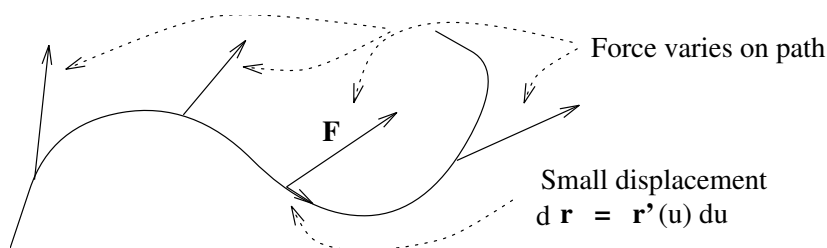
$$\mathbf{F} \cdot \Delta \mathbf{r}.$$

If the force is not constant, this formula will still be approximately valid if the magnitude of the displacement $\Delta \mathbf{r}$ is sufficiently small.



We are now ready to put all this together for the most general situation. Suppose we have a force \mathbf{F} which may vary from point to point, and it is exerted on a particle moving on a path given parametrically by $\mathbf{r} = \mathbf{r}(u), a \leq u \leq b$. (We assume also that the representation is one-to-one, although it turns out that this is not strictly necessary here.) The correct mathematical quantity to use for the work done is the integral

$$\int_a^b \mathbf{F}(\mathbf{r}(u)) \cdot \mathbf{r}'(u) du. \quad (20)$$



The idea behind this formula is that $d\mathbf{r} = \mathbf{r}'(u)du$ is the displacement along the curve produced by a small increment du in the parameter, so $\mathbf{F} \cdot d\mathbf{r} = \mathbf{F} \cdot \mathbf{r}'(u)du$ is a good approximation to the work done in that displacement. As usual, the integral sign suggests a summing up of the incremental contributions to the total work.

Example 20 Let \mathcal{C} be a circle of radius R centered at the origin and traversed in the counter-clockwise direction. \mathcal{C} may be represented parametrically by

$$\mathbf{r} = (R \cos \theta)\mathbf{i} + (R \sin \theta)\mathbf{j}, \quad 0 \leq \theta \leq 2\pi.$$

Suppose the force is given by $\mathbf{F} = -y\mathbf{i} + x\mathbf{j}$. To calculate the work by formula (20), we calculate

$$\mathbf{r}'(\theta) = (-R \sin \theta)\mathbf{i} + (R \cos \theta)\mathbf{j},$$

and note that since $x = R \cos \theta, y = R \sin \theta$ on the circle, we have

$$\mathbf{F} = -y\mathbf{i} + x\mathbf{j} = (-R \sin \theta)\mathbf{i} + (R \cos \theta)\mathbf{j}.$$

Hence,

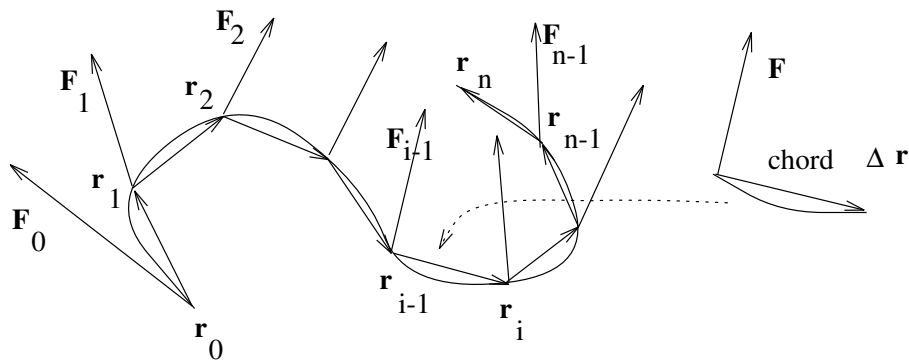
$$W = \int_0^{2\pi} (R^2 \cos^2 \theta + R^2 \sin^2 \theta) d\theta = \int_0^{2\pi} R^2 d\theta = 2\pi R^2.$$

Note that had we not substituted for x and y in terms of θ in the expression for \mathbf{F} , we would have gotten a meaningless answer with x 's and y 's in it. Generally, when working such problems, one must be sure that the relevant quantities in the integrand have all been expressed in terms of the parameter.

Some additional discussion will clarify the meaning of formula (20). Suppose a curve \mathcal{C} is represented parametrically by a vector function $\mathbf{r} = \mathbf{r}(u), a \leq u \leq b$, and suppose a force \mathbf{F} is defined on \mathcal{C} but may vary from point to point. Choose a sequence of points on the curve with position vectors $\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_n$ by subdividing the parameter domain

$$a = u_0 < u_1 < u_2 < \dots < u_{n-1} < u_n = b$$

and letting $\mathbf{r}_i = \mathbf{r}(u_i)$. Let $\Delta \mathbf{r}_i = \mathbf{r}_i - \mathbf{r}_{i-1}$.



Then it is plausible that the work done moving along the curve from \mathbf{r}_{i-1} to \mathbf{r}_i is approximated by $\mathbf{F}(\mathbf{r}_{i-1}) \cdot \Delta \mathbf{r}_i$, at least if $|\Delta \mathbf{r}_i|$ is small. Moreover, adding it all up,

we have

$$W \approx \sum_{i=1}^n \mathbf{F}(\mathbf{r}_{i-1}) \cdot \Delta \mathbf{r}_i.$$

Just as in the case of arc length, the right hand side approaches the integral in formula (20) as a limit as $n \rightarrow \infty$ and as the points get closer together.

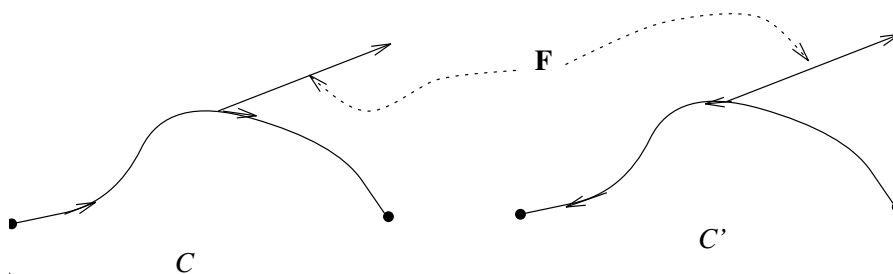
We introduce the following notation for the limit of the sum

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \lim \sum_{i=1}^n \mathbf{F}(\mathbf{r}_{i-1}) \cdot \Delta \mathbf{r}_i = \int_a^b \mathbf{F}(\mathbf{r}(u)) \cdot \mathbf{r}'(u) du,$$

and call it a *line integral*. It first appears in physics in discussions of work and energy where \mathbf{F} represents a force, but it may be thought of as a mathematical entity defined for any vector function of position \mathbf{F} .

The curve \mathcal{C} and ultimately the line integral have been discussed in terms of a specific parametric representation $\mathbf{r} = \mathbf{r}(u)$. As in the case of arc length, it is possible to show that different parametric representations produce the same answer *except for one additional difficulty*. In the formula for arc length, we dealt with $|\Delta \mathbf{r}|$, so the *direction* in which the curve was traced was not relevant. In the case of line integrals, we use $\mathbf{F} \cdot \Delta \mathbf{r}$, so reversing the direction changes the sign of $\Delta \mathbf{r}$ and hence it changes the sign of the line integral. Line integrals, in fact, must be defined for *oriented* curves for which one of the two possible directions has been specified. If \mathcal{C}' and \mathcal{C} are the same curve but traversed in opposite directions then

$$\int_{\mathcal{C}'} \mathbf{F} \cdot d\mathbf{r} = - \int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r}.$$

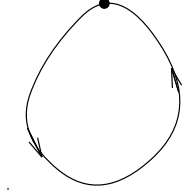


There are several different notations for line integrals. For example, you may see

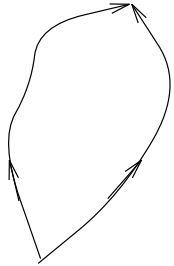
$$\int_C \mathbf{F} \cdot d\mathbf{l} \quad \text{or} \quad \int_C \mathbf{F} \cdot ds.$$

The ' ds ' suggests a vector version of the element of arc length $ds = |\mathbf{r}'(t)|dt$. We can also write formally, $\mathbf{F} = F_x \mathbf{i} + F_y \mathbf{j} + F_z \mathbf{k}$, $d\mathbf{r} = dx \mathbf{i} + dy \mathbf{j} + dz \mathbf{k}$, so $\mathbf{F} \cdot d\mathbf{r} =$

Start = Finish



B



A

 $F_x dx + F_y dy + F_z dz$, and

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_C F_x dx + F_y dy + F_z dz.$$

The notation

$$\oint \mathbf{F} \cdot d\mathbf{r}$$

is sometimes used to denote a line integral for an unspecified closed path (i.e., one which starts and ends at the same point.) Finally, in some cases the line integral depends only on the endpoints of A and B of the path, so you may see the notation

$$\int_A^B \mathbf{F} \cdot d\mathbf{r}$$

or something equivalent. We shall discuss some of these concepts and special notations later in this course.

Geometric Reasoning

Example 20, revisited In applying formula (20) to Example 20, you may have noticed that, on the circle, the force

$$\mathbf{F} = -R \sin \theta \mathbf{i} + R \cos \theta \mathbf{j}$$

and the displacement

$$d\mathbf{r} = \mathbf{r}'(\theta)d\theta = (-R \sin \theta \mathbf{i} + R \cos \theta \mathbf{j})d\theta$$

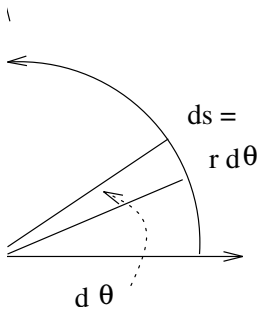
have the same direction. If we had been able to visualize this geometric relationship, we could have calculated the line integral more directly. Thus,

$$\mathbf{F} \cdot d\mathbf{r} = |\mathbf{F}||d\mathbf{r}| \cos 0 = |\mathbf{F}||d\mathbf{r}|.$$

On the other hand, $|\mathbf{F}| = \sqrt{(-y)^2 + x^2} = \sqrt{R^2} = R$, and $|d\mathbf{r}| = ds = R d\theta$. (The length of arc on a circle is always the radius of the circle times the subtended angle). Hence, we could have written directly

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_0^{2\pi} R R d\theta = 2\pi R^2.$$

You will often see this approach used by physicists or engineers. Note that it emphasizes the geometric or physical significance of the quantities, and it often makes clear underlying simplicity which might otherwise be hidden by complicated formulas. The approach used previously seems more straightforward, but that is because the the geometric (or physical) part of the problem had already been solved for you by giving you the parametric representation. For a real problem, you would have to do that yourself.



To approach line integral problems geometrically, it is useful to introduce one final bit of notation. We can think of the displacement $d\mathbf{r}$ as connecting two very close points on the curve, or to a high degree of approximation as being tangent to the curve. We have already introduced the notation ds for the magnitude of $d\mathbf{r}$, and we can specify its direction by giving the unit vector \mathbf{T} which is tangent to the curve at the point and which points in the preferred direction along the curve. Then $d\mathbf{r} = \mathbf{T} ds$, and we may write

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_C \mathbf{F} \cdot \mathbf{T} ds.$$

It is important to note that this last formula is only useful if you argue geometrically.

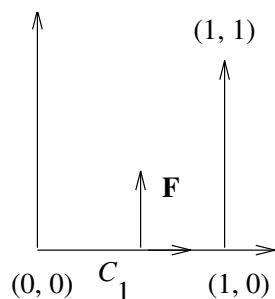
Example 21 Suppose $\mathbf{F} = 2x\mathbf{j}$ and let C be the path which starts at $(0, 0)$, moves to $(1, 0)$ and then moves to $(1, 1)$. (See the diagram.) In this case, the path consists of two segments C_1 and C_2 joined at a corner where the direction changes discontinuously. Clearly, the proper definition of the line integral in this situation should be

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_{C_1} \mathbf{F} \cdot d\mathbf{r} + \int_{C_2} \mathbf{F} \cdot d\mathbf{r}.$$

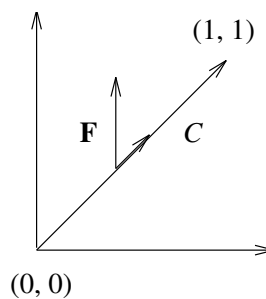
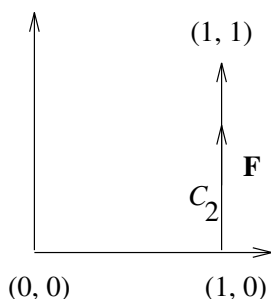
For the first segment C_1 , $\mathbf{F} = 2x\mathbf{j}$ is perpendicular to \mathbf{T} (hence, to $d\mathbf{r}$), so $\mathbf{F} \cdot d\mathbf{r} = 0$. Thus the first integral vanishes. For the second segment, C_2 , $\mathbf{F} = 2x\mathbf{j} = 2\mathbf{j}$, and $d\mathbf{r} = \mathbf{T} ds = \mathbf{j} dy$. \mathbf{F} and \mathbf{T} point the same way, so $\mathbf{F} \cdot d\mathbf{r} = 2ds = 2dy$. Hence,

$$\int_{C_2} \mathbf{F} \cdot d\mathbf{r} = \int_0^1 2 dy = 2y|_0^1 = 2.$$

To find the total answer, add up the answers for the two segments to get $0 + 2 = 2$.



Example 5



Example 6

Example 22 Let $\mathbf{F} = 2x\mathbf{j}$ as in the previous example, but let C be the straight line segment from $(0, 0)$ to $(1, 1)$. Then, $\mathbf{F} = 2x\mathbf{j}$ makes angle $\pi/4$ with \mathbf{T} (hence, with $d\mathbf{r}$). Thus,

$$\mathbf{F} \cdot d\mathbf{r} = |\mathbf{F}| ds \cos\left(\frac{\pi}{4}\right) = 2x ds \frac{1}{\sqrt{2}}.$$

Choosing x as the parameter, we may write $ds = \sqrt{2} dx$, so

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_0^1 \frac{2x\sqrt{2}dx}{\sqrt{2}} = \int_0^1 2x dx = x^2 \Big|_0^1 = 1.$$

Here is another slightly different way to do the calculation. We have $\mathbf{F} = 2x\mathbf{j}$, $d\mathbf{r} = dx\mathbf{i} + dy\mathbf{j}$. Hence, $\mathbf{F} \cdot d\mathbf{r} = 2x dy$. It would seem appropriate to choose y as parameter, so since $x = y$ on the given line, we have

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_0^1 2y dy = 1.$$

Finally, the calculation could be done using formula (20) as follows. The given line segment can be described by the parametric representation $\mathbf{r} = t\mathbf{i} + t\mathbf{j}$, $0 \leq t \leq 1$. Then $\mathbf{r}'(t) = \mathbf{i} + \mathbf{j}$, and on the line $\mathbf{F} = 2x\mathbf{j} = 2t\mathbf{j}$. Hence, by formula (20),

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_0^1 (2t\mathbf{j}) \cdot (\mathbf{i} + \mathbf{j}) dt = \int_0^1 2t dt = 1.$$

Polar Coordinates In polar coordinates, the velocity vector is given by

$$\frac{d\mathbf{r}}{dt} = \frac{dr}{dt}\mathbf{u}_r + r\frac{d\theta}{dt}\mathbf{u}_\theta$$

so we may write symbolically

$$d\mathbf{r} = \frac{d\mathbf{r}}{dt} dt = dr \mathbf{u}_r + r d\theta \mathbf{u}_\theta.$$

Hence, if $\mathbf{F} = F_r\mathbf{u}_r + F_\theta\mathbf{u}_\theta$ is resolved into polar components, we may write

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_C F_r dr + F_\theta r d\theta.$$

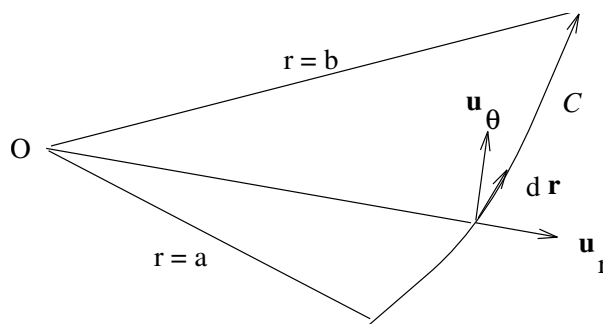
Example Let $\mathbf{F} = -\frac{1}{r^2}\mathbf{u}_r$. Such a force has magnitude inversely proportional to the square of the distance to the origin and is directed toward the origin. (Does it remind you of anything?) Let C be any path whatsoever, which starts at a point with $r = a$ and ends up at a point with $r = b$. We have

$$\mathbf{F} \cdot d\mathbf{r} = \left(-\frac{1}{r^2}\mathbf{u}_r\right) \cdot (dr \mathbf{u}_r + r d\theta \mathbf{u}_\theta) = -\frac{1}{r^2} dr.$$

(In other words, only the radial component of the displacement counts since the force is entirely radial in direction.) If we choose r as the parameter, we have

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_a^b -\frac{1}{r^2} dr = \left. \frac{1}{r} \right|_a^b = \frac{1}{b} - \frac{1}{a}.$$

In particular, the answer does not depend on the path, only on the values of r at the start and finish.



Note that the argument (and diagram) presumes that the particle moves in such a way that the radius always increases. Otherwise, it would not make sense to use r as the parameter, since there should be exactly one point on the curve for each value of the parameter. Can you see how to make the argument work if the path is allowed to meander in such a way that r sometimes increases and sometimes decreases?

Exercises for 1.6.

- Find the arc length of each of the following paths for the given parameter interval.
 - $\mathbf{r} = 2 \sin t \mathbf{i} - 2 \cos t \mathbf{j} + 2t \mathbf{k}$, from $t = 0$ to $t = 2\pi$.
 - $\mathbf{r} = \langle e^{-t} \cos t, e^{-t} \sin t, e^{-t} \rangle$, from $t = 0$ to $t = 2$.
 - $x = t^2$, $y = 4t + 3$, $z = -4 \ln t$, from $t = 1$ to $t = 2$.
 - $x = 4t$, $y = 3$, $z = 2t^2$, from $t = 0$ to $t = 3$.
- Let \mathcal{C} be the graph of the function defined by $y = f(x)$ on the interval $a \leq x \leq b$. Use the parametric representation $\mathbf{r} = x \mathbf{i} + f(x) \mathbf{j}$ to derive the formula

$$L(\mathcal{C}) = \int_a^b \sqrt{1 + f'(x)^2} dx.$$

- (Optional, for those with an interest in mathematical rigor). Suppose a curve \mathcal{C} is represented parametrically by $\mathbf{r} = \mathbf{r}_1(u)$, $a \leq u \leq b$ and $\mathbf{r} = \mathbf{r}_2(v)$, $c \leq v \leq d$. Under reasonable assumptions on the parametric representations, it may be shown that there is a functional relation $u = p(v)$, $c \leq v \leq d$ such that p is a differentiable function with continuous derivative, $p'(u) \geq 0$, $p(c) = a$, $p(d) = b$, and $\mathbf{r}_1(p(v)) = \mathbf{r}_2(v)$ for $c \leq v \leq d$. Show that

$$\int_a^b |\mathbf{r}'_1(u)| du = \int_c^d |\mathbf{r}'_2(v)| dv.$$

Hint: Use the chain rule

$$\frac{d\mathbf{r}_1(p(v))}{dv} = \frac{d\mathbf{r}_1(u)}{du} \frac{dp(v)}{dv} \quad \text{where } u = p(v),$$

and apply the change of variables formula for integrals.

4. Evaluate $\int_C \mathbf{F} \cdot \mathbf{T} ds$ in each case.
 - (a) $\mathbf{F} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$, $\mathbf{r} = \langle t, t, 2 \rangle$, $0 \leq t \leq 1$.
 - (b) $\mathbf{F} = yz\mathbf{i} + xz\mathbf{j} + xy\mathbf{k}$, $x = \sin t$, $y = 2 \cos t$, $z = t$, $0 \leq t \leq \pi$.
 - (c) $\mathbf{F} = y\mathbf{i} + z\mathbf{j} + x\mathbf{k}$, $\mathbf{r} = t\mathbf{i} + t^2\mathbf{j} + t^3\mathbf{k}$, $0 \leq t \leq 1$.
5. Let $\mathbf{F} = -y\mathbf{i}$, and let \mathcal{C} be the rectilinear path starting at $(0, 1)$, going to $(0, 3)$ and ending at $(3, 3)$. Find $\int_C \mathbf{F} \cdot d\mathbf{r}$.
6. Let $\mathbf{F} = \mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$. Compute $\int_C \mathbf{F} \cdot d\mathbf{r}$ for each of the following paths:
 - (a) the linear path which goes directly from the origin to the point $(2, 1, 3)$,
 - (b) the rectilinear path which goes from the origin to $(2, 0, 0)$, then to $(2, 1, 0)$ and finally to $(2, 1, 3)$.
7. Let $\mathbf{F} = y\mathbf{i}$ and let \mathcal{C} be the path which starts at $(-1, 1)$ and follows the parabola $y = x^2$ to $(1, 1)$. Find $\int_C \mathbf{F} \cdot d\mathbf{r}$.
8. A workman at the top of a ten story building (50 m high) needs to lower a 10 kg box to a point on the ground 50 m from the base of the building. To do this, he constructs a variety of slides. Show that the work done by gravity (with force per unit mass $9.8\mathbf{j}$) is identical if the slide is
 - (a) A freefall chute, and the box is then pushed 50 m horizontally.
 - (b) A straight line of slope 1.
 - (c) One quarter of a circle, with its center at the base of the building (concave down).
 - (d) One quarter of a circle, with its center 50 m above the final point (concave up).

Interesting diversion: Are the times involved the same? Hint: Consider the ISP toy with two balls on two slightly different tracks.
9. Suppose $\mathbf{F} = r\mathbf{u}_\theta$. (a) Calculate $\int_C \mathbf{F} \cdot d\mathbf{r}$ for \mathcal{C} a circle of radius R centered at the origin and traversed once counter-clockwise. (This was done differently as an Example in the text. Do you recognize it?) (b) Calculate the line integral for the same \mathbf{F} , but for the path given parametrically by $\mathbf{r} = t\mathbf{i} + t\mathbf{j}$, $0 \leq t < \infty$. In principle, the fact that the path is unbounded could cause a problem. Why doesn't it in this case?

10. Let $\mathbf{F} = \mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$, and let \mathcal{C} be the circle of radius R lying in the plane $z = h$ and centered at the point $(0, 0, h)$. Assume \mathcal{C} is traversed in the counter-clockwise direction when viewed from above. Calculate $\int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r}$.
11. Let $\mathbf{F} = \mathbf{j}$ and let \mathcal{C} be the quarter of the circle $x^2 + y^2 = R^2$ in the first quadrant of the x, y -plane. Assume \mathcal{C} is traversed counter-clockwise. Find $\int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r}$. (Hint: The angle between \mathbf{F} and $d\mathbf{r}$ is the polar angle θ .) What if we use $\mathbf{F} = \mathbf{i}$ instead? How about $x\mathbf{i}$?

Chapter 2

Differential Equations, Preview

2.1 Some Elementary Differential Equations

Later in this course, you will study differential equations in a systematic way, but before that you are likely to encounter some simple differential equations in Physics and other courses. Hence, we shall give a brief introduction to the subject here.

A differential equation is an equation involving an independent variable, a dependent variable, and derivatives of the latter. The *order* of a differential equation is the highest order of any derivative which appears in it. For example,

$$\frac{dx}{dt} = ax \tag{22}$$

is an example of a *first order* differential equation, while

$$\frac{d^2x}{dt^2} = -Kx \tag{23}$$

is an example of a *second order* differential equation.

To solve a differential equation, you must find a *function* $x = x(t)$, expressing the dependent variable in terms of the independent variable, which when substituted in the equation yields a true identity. For example, substituting $x = Ce^{at}$ (where C is any constant) and $dx/dt = Ca e^{at}$ in equation (22) yields

$$Ca e^{at} = a(Ce^{at})$$

which is true. Hence, the function given by $x(t) = Ce^{at}$ is a solution. Note the following very important point. If someone is kind enough to suggest a solution to

you, it is not necessary to go through any procedure to “find” that solution. All you need to do is to check that it works.

If you don’t have any idea of what a solution to a given differential equation might be, then you need some method to try to find a solution. How one goes about this is a vast subject, and we shall go into it later, as mentioned above. For the moment, we describe only the method of *separation of variables* which works in many simple examples. We illustrate it by solving equation (22).

$$\begin{aligned}
 \frac{dx}{dt} &= ax \\
 \frac{dx}{x} &= a dt && \text{Variables are separated formally} \\
 \int \frac{dx}{x} &= \int a dt && \text{Take antiderivatives} \\
 \ln |x| &= at + c \\
 |x| &= e^{at+c} = e^{at} e^c && \text{Exponentiate} \\
 x &= \pm e^c e^{at}.
 \end{aligned}$$

However, $\pm e^c$ is just some constant which we may call C . Thus we find that $x = Ce^{at}$ is a solution.

Note that when integrating there should be an undetermined constant on each side of the equation. However, these constants can be combined in one by transposition, and that is what we did by putting the $+c$ only on the right side of the equation in the fourth line.

Whenever the variables can be so separated, the above procedure produces a *general solution* of the equation. There are some technical difficulties with the method. (For example, a denominator might vanish after the variables have been separated. Also, the formal separation procedure needs some rigorous justification.) However, these difficulties can usually be resolved and one can make a convincing argument that *any solution* will be of the form obtained by the method.

Note that in the example, the solution produced one *arbitrary* constant C , i.e., one constant which is not otherwise determined by the equation. (a is also a constant, but it was assumed known in the original equation.) You can think of that constant arising from integration. To complete the solution, it is necessary to determine the constant. There are many ways to do this, depending on the formulation of the problem giving rise to the differential equation. One of the easiest is to give the value x_0 of the dependent variable for one specified value t_0 of the independent variable and then solve the resulting equation for the constant C . (This is called satisfying an *initial condition*.) For example, suppose we are given that $x = 10$ when $t = 0$. Substituting these values yields

$$10 = Ce^{a(0)} = Ce^0 = C$$

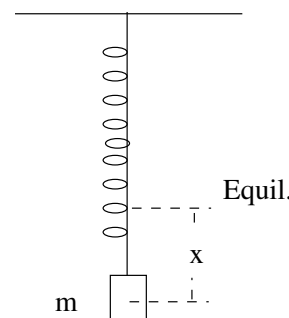
or $C = 10$. With that initial condition, the solution is $x = 10e^{at}$.

Equation (22) describes processes of growth or decay. x might represent the size of a population, the number of radioactive atoms, or something similar. The equation asserts that the growth rate (decay rate if $a < 0$) is proportional to the amount present. In the case of population growth, a is often called the “birth rate”. In the case of radioactive decay, $a < 0$, so we put $a = -\gamma$, and γ tells us the *proportion* of atoms present at a given time which will decay at that time. In that case, γ is often expressed in terms of the *half life* T which is the time for the x to decrease to half its initial value. You might check your proficiency in the use of exponentials and logarithms by checking the following formula.

$$\gamma = \frac{\ln 2}{T}.$$

Equation (23), $\frac{d^2x}{dt^2} = -Kx$, arises in studying *simple harmonic motion*. For example, the motion of a particle of mass m oscillating at the end of a spring with spring constant k is governed by such an equation with $K = k/m$, if we ignore friction. (23) is much harder to solve than (22). One must distinguish the cases $K > 0$ and $K < 0$, but the former has more interesting applications (as in simple harmonic motion), so we shall assume $K > 0$. The solution proceeds by separation of variables as follows. First, put $v = dx/dt$. Then the equation becomes

$$\begin{aligned} \frac{dv}{dt} &= -Kx \\ \frac{dv}{dx} \frac{dx}{dt} &= -Kx && \text{using the chain rule} \\ \text{or } \frac{dv}{dx} v &= -Kx \\ v \, dv &= -Kx \, dx && \text{separating the variables} \\ \frac{v^2}{2} &= -K \frac{x^2}{2} + c \\ v^2 &= -Kx^2 + 2c. \end{aligned}$$



However, we might just as well rename the constant $2c$ and call it C_1 . Note that since $K > 0$, we have $C_1 = v^2 + Kx^2 \geq 0$. If $C_1 = 0$, it follows that $v = x = 0$ for all t . That is certainly possible, but it isn't very interesting, so we assume $C_1 > 0$. Now, recalling that $v = dx/dt$, we obtain

$$\begin{aligned} v = \frac{dx}{dt} &= \pm \sqrt{C_1 - Kx^2} \\ \frac{dx}{\sqrt{C_1 - Kx^2}} &= \pm dt && \text{separating variables} \\ \int \frac{dx}{\sqrt{C_1 - Kx^2}} &= \pm t + C_2 && \text{integrating} \\ \frac{1}{\sqrt{K}} \cos^{-1} \left(\sqrt{\frac{K}{C_1}} x \right) &= \pm t + C_2. \end{aligned}$$

Note that we need $K > 0$ and $C_1 > 0$ for the last integration to be valid. Continuing, we have

$$\begin{aligned}\cos^{-1}\left(\sqrt{\frac{K}{C_1}}x\right) &= \pm\sqrt{K}t + \sqrt{K}C_2 \quad \text{or} \\ x &= \sqrt{\frac{C_1}{K}} \cos(\pm\sqrt{K}t + \sqrt{K}C_2) \\ &= \sqrt{\frac{C_1}{K}} \cos(\sqrt{K}t \pm \sqrt{K}C_2).\end{aligned}$$

However, we may define $A = \sqrt{\frac{C_1}{K}}$ (so $A > 0$) and $\delta = \pm\sqrt{K}C_2$ to obtain the general solution

$$x = A \cos(\sqrt{K}t + \delta). \quad (24)$$

Note that this solution has *two* arbitrary constants arising from the two integrations which had to be performed. Generally, the solution of an n th order equation involves n arbitrary constants.

The above derivation depended strongly on the assumption $K > 0$. For $K < 0$, we would not be able to conclude that $C_1 > 0$, and the integration step which resulted in $\cos^{-1}(\sqrt{\frac{K}{C_1}}x)$ on the left would not be valid. If you are ambitious, you might try doing the integration under the assumption $K < 0$ to see what you get. Later in this course, we shall derive other methods to solve this equation whatever the sign of K .

To determine the constants, one may again specify initial conditions. However, specifying the value x_0 at t_0 will yield only one equation for the two constants. Hence, one needs an additional condition to obtain another equation. To get that, one commonly specifies the derivative v_0 at t_0 . For example, suppose $K = 3$, and suppose $x = 1, dx/dt = -1$ at $t = 0$. Then (24) yields the two equations

$$\begin{aligned}1 &= A \cos(\delta) \\ -1 &= -A\sqrt{3} \sin(\delta).\end{aligned}$$

Dividing the second equation by the first, yields

$$\tan \delta = \frac{1}{\sqrt{3}}$$

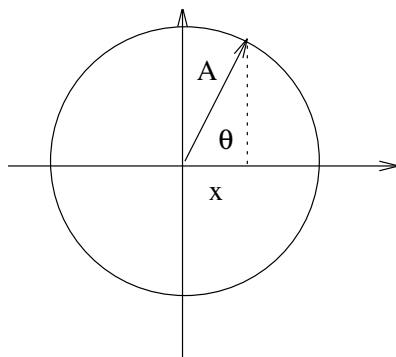
from which we conclude $\delta = \pi/6$ or $-5\pi/6$. Since both $\sin \delta$ and $\cos \delta$ are positive, this yields $\delta = \pi/6$. Putting this in the first equation yields

$$1 = A \cos\left(\frac{\pi}{6}\right) = A \frac{\sqrt{3}}{2}$$

so $A = 2/\sqrt{3}$. We conclude that the solution satisfying these initial conditions is

$$x = \frac{2}{\sqrt{3}} \cos\left(\sqrt{3}t + \frac{\pi}{6}\right).$$

The quantity A in equation (24) is called the *amplitude* of the solution, and δ is called the *phase*. One way to visualize the solution is as follows. Consider a point moving on a circle of radius A centered at the origin.



$$\theta = \omega t + \delta$$

The projection of that point on the x -axis oscillates back according to (24). δ gives the angle the position vector of the point makes with the x -axis at $t = 0$. The quantity $\omega = \sqrt{K}$ gives the angular velocity of the point on the circle. ω is also called the *angular frequency* of the oscillation. The *period* $T = 2\pi/\omega$ is the time required for one circuit (one complete oscillation). The frequency $f = 1/T = \omega/2\pi$ is the number of circuits (oscillations) per unit time. The frequency is usually measured in Hertz (Hz), i.e., oscillations per second.

One often sees the solution written differently in the form obtained by expanding (24) as follows.

$$\begin{aligned} x &= A \cos(\omega t + \delta) \\ &= A \cos(\omega t) \cos \delta - A \sin(\omega t) \sin \delta \\ x &= C \cos(\omega t) + D \sin(\omega t) \end{aligned}$$

where $C = A \cos \delta$ and $D = -A \sin \delta$.

You should note that had you suspected that $x = A \cos(\omega t + \delta)$ is a solution, you could have checked it by substituting into (23), and avoided most of the work in the derivation above. You might argue that one would have little reason to suspect such a solution might work. That is true if one is thinking in purely mathematical terms, but given that one knows the differential equation arises from a physical problem in which one observes oscillatory behavior, it is not entirely unreasonable to guess at a solution like (24). Of course, without going through a more refined analysis, you could not be absolutely sure that (24) encompasses all possible solutions (i.e., that it is sufficiently general), but the fact that it involves *two* arbitrary constants would strongly suggest that such is the case.

Remember the moral of the above discussion. If somehow or other a solution to a problem is suggested to you, you don't have to bother "deriving" that solution by a "mathematical" procedure. If the solution works, and if in addition, you have reason to believe there is only one solution which can work, then the one you have must be it.

Exercises for 2.1.

- Find general solutions for the following differential equations.

(a) $\frac{dy}{dx} = 4x^2y^2$.

(b) $\frac{dy}{dx} = \frac{x}{y}$.

(c) $\frac{dy}{dx} = \frac{1+x^2}{1+y^2}$. (Don't try to express y as a function of x .)

(d) $\frac{dy}{dx} = (xy)^{3/2}$.

- Solve the initial value problem

$$\frac{dy}{dx} = y^2, \text{ where } y(0) = 2$$

- The arctic fox population in a certain habitat is given by the equation

$$\frac{dP}{dt} = -k\sqrt{P}$$

- If the initial population (at $t = 0$) is P_0 , find the general solution.
 - If there are 100 foxes initially, and 25 remain after 6 weeks, how long until extinction? Do you see anything wrong with this approach?
- The bacterial species *E. Coli* doubles in number every twenty minutes. If a single colony is present initially, how long will a biologist have to wait before having one million colonies? One billion colonies?
 - Show that the time T required for $e^{-\gamma t}$ to drop to half its value at $t = 0$ is

$$T = \frac{\ln 2}{\gamma}.$$

- Archaeologists recently uncovered a relic purported to date from 0 A.D. Several independent laboratories used carbon dating to analyze the sample. If the fabric contains 6.0×10^{11} atoms of ^{14}C per gram, and modern fabric of the same type contains 7.0×10^{11} atoms of ^{14}C per gram, is it authentic?

7. Electricity drains from a capacitor according to the equation

$$\frac{dV}{dt} = -\frac{1}{20}V.$$

Solve the equation in terms of $V(0) = V_0$. Find how long it will take for V to be reduced to one hundredth its value at $t = 0$. Is V ever zero?

8. *Newton's Law of Cooling* states that the rate at which an object's temperature changes is proportional to the temperature difference between the object and its surroundings. In other words,

$$\frac{dT}{dt} = -k(T - T_s),$$

where T is the temperature of the object, and T_s is the temperature of its surroundings (a constant).

- (a) Solve this equation for T .
- (b) A steel ingot at 1000 K is exposed to air at 300 K. If it cools to 800 K in one hour, what will its temperature be after five hours? How long will it take until equilibrium ($T = T_s$) is attained?
9. In the solution of the equation $d^2x/dt^2 = -Kx$, it was noted that the constant $C_1 = 2c > 0$ if $K > 0$. However, the reasoning did not exclude the case $C_1 = 2c = 0$. How might we exclude that case? Hint: What would $C_1 = 0$ say about x and v if $K > 0$?
10. A mass on a spring undergoes simple harmonic motion.
- (a) If the frequency is 6 Hz, $x(0) = 5$, and $x'(0) = 0$, find the amplitude A of the motion.
- (b) If the period is 4 seconds, $x(0) = 0$, and $x'(0) = 5$, find the amplitude A of the motion. If the mass is 10 kg, what is the force applied by the spring at $t = 0$?
- (c) If the amplitude is 10, the mass starts at the origin, and $x'(0) = 1$, find the frequency and period of motion.
- (d) Find the phases for (a), (b), and (c).
11. A pendulum at the end of a (massless) rod of length L satisfies the differential equation $\frac{d^2\theta}{dt^2} = -\frac{g}{L} \sin \theta$. Show that

$$\left(\frac{d\theta}{dt}\right)^2 - \frac{2g}{L} \cos \theta = C$$

where C is a constant. Solve for $\frac{d\theta}{dt}$ and try to solve the resulting first order differential equation explicitly. If you can't integrate the resulting expression, don't be surprised.

Chapter 3

Differential Calculus of Functions of n Variables

We want to develop the calculus necessary to discuss functions of many variables. We shall start with functions $f(x, y)$ of two independent variables and functions $f(x, y, z)$ of three independent variables. However, in general, we need to consider functions $f(x_1, x_2, \dots, x_n)$ of any number of independent variables. We shall use the notation \mathbf{R}^n as before to stand for the set of all n -tuples (x_1, x_2, \dots, x_n) with real entries x_i . For $n = 2$, we shall identify \mathbf{R}^2 with the plane and, for $n = 3$, we shall identify \mathbf{R}^3 with space. We shall use the old fashioned term *locus* to denote the set of all points satisfying some equation or condition.

3.1 Graphing in \mathbf{R}^n

We shall encounter equations involving two, three, or more variables. As you know, an equation of the form

$$f(x, y) = C$$

may be viewed as defining a curve in the plane. For example, $ax + by = c$ has plane locus a line, while $x^2 + y^2 = R^2$ has plane locus a circle of radius R centered at the origin. Similarly, an equation involving three variables

$$f(x, y, z) = C$$

may be thought of as defining a *surface* in space. Thus, we saw previously that the locus in \mathbf{R}^3 of a linear equation

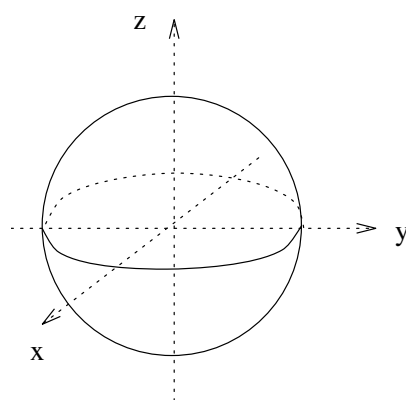
$$ax + by + cz = d$$

(where not all a, b , and c are zero) is a plane. If we use more complicated equations, we get more complicated surfaces.

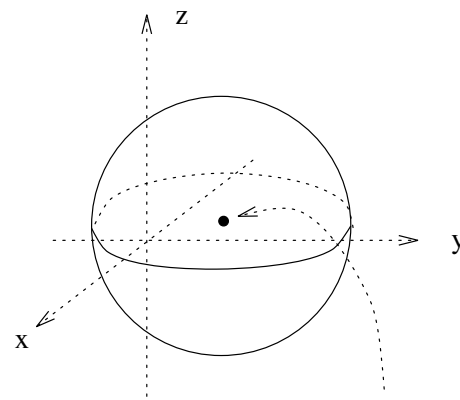
Example 23 The equation

$$x^2 + y^2 + z^2 = R^2$$

may be rewritten $|\mathbf{r}| = \sqrt{x^2 + y^2 + z^2} = R$, so it asserts that the point with position vector \mathbf{r} is at distance R from the origin. Hence, the locus of all such points is a *sphere* of radius R centered at the origin.



Sphere centered at $(0, 0, 0)$



Sphere centered at $(-1, 2, 0)$

Example 24 Consider the locus of the equation

$$x^2 + 2x + y^2 - 4y + z^2 = 20.$$

This is also a sphere, but one not centered at the origin. To see this, *complete the squares* for the terms involving x and y .

$$\begin{aligned} x^2 + 2x + 1 + y^2 - 4y + 4 + z^2 &= 10 + 1 + 4 = 25 \\ (x + 1)^2 + (y - 2)^2 + z^2 &= 5^2. \end{aligned}$$

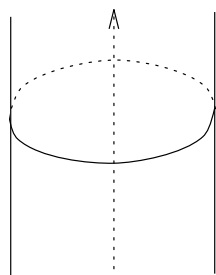
This asserts that the point with position vector $\mathbf{r} = \langle x, y, z \rangle$ is 5 units from the point $(-1, 2, 0)$, i.e., it lies on a sphere of radius 5 centered at $(-1, 2, 0)$.

Example 25 Consider the locus of the equation $z = x^2 + y^2$ (which could also be written $x^2 + y^2 - z = 0$.) To see what this looks like, we consider its intersection with various planes. Its intersection with the y, z -plane is obtained by setting $x = 0$ to get $z = y^2$. This is a parabola in the y, z -plane. Similarly, its intersection with the x, z -plane is the parabola given by $z = x^2$. To fill in the picture, consider intersections with planes parallel to the x, y -plane. Any such plane has equation $z = h$, so the intersection has equation $x^2 + y^2 = h = (\sqrt{h})^2$, which you should

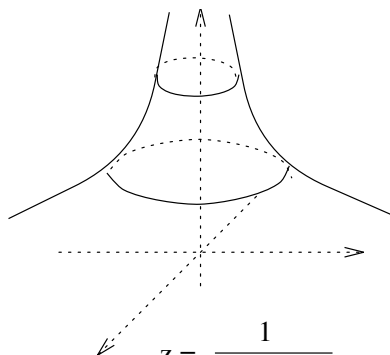
recognize as a circle of radius \sqrt{h} , at least if $h > 0$. Note that the circle is centered at $(0, 0, h)$ on the z -axis since it lies in the plane $z = h$. If $z = h = 0$, the circle reduces to a single point, and for $z = h < 0$, there is no locus. The surface is “bowl” shaped. It is called a *circular paraboloid*.

Graphing a surface in \mathbf{R}^3 by sketching its traces on various planes is a useful strategy. In order to be good at it, you need to know the basics of plane analytic geometry so you can recognize the resulting curves. In particular, you should be familiar with the elementary facts concerning *conic sections*, i.e., ellipses, hyperbolas, and parabolas. Edwards and Penney, 3rd Edition, Chapter 10 is a good reference for this material.

Example 26 Consider the locus in space of $\frac{x^2}{4} + \frac{y^2}{9} = 1$. Its intersection with a plane $z = h$ parallel to the x, y -plane is an ellipse centered on the z -axis and with semi-minor and semi-major axes 2 and 3. The surface is a *cylinder* perpendicular to the x, y -plane with elliptical cross sections. Note that the locus *in space* is not just the ellipse in the x, y -plane with the same equation.



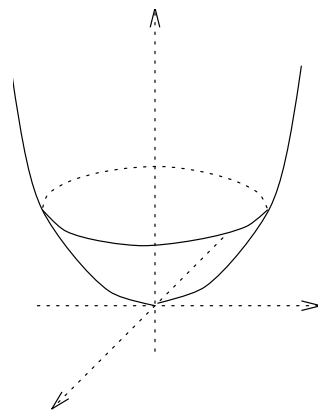
$$\frac{x^2}{4} + \frac{y^2}{9} = 1$$



$$z = \frac{1}{x^2 + y^2}$$

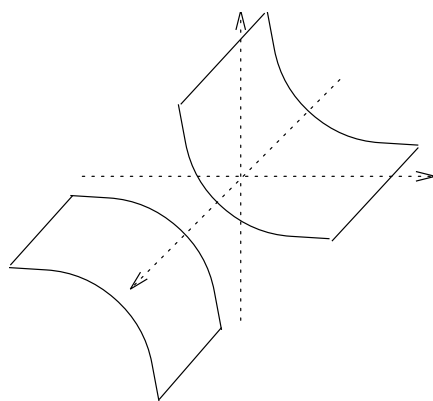
Example 27 Consider the locus in space of the equation $z = \frac{1}{x^2 + y^2}$. Its intersection with the plane $z = h$ (for $h > 0$) is the circle with equation $x^2 + y^2 = 1/h = (\sqrt{1/h})^2$. The surface does not intersect the x, y -plane itself ($z = 0$) nor any plane below the x, y -plane. Its intersection with the x, z -plane ($y = 0$) is the curve $z = 1/x^2$ which is asymptotic to the x -axis and to the positive z -axis. Similarly, for its intersection with the y, z -plane. The surface flattens out and approaches the x, y -plane as $r = \sqrt{x^2 + y^2} \rightarrow \infty$. It approaches the positive z -axis as $r \rightarrow 0$.

Example 28 Consider the locus in space of the equation $yz = 1$. Its intersection with a plane parallel to the y, z -plane ($x = d$) is a hyperbola asymptotic to the y and z axes. The surface is perpendicular to the y, z -plane. Such a surface is

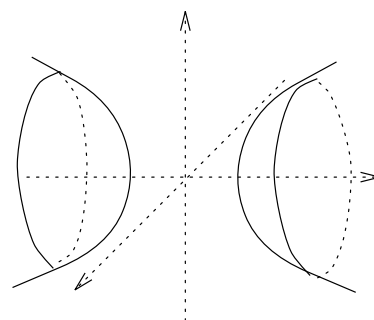


also called a *cylinder* although it doesn't close upon itself as the elliptical cylinder considered above.

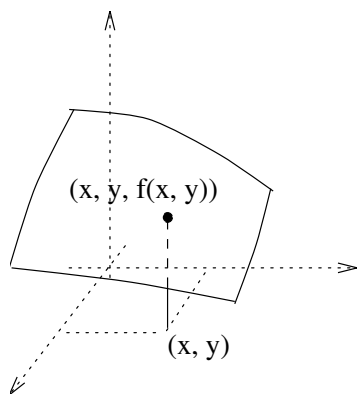
Example 29 Consider the locus of the equation $x^2 + z^2 = y^2 - 1$. For each plane parallel to the x, z -plane ($y = c$), the intersection is a circle $x^2 + z^2 = c^2 - 1 = (\sqrt{c^2 - 1})^2$ centered on the y -axis, at least of $c^2 > 1$. For $y = c = \pm 1$, the locus is a point, and for $-1 < y = c < 1$, the locus is empty. In addition, the intersection of the surface with the x, y -plane ($z = 0$) is the hyperbola with equation $x^2 - y^2 = -1$, and similarly for its intersection with the y, z -plane. The surface comes in two pieces which open up as “bowls” centered on the positive and negative y -axes. The surface is called a *hyperboloid of 2 sheets*.



$$yz = 1$$



$$x^2 + z^2 = y^2 - 1$$

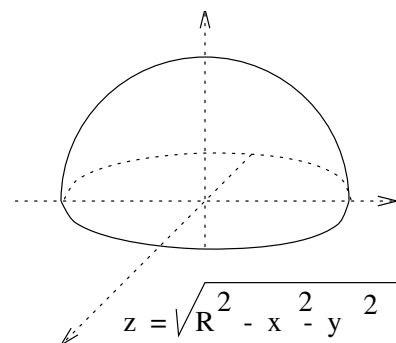


Graphs of Functions For a scalar function f of one independent variable, the *graph of the function* is the set of all points in \mathbf{R}^2 of the form $(x, f(x))$ for x in the domain of the function. (The domain of a function is the set of values of the independent variable for which the function is defined.) In other words, it is the locus of the equation $y = f(x)$. It is generally a curve in the plane.

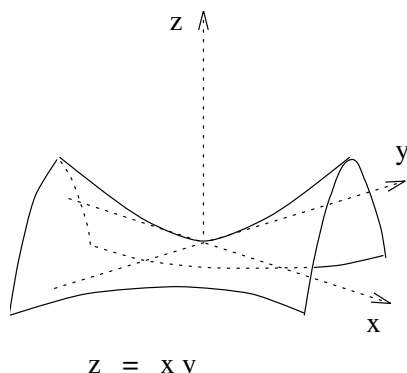
We can define a similar notion for a scalar function f of two independent variables. The graph is the set of points in \mathbf{R}^3 of the form $(x, y, f(x, y))$ for (x, y) a point in the domain of the function. In other words, it is the locus of the equation $z = f(x, y)$, and it is generally a surface in space. The graph of a function is often useful in understanding the function.

We have already encountered several examples of graphs of functions. For example, the locus of $z = x^2 + y^2$ is the graph of the function f defined by $f(x, y) = x^2 + y^2$. Similarly, the locus of $z = 1/(x^2 + y^2)$ is the graph of the function f defined by $f(x, y) = 1/(x^2 + y^2)$ for $(x, y) \neq (0, 0)$. Note that in the first case there need be no restriction on the domain of the function, but in the second case $(0, 0)$ was omitted.

In some of the other examples, the locus of the equation cannot be considered the graph of a function. For example, the equation $x^2 + y^2 + z^2 = R^2$ cannot be solved uniquely for z in terms of (x, y) . Indeed, we have $z = \pm\sqrt{R^2 - x^2 - y^2}$, so that two possible functions suggest themselves. $z = f_1(x, y) = \sqrt{R^2 - x^2 - y^2}$ defines a function with graph the *top hemisphere* of the sphere, while $z = f_2(x, y) = -\sqrt{R^2 - x^2 - y^2}$ yields the lower hemisphere. (Note that for either of the functions the relevant domain is the set of points on or inside the circle $x^2 + y^2 = R^2$. For points outside that circle, the expression inside the square root is negative, so, since we are only talking about functions assuming real values, such points must be excluded.)



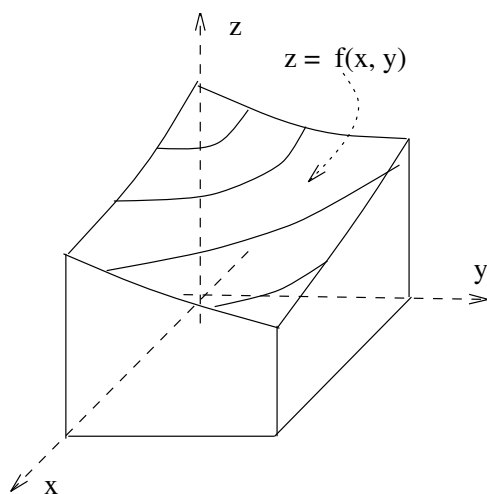
Example 30 Let $f(x, y) = xy$ for all (x, y) in \mathbf{R}^2 . The graph is the locus of the equation $z = xy$. We can sketch it by considering traces on various planes. Its intersection with a plane parallel to the x, y -plane ($z = \text{constant}$) is a hyperbola asymptotic to lines parallel to the x and y axes. For $z > 0$, the hyperbola is in the first and third quadrants of the plane, but for $z < 0$ it is in the second and fourth quadrants. For $z = 0$, the equation is $xy = 0$ with locus consisting of the x -axis ($y = 0$) and the y -axis ($x = 0$). Thus, the graph intersects the x, y -plane in two straight lines. The surface is generally shaped like an “infinite saddle”. It is called a *hyperbolic paraboloid*. It is clear where the term “hyperbolic” comes from. Can you see any parabolas? (Hint: Try planes perpendicular to the x, y -plane with equations of the form $y = mx$.)



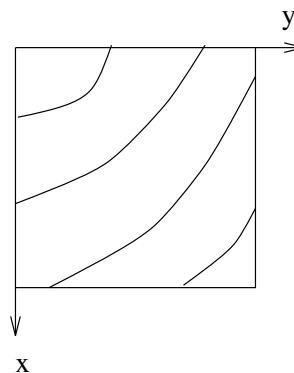
Example 31 Let $f(x, y) = x/y$ for $y \neq 0$. Thus, the domain of this function consists of all points (x, y) not on the x -axis ($y = 0$). The trace in the plane $y = c, c \neq 0$ is the line $z = (1/c)x$ with slope $1/c$. Similarly, the trace in the plane $z = c, c \neq 0$ is the line $y = (1/c)x$. Finally, the trace in the plane $x = c$, is the hyperbola $z = c/y$. Even with this information you will have some trouble visualizing the graph. However, the equation $z = x/y$ can be rewritten $yz = x$. By permuting the variables, you should see that the locus of $yz = x$ is similar to the saddle shaped surface we just described, but oriented differently in space. However, the saddle is not quite the graph of the function since it contains the z -axis ($y = x = 0$) but the

graph of the function does not. In general, the graph of a function, since it consists of points of the form $(x, y, f(x, y))$, cannot contain points with the same values for x and y but different values for z . In other words, any line parallel to the z -axis can intersect such a graph at most once.

Sketching graphs of functions, or more generally loci of equations in x, y , and z , is not easy. One approach drawn from the study of topography is to interpret the equation $z = f(x, y)$ as giving the *elevation* of the surface, viewed as a hilly terrain, above a reference plane. (Negative elevation $f(x, y)$ is interpreted to mean that the surface dips below the reference plane.) For each possible elevation c , the intersection of the plane $z = c$ with the graph yields a curve $f(x, y) = c$. This curve is called a *level curve*, and we draw a 2-dimensional map of the graph by sketching the level curves and labeling each by the appropriate elevation c . Of course, there are generally infinitely many level curves since there are infinitely many possible values of z , but we select some subset to help us understand the topography of the surface.



Contour lines on surface



Projected level curves in plane

Example 32 The level curves of the surface $z = xy$ have equations $xy = c$ for various c . They form a family of hyperbolas, each with two branches. For $c > 0$, these hyperbolas fill the first and third quadrants, and for $c < 0$ they fill the second and fourth quadrants. For $c = 0$ the x and y axes together constitute the level “curve”. See the diagram.

You can see that the region around the origin $(0, 0)$ is like a “mountain pass” with the topography rising in the first and third quadrants and dropping off in the second and fourth quadrants. In general a point where the graph behaves this way is called

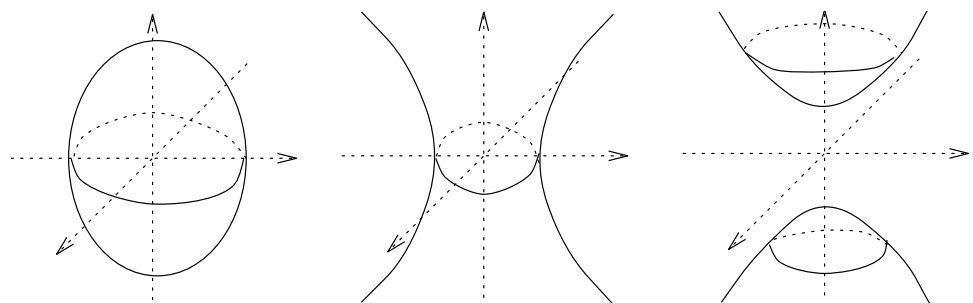
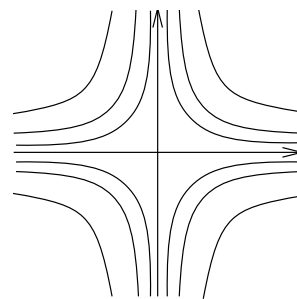
a *saddle point*. Saddle points indicate the added complexity which can arise when one goes from functions of one variable to functions of two or more variables. At such points, the function can be considered as having a maximum or a minimum depending on where you look.

Quadric Surfaces One important class of surfaces are those defined by quadratic equations. These are analogues in three dimensions of conics in two dimensions. They are called *quadric surfaces*. We describe here *some* of the possibilities. You can verify the pictures by using the methods described above.

Consider first equations of the form

$$\pm \frac{x^2}{a^2} \pm \frac{y^2}{b^2} \pm \frac{z^2}{c^2} = 1$$

If all the signs are positive, the surface is called an *ellipsoid*. Planes perpendicular to one of the coordinate axes intersect it in ellipses (if they intersect at all). However, at the extremes these ellipses degenerate into the points $(\pm a, 0, 0)$, $(0, \pm b, 0)$, and $(0, 0, \pm c)$.



Ellipsoid

Hyperboloid of one sheet

Hyperboloid of two sheets

If exactly one of the signs are negative, the surface is called a *hyperboloid of one sheet*. It is centered on one axis (the one associated to the negative coefficient in the equation) and it opens up in both positive and negative directions along that axis. Its intersection with planes perpendicular to that axis are ellipses. Its intersections with planes perpendicular to the other axes are hyperbolas.

If exactly two of the signs are negative, the surface is called a *hyperboloid of two sheets*. It is centered on one axis (associated to the positive coefficient). For example, suppose the equation is

$$-\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1.$$

For $y < -b$ or $y > b$, the graph intersects a plane perpendicular to the y -axis in an ellipse. For $y = \pm b$, the intersection is the point $(0, \pm b, 0)$. (These two points are called vertices of the surface.) For $-b < y < b$, there is no intersection with a plane perpendicular to the y -axis.

The above surfaces are called *central quadrics*. Note that for the hyperboloids, with equations in standard form as above, the number of sheets is the same as the number of minus signs.

Consider next equations of the form

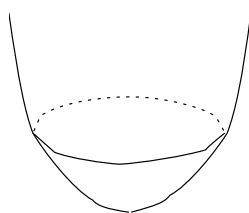
$$z = \pm \frac{x^2}{a^2} \pm \frac{y^2}{b^2}$$

(or similar equations obtained by permuting x, y and z .)

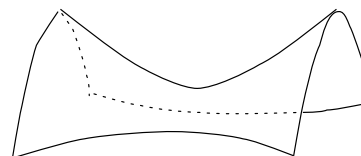
If both signs are the same, the surface is called an *elliptic paraboloid*. If both signs are positive, it is centered on the positive z -axis and its intersections with planes perpendicular to the positive z -axis are a family of similar ellipses which increase in size as z increases. If both signs are negative, the situation is similar, but the surface lies below the x, y plane.

If the signs are different, the surface is called a *hyperbolic paraboloid*. Its intersection with planes perpendicular to the z -axis are hyperbolas asymptotic to the lines in those planes parallel to the lines $x/a = \pm y/b$. Its intersection with the x, y -plane is just those two lines. The surface has a saddle point at the origin.

The locus of the equation $z = cxy, c \neq 0$ is also a hyperbolic paraboloid, but rotated so it intersects the x, y -plane in the x and y axes.



Elliptic paraboloid



Hyperbolic paraboloid

Finally, we should note that many so called “degenerate conics” are loci of quadratic equations. For example, consider

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 0$$

which may be solved to obtain

$$z = \pm c \sqrt{\frac{x^2}{a^2} + \frac{y^2}{b^2}}.$$

The locus is a double cone with elliptical cross sections and vertex at the origin.

Generalizations In general, we will want to study functions of any number of independent variables. For example, we may define the graph of a scalar valued function f of three independent variables to be the set of all points in \mathbf{R}^4 of the form $(x, y, z, f(x, y, z))$. Such an object should be considered a three dimensional subset of \mathbf{R}^4 , and it is certainly not easy to visualize. It is more useful to consider the analogues of level curves for such functions. Namely, for each possible value c attained by the function, we may consider the locus in \mathbf{R}^3 of the equation $f(x, y, z) = c$. This is generally a surface called a *level surface* for the function.

Examples For $f(x, y, z) = x^2 + y^2 + z^2$, the level surfaces are concentric spheres centered at the origin if $c > 0$. For $c = 0$ the level ‘surface’ is not really a surface at all; it just consists of the point at the origin. (What if $c < 0$?)

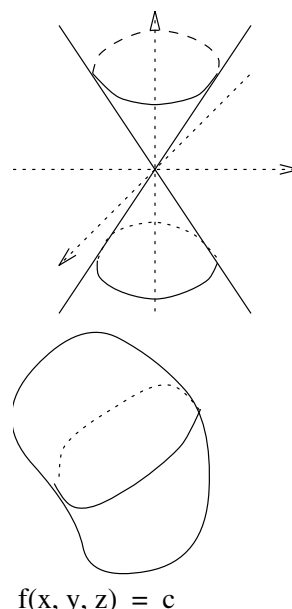
For $f(x, y, z) = x^2 + y^2 - z^2$, the level surfaces are either hyperboloids of one sheet if $c > 0$ or hyperboloids of two sheets if $c < 0$. (What if $c = 0$?)

For functions of four or more variables, geometric interpretations are even harder to come by. If $f(x_1, x_2, \dots, x_n)$ denotes a function of n variables, the locus in \mathbf{R}^n of the equation $f(x_1, x_2, \dots, x_n) = c$ is called a level set, but one doesn’t ordinarily try to visualize it geometrically.

Instead of talking about many independent variables, it is useful to think instead of a single independent variable which is a *vector*, i.e., an element of \mathbf{R}^n for some n . In the case $n = 2, 3$, we usually write $\mathbf{r} = \langle x, y \rangle$ or $\mathbf{r} = \langle x, y, z \rangle$ so $f(x, y)$ or $f(x, y, z)$ would be written simply $f(\mathbf{r})$. If $n > 3$, then one often denotes the variables x_1, x_2, \dots, x_n and denotes the vector (i.e., element of \mathbf{R}^n) by $\mathbf{x} = (x_1, x_2, \dots, x_n)$. Then $f(x_1, x_2, \dots, x_n)$ becomes simply $f(\mathbf{x})$. The case of a function of a single real variable can be subsumed in this formalism by allowing the case $n = 1$. That is, we consider a scalar x to be just a vector of dimension 1, i.e., an element of \mathbf{R}^1 .

When we talked about kinematics, we considered *vector valued* functions $\mathbf{r}(t)$ of a single independent variable. Thus we see that it makes sense to consider in general functions of a vector variable which can also assume vector values. We indicate this by the notation $f : \mathbf{R}^n \rightarrow \mathbf{R}^m$. That shall mean that the domain of the function f is a subset of \mathbf{R}^n while the set of values is a subset of \mathbf{R}^m . Thus, $n = m = 1$ would yield a scalar function of one variable, $n = 2, m = 1$ a scalar function of two variables, and $n = 1, m = 3$ a vector valued function of one scalar variable. We shall have occasion to consider several other special cases in detail.

There is one slightly non-standard aspect to the above notation. In ordinary usage



in mathematics, “ $f : \mathbf{R}^n \rightarrow \mathbf{R}^m$ ” means that \mathbf{R}^n is the entire domain of the function f , whereas we are taking it to mean that the domain is some subset. We do this mostly to save writing since usually the domain will be almost all of \mathbf{R}^n or at least some significant chunk of it. What we want to make clear by the notation is the dimensionality of both the independent and dependent variables.

Exercises for 3.1.

You are encouraged to make use of the available computer software (e.g., Maple, Mathematica, etc.) to help you picture the graphs in the following problems.

1. State the largest possible domain for the function
 - (a) $f(x, y) = e^{x^2 - y^2}$
 - (b) $f(x, y) = \ln(y^2 - x^2 - 2)$
 - (c) $f(x, y) = \frac{x^2 - y^2}{x - y}$
 - (d) $f(x, y, z) = \frac{1}{xyz}$
 - (e) $f(x, y, z) = \frac{1}{\sqrt{z^2 - x^2 - y^2}}$
2. Describe the graph of the function described by
 - (a) $f(x, y) = 5$
 - (b) $f(x, y) = 2x - y$
 - (c) $f(x, y) = 1 - x^2 - y^2$
 - (d) $f(x, y) = 4 - \sqrt{x^2 + y^2}$
 - (e) $f(x, y) = \sqrt{24 - 4x^2 - 6y^2}$
3. Sketch selected level curves for the functions given by
 - (a) $f(x, y) = x + y$
 - (b) $f(x, y) = x^2 + 9y^2$
 - (c) $f(x, y) = x - y^2$
 - (d) $f(x, y) = x - y^3$
 - (e) $f(x, y) = x^2 + y^2 + 4x + 2y + 9$
4. Describe selected level surfaces for the functions given by
 - (a) $f(x, y, z) = x^2 + y^2 - z$
 - (b) $f(x, y, z) = x^2 + y^2 + z^2 + 2x - 2y + 4z$
 - (c) $f(x, y, z) = z^2 - x^2 - y^2$

5. Describe the quadric surfaces which are loci in \mathbf{R}^3 of the following equations.
- (a) $x^2 + y^2 = 16$
 - (b) $z^2 = 49x^2 + y^2$
 - (c) $z = 25 - x^2 - y^2$
 - (d) $x = 4y^2 - z^2$
 - (e) $4x^2 + y^2 + 9z^2 = 36$
 - (f) $x^2 + y^2 - 4z^2 = 4$
 - (g) $9x^2 + 4y^2 - z^2 = 36$
 - (h) $9x^2 - 4y^2 - z^2 = 36$
6. Describe the traces of the following functions in the given planes
- (a) $z = xy$, in horizontal planes $z = c$
 - (b) $z = x^2 + 9y^2$ in vertical planes $x = c$ or $y = c$
 - (c) $z = x^2 + 9y^2$ in horizontal planes $z = c$
7. Describe the intersection of the cone $x^2 + y^2 = z^2$ with the plane $z = x + 1$.
8. Let $\mathbf{f} : \mathbf{R}^2 \rightarrow \mathbf{R}^2$.
- (a) Try to invent a definition of ‘graph’ for such a function. For what n would it be a subset of \mathbf{R}^n ?
 - (b) Try to invent a definition of ‘level set’ for such a function. For what n would it be a subset of \mathbf{R}^n ?

3.2 Limits and Continuity

Most users of mathematics don’t worry about things that might go wrong with the functions they use to represent physical quantities. They tend to assume that functions are differentiable when derivatives are called for (except possibly for a finite set of isolated points), and they assume all functions which need to be integrated are continuous so the integrals will exist. For much of the period during which Calculus was developed (during the 17th and 18th centuries), mathematicians also did not bother themselves with such matters. Unfortunately, during the 19th century, mathematicians discovered that general functions could behave in unexpected and subtle ways, so they began to devote much more time to careful formulation of definitions and careful proofs in analysis. This is an aspect of mathematics which is covered in courses in real analysis, so we won’t devote much time to it in this course. (You may have noticed that we didn’t worry about the existence of derivatives in our discussion of velocity and acceleration.) However, for functions of several variables, lack of rigor can be more troublesome than in the one variable case, so we

briefly devote some attention to such questions. In this section, we shall discuss the concepts of *limit* and *continuity* for functions $f : \mathbf{R}^2 \rightarrow \mathbf{R}$. The big step, it turns out, is going from one independent variable to two. Once you understand that, going to three or more independent variables introduces few additional difficulties.

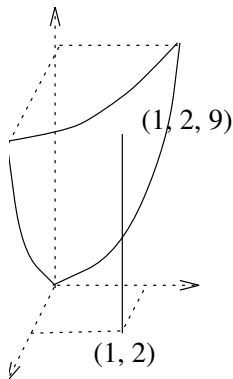
Let $\mathbf{r}_0 = \langle x_0, y_0 \rangle$ be (the position vector of) a point in the domain of the function f . We want to define the concept to be expressed symbolically

$$\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = L \quad \text{or} \quad \lim_{(x,y) \rightarrow (x_0, y_0)} f(x, y) = L.$$

We start with two examples which illustrate the concept and some differences from the single variable case.

Example 33 Let $f(x, y) = x^2 + 2y^2$, and consider the nature of the graph of f near the point $(1, 2)$. As we saw in the previous section, the graph is an elliptic paraboloid, the locus of $z = x^2 + 2y^2$. In particular, the surface is quite smooth, and if (x, y) is a point in the domain *close to* $(1, 2)$, then $f(x, y)$ will be very close to the value of the function there, $f(1, 2) = 1^2 + 2(2^2) = 9$. Thus, it makes sense to assert that

$$\lim_{(x,y) \rightarrow (1,2)} x^2 + 2y^2 = 9.$$



In Example 33, the limit was determined simply by evaluating the function at the desired point. You may remember that in the single variable case, you cannot always do that. For example, putting $x = 0$ in $\sin x/x$ yields the meaningless expression $0/0$, but $\lim_{x \rightarrow 0} \sin x/x$ is known to be 1. Usually, it requires some ingenuity to find such examples in the single variable case, but the next example shows that fairly simple formulas can lead to unexpected difficulties for functions of two or more variables.

Example 34 Let

$$f(x, y) = \frac{x^2 - y^2}{x^2 + y^2} \quad \text{for } (x, y) \neq (0, 0).$$

What does the graph of this function look like in the vicinity of the point $(0, 0)$? (Since, $(0, 0)$ is not in the domain of the function, it does not make sense to talk about $f(0, 0)$, but we can still seek a ‘limit’.) The easiest way to answer this question is to switch to polar coordinates. Using $x = r \cos \theta$, $y = r \sin \theta$, we find

$$f(\mathbf{r}) = f(x, y) = \frac{r^2 \cos^2 \theta - r^2 \sin^2 \theta}{r^2 \cos^2 \theta + r^2 \sin^2 \theta} = \cos^2 \theta - \sin^2 \theta = \cos 2\theta.$$

Thus, $f(\mathbf{r}) = f(x, y)$ is independent of the polar coordinate r and depends only on θ . As $r = |\mathbf{r}| \rightarrow 0$ with θ fixed, $f(\mathbf{r})$ is constant, and equal to $\cos 2\theta$, so, if it ‘approaches’ a limit, that limit would have to be $\cos 2\theta$. Unfortunately, $\cos 2\theta$ varies between -1 and 1 , so it does not make sense to say $f(\mathbf{r})$ has a limit as $\mathbf{r} \rightarrow \mathbf{0}$. You can get some idea of what the graph looks like by studying the level curves which

are pictured in the diagram. For each value of θ , the function is constant, so the level curves consist of rays emanating from the origin, as indicated. On any such ray, the graph is at some constant height z with z taking on *every value* between -1 and $+1$.

In general, the statement

$$\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = L$$

will be taken to mean that $f(\mathbf{r})$ is *close to* L whenever \mathbf{r} is *close to* \mathbf{r}_0 . As in the case of functions of a single scalar variable, this can be made completely precise by the following ‘ ϵ, δ ’ definition.

For each number $\epsilon > 0$, there is a number $\delta > 0$ such that

$$0 < |\mathbf{r} - \mathbf{r}_0| < \delta \quad \text{implies} \quad |f(\mathbf{r}) - L| < \epsilon.$$

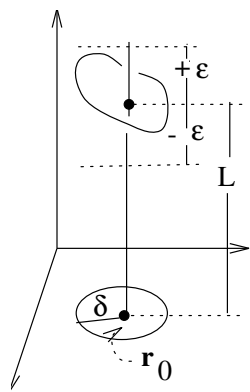
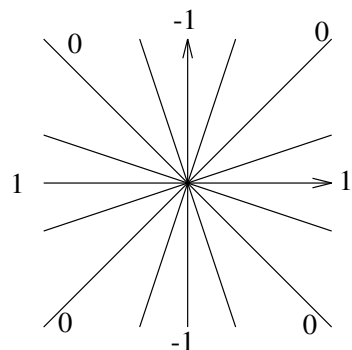
In this statement, $|\mathbf{r} - \mathbf{r}_0| < \delta$ asserts that the distance from \mathbf{r} to \mathbf{r}_0 is less than δ . Since δ is thought of as small, the inequality makes precise the meaning of ‘ \mathbf{r} is close to \mathbf{r}_0 ’. Similarly, $|f(\mathbf{r}) - L| < \epsilon$ catches the meaning of ‘ $f(\mathbf{r})$ is close to L ’. Note that we never consider the case $\mathbf{r} = \mathbf{r}_0$, so the value of $f(\mathbf{r}_0)$ is not relevant in checking the limit as $\mathbf{r} \rightarrow \mathbf{r}_0$. (It is not even necessary that $f(\mathbf{r})$ be well defined at $\mathbf{r} = \mathbf{r}_0$.)

Limits for functions of several variables behave formally much the same as limits for functions of one variable. Thus, you may calculate the limit of a sum by taking the sum of the limits, and similarly for products and quotients (except that for quotients the limit of the denominator should not be zero). The understanding you gained of these matters in the single variable case should be an adequate guide to what to expect for several variables. If you never really understood all this before, we won’t enlighten you much here. You will have to wait for a course in real analysis for real understanding.

Continuity In Example 33, the limit was determined simply by evaluating the function at the point. This is certainly not always possible because the value of the function may be irrelevant or there may be no meaningful way to attach a value. Functions for which it is always possible to find the limit this way are called *continuous*. (This is the same notion as for functions of a single scalar variable). More precisely, we say that f is continuous at a point \mathbf{r}_0 if the point is in its domain (i.e., $f(\mathbf{r}_0)$ is defined) and

$$\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = f(\mathbf{r}_0).$$

Points at which this fails are called *discontinuities* or sometimes *singularities*. (The latter term is also sometimes reserved for less serious kinds of mathematical pathology.) It sometimes happens, that a function f has a well defined limit L at a point \mathbf{r}_0 which does not happen to be in the domain of the function, i.e., $f(\mathbf{r}_0)$ is not defined. (In the single variable case, $\sin x/x$ at $x = 0$ is a good example.) Then we



can extend the domain of the function to include the point \mathbf{r}_0 by defining $f(\mathbf{r}_0) = L$. Thus the original function had a discontinuity, but it can be eliminated simply by extending the definition of the function. In this case, the discontinuity is called *removable*. As Example 34 shows, there are functions with discontinuities which cannot be defined away no matter what you try.

A function without discontinuities is called continuous. Continuous functions have graphs which look reasonably smooth. They don't have big holes or sudden jumps, but as we shall see later, they can still look pretty bizarre. Usually, just knowing that a function is continuous won't be enough to make it a good candidate to represent a physical quantity. We shall also want to be able to take derivatives and do the usual things one does in differential calculus, but as you might expect, this is somewhat more involved than it is in the single variable case.

Exercises for 3.2.

1. Use the examples of limits and the definition of continuity in this section to find the following, if they exist.

$$(a) \lim_{(x,y) \rightarrow (-1,1)} e^{-xy}$$

$$(b) \lim_{(x,y) \rightarrow (0,0)} \sin \sqrt{1 - x^2 - y^2}$$

$$(c) \lim_{(x,y,z) \rightarrow (0,0,0)} e^{-\frac{1}{x^2+y^2+z^2}}$$

$$(d) \lim_{(x,y,z) \rightarrow (0,0,1)} \ln \frac{xyz}{\sqrt{1 - x^2 - y^2 - z^2}}$$

2. Change from rectangular to polar coordinates to determine the following limits when they exist.

$$(a) \lim_{(x,y) \rightarrow (1,0)} (x^2 + y^2) \sqrt{x^2 + y^2}$$

$$(b) \lim_{(x,y) \rightarrow (0,0)} \frac{xy}{\sqrt{x^2 + y^2}}$$

$$(c) \lim_{(x,y) \rightarrow (0,0)} \frac{x^2 - y^2}{x^2 + y^2}$$

3.3 Partial Derivatives

Given a function f of two or more variables, its *partial derivative* with respect to one of the independent variables is what is obtained by differentiating with respect to that variable while keeping all other variables constant.

Example In thermodynamics, the function defined by

$$p = f(v, T) = k \frac{T}{v}$$

expresses the pressure p in terms of the volume v and the temperature T in the case of an ‘ideal gas’. Here v and T are considered to be independent variables, and k is a constant. ($k = nR$ where n is the number of moles of the gas and R is a physical constant.) The partial derivative with respect to v (keeping T constant) is

$$-k \frac{T}{v^2}$$

while the partial derivative with respect to T (keeping v constant) is

$$k \frac{1}{v}.$$

Notation One uses a variety of notations for partial derivatives. For example, for a function f of two variables,

$$\frac{\partial f}{\partial x}(x, y) \quad \text{and} \quad f_x(x, y)$$

are used to denote the partial derivative with respect to x (y kept constant).

Example

$$\begin{aligned} f(x, y) &= 2x + \sin(xy) \\ f_x(x, y) &= 2 + \cos(xy) \quad y = 2 + y \cos(xy) \\ f_y(x, y) &= 0 + \cos(xy) \quad x = x \cos(xy). \end{aligned}$$

In some circumstances, the variable names may change frequently in the discussion, so the partial derivative is indicated by an numerical subscript giving the position of the relevant variable. Thus,

$$f_2(x, y, z, t)$$

denotes the partial derivative of $f(x, y, z, t)$ with respect to the second variable, in this case y . In thermodynamics, one may see things like

$$\left(\frac{\partial p}{\partial v} \right)_T$$

which is interpreted as follows. It is supposed there is a functional dependence $p = p(v, T)$ and the notation represents the partial derivative of this function with respect to v with T kept constant.

It should be emphasized that just as in calculus of one variable, it is only *functions* which can have derivatives (partial or not). It does not make sense to ask for the

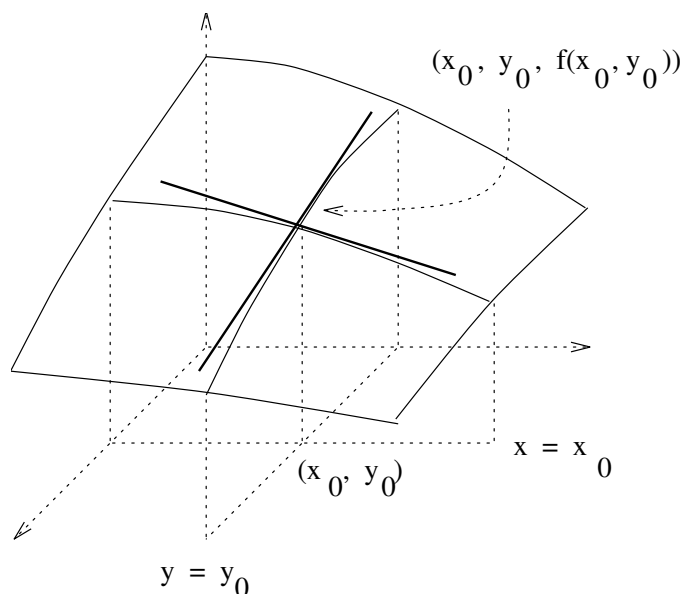
rate of change of one variable with respect to another without assuming there is a specific functional relation between the two. In the many variable case, since there are other variables which may be interrelated in complex ways, it is specially important to get this distinction straight.

If $f(x, y)$ describes a function of two variables, its partial derivatives could in fact be defined directly by

$$f_x(x, y) = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x}$$

$$f_y(x, y) = \lim_{\Delta y \rightarrow 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}.$$

Geometric Interpretation for $f : \mathbf{R}^2 \rightarrow \mathbf{R}$ Let f be a function with domain a subset of \mathbf{R}^2 and assuming scalar values. Fix a point (x_0, y_0) in the domain of f . We shall give geometric interpretations of $f_x(x_0, y_0)$ and $f_y(x_0, y_0)$ in terms of the graph of the function f . First consider the function of x given by $f(x, y_0)$ (i.e., y is kept constant, and x varies in the vicinity of $x = x_0$). The graph of $z = f(x, y_0)$ may be viewed as the curve in which the plane $y = y_0$ intersects the graph of f . It is called the sectional curve in the x -direction. The partial derivative $f_x(x_0, y_0)$ is the slope of this curve for $x = x_0$. In other words, it is the slope of the tangent line to the curve at the point $(x_0, y_0, f(x_0, y_0))$ on the graph. Similarly, fixing $x = x_0$, and letting y vary leads to the sectional curve in the y -direction. (It is the intersection of the plane $x = x_0$ with the graph of the function.) Its slope for $y = y_0$ is the partial derivative $f_y(x_0, y_0)$. Study the diagram to see how the two sectional curves and their tangents at the common point $(x_0, y_0, f(x_0, y_0))$ are related to one another. Note in particular that they lie in two mutually perpendicular planes.



The two tangent lines to the sectional curves determine a plane through the point $(x_0, y_0, f(x_0, y_0))$. It is reasonable to think of it as being *tangent* to the surface at that point. Put $z_0 = f(x_0, y_0)$. From the above discussion, it is clear that the first sectional tangent (in the x -direction) may be characterized by the equations

$$z - z_0 = f_x(x_0, y_0)(x - x_0), \quad y = y_0.$$

This characterizes it as the intersection of *two planes*. Similarly, the other sectional tangent (in the y -direction) may be characterized by the equations

$$z - z_0 = f_y(x_0, y_0)(y - y_0), \quad x = x_0.$$

However, the plane characterized by the equation

$$z - z_0 = f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) \quad (25)$$

contains both these lines, the first by intersecting with $y = y_0$ and the second by intersection with $x = x_0$. It follows that (25) is the equation of the desired tangent plane.

Example Let $f(x, y) = x^2 - y^2$. We find the tangent plane at $(1, 1, 0)$ ($x_0 = 1, y_0 = 1, z_0 = 1^2 - 1^2 = 0$). We have

$$\begin{aligned} f_x(x, y) &= 2x = 2 \quad \text{at } x = 1, y = 1, \\ f_y(x, y) &= -2y = -2 \quad \text{at } x = 1, y = 1. \end{aligned}$$

Hence, the tangent plane has equation

$$z - 0 = 2(x - 1) + (-2)(y - 1)$$

or

$$2x - 2y - z = 0$$

You should try to sketch the surface and the tangent plane. You may find the picture somewhat surprising.

Exercises for 3.3.

1. Compute all first order partial derivatives of the following functions
 - (a) $f(x, y) = x^5 - 4x^4 + 2x + 10$
 - (b) $f(x, y) = x \cos y + y \cos x$
 - (c) $f(x, y) = \ln(x^2 - y^2)$
 - (d) $f(x, y, z) = x^2 y^3 z^4$
 - (e) $f(u, v, w) = ue^w + ve^v + we^u$
2. Find an equation for the tangent plane to the graph of the function at the given point:
 - (a) $f(x, y) = x^2 - y^2$; $P(5, 4, 9)$
 - (b) $f(x, y) = \cos \frac{\pi xy}{4}$; $P(4, 2, 1)$
 - (c) $f(x, y) = xy$; $P(3, -2, -6)$
3. The ideal gas law may be written $pv = nRT$. Show that

$$\left(\frac{\partial p}{\partial v}\right)_T \left(\frac{\partial v}{\partial T}\right)_p \left(\frac{\partial T}{\partial p}\right)_v = -1$$

Hint: To find $\partial p / \partial v$, for example, you could solve for p in terms of v and T as above in the text. Alternatively, you could *assume* that p is a function of v and T , and differentiate the equation $pv = nRT$ with respect to v treating v and T as independent variables.

4. Let (x_0, y_0, z_0) be a point on the upper cone defined by the equation $z^2 = x^2 + y^2$.
 - (a) Find an equation for the tangent plane at (x_0, y_0, z_0) . Hint: You may solve for z in terms of x and y and determine the partial derivatives $\partial z / \partial x$ and $\partial z / \partial y$ at the point (x_0, y_0, z_0) . Alternatively, you may *assume* z is a function of x and y and differentiate the equation $z^2 = x^2 + y^2$ with respect to x and y treating them as independent variables.
 - (b) Show that the plane determined in part (a) passes through the origin. Is this reasonable on geometric grounds?

5. Newton's Method is an iterative process used to solve systems of equations of the form $f(x) = 0$. There is a generalization for a system of two equations in two unknowns: $f(x, y) = 0$, $g(x, y) = 0$. Each equation has as locus a curve in the x, y -plane. Consider a *guess* (x_0, y_0) for the point of intersection of these two curves. The point $(x_0, y_0, f(x_0, y_0))$ is on the surface $z = f(x, y)$ and the point $(x_0, y_0, g(x_0, y_0))$ is on the second surface $z = g(x, y)$. The tangent planes to these surfaces at these two points and the xy -plane intersect in a point (x_1, y_1) which—we hope—is closer to the desired intersection than (x_0, y_0) .

(a) Verify that the the following equations give the coordinates of (x_1, y_1) .

$$\begin{aligned}x_1 &= x_0 - \frac{f(x_0, y_0)g_y(x_0, y_0) - g(x_0, y_0)f_y(x_0, y_0)}{f_x(x_0, y_0)g_y(x_0, y_0) - g_x(x_0, y_0)f_y(x_0, y_0)} \\y_1 &= y_0 - \frac{g(x_0, y_0)f_x(x_0, y_0) - f(x_0, y_0)g_x(x_0, y_0)}{f_x(x_0, y_0)g_y(x_0, y_0) - g_x(x_0, y_0)f_y(x_0, y_0)}.\end{aligned}$$

- (b) Are there circumstances in which no such point will exist?
- (c) The above process may be iterated to find a solution of a pair of equations. Choose some arbitrary guess, use the above equations to obtain another guess, do the same for the new guess to obtain a third guess, etc. Continue this process until you have sufficient accuracy for your purposes.

Apply this method to the system $x^3 + y^2 + 2xy^2 = 0$, $x^2 - y^2 + 4 = 0$. Try 3 or 4 iterations.

(d) Clearly, you would be better off doing part (c) on a computer. Write a computer program to run the algorithm. Run the program for many more iterations.

3.4 First Order Approximation and the Gradient

Most functions cannot be calculated directly, and so one uses *approximations* which are accurate enough for one's needs. For example, the statement

$$e = 2.71828$$

is presumably accurate to 5 decimal places, but it is certainly not an exact equality. (In fact, e cannot be given exactly by any finite decimal. Do you know why?) You may have learned in your previous calculus course that e^x may be represented in general by an *infinite series*

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots$$

In calculations, you use as many terms as necessary to get the accuracy you need. In general, many interesting functions can be represented by *power series*, that is we have

$$f(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n + \cdots$$

(Do you know what a_n is and how it is related to the function f ? Refer to the chapter on *Taylor series* in your one variable Calculus book.) The simplest kind of approximation is the *linear approximation* where one ignores all terms of degree higher than one. The linear approximation is intimately tied up with the notion of derivative. We review what you learned in one variable calculus about derivatives, approximation, and tangent lines.

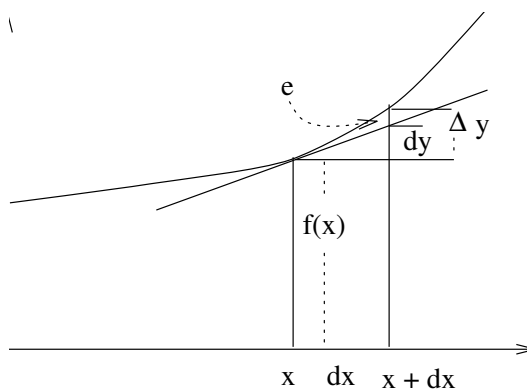
Let $f : \mathbf{R} \rightarrow \mathbf{R}$ denote a function of a single variable. The relation between the value of the function $y = f(x)$ at one point and its value $y + \Delta y = f(x + \Delta x)$ at a nearby point is given approximately by

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x.$$

The quantity $f'(x)\Delta x$ may be interpreted geometrically as the change in y along the tangent line. (See the diagram.) It is also sometimes expressed in *differential notation*

$$dy = f'(x)dx$$

where for consistency we put $\Delta x = dx$. Here, dy is only an approximation for the true change Δy , but one often acts as though they were equal. (Differentials are a wonderful tool for doing calculations, but sometimes it requires great ingenuity to see why the calculations really work.)



If we want to be completely accurate, we should write instead

$$f(x + \Delta x) = f(x) + f'(x)\Delta x + e(x, \Delta x)$$

where e is an “error term” representing the difference between the value of the y -coordinate on the graph of the function and y -coordinate on the tangent line. Since

the tangent line is very close to the graph, we expect e to be very small, at least if Δx is small.

One way to think of this is that there is an infinite expansion in “higher order terms” which involve powers of Δx

$$f(x + \Delta x) = f(x) + f'(x)\Delta x + \dots$$

and the tangential approximation ignores all terms of degree greater than one. e would be the sum of these additional terms. (This may have been brought up in your previous Calculus course as part of a discussion of Taylor’s formula.)

To proceed further, we need to be careful about what we mean by $e(x, \Delta x)$ “being very small”. Indeed, all the incremental quantities will be small, so we must ask “small compared to what?” To answer this question, rewrite (26) by transposing and dividing by Δx

$$\frac{f(x + \Delta x) - f(x)}{\Delta x} = f'(x) + \frac{e(x, \Delta x)}{\Delta x}.$$

Letting $\Delta x \rightarrow 0$, we see that the left hand side approaches $f'(x)$ as a limit, so it follows that

$$\lim_{\Delta x \rightarrow 0} \frac{e(x, \Delta x)}{\Delta x} = 0.$$

This says that if Δx is small, the ratio $e/\Delta x$ will be small, i.e., that e is small even when compared to Δx . A simple example will illustrate this.

Example 35 Let $f(x) = x^3$, $x = 2$. Then

$$f(x + \Delta x) = (2 + \Delta x)^3 = 8 + 12\Delta x + 6\Delta x^2 + \Delta x^3.$$

Here, $f'(x) = f'(2) = 3 \cdot 2^2 = 12$. Hence, $f(2) + f'(2)\Delta x = 8 + 12\Delta x$, and $e(2, \Delta x) = 6\Delta x^2 + \Delta x^3$. Indeed,

$$\frac{e}{\Delta x} = 6\Delta x + \Delta x^2.$$

Thus, if $\Delta x = .01$ (a fairly small number), $e/\Delta x = .0601$, and $e = .000601$ which is quite a bit smaller than $\Delta x = .01$.

The theory of linear approximation for functions of two (or more variables) is similar, but complicated by the fact that more things are allowed to vary. We start with an example.

Example 36 Let $f(\mathbf{r}) = f(x, y) = x^2 + xy$, and let $\mathbf{r} = (x, y) = (1, 2)$. Then, proceeding as in Example 35, we have

$$\begin{aligned} f(x + \Delta x, y + \Delta y) &= f(1 + \Delta x, 2 + \Delta y) \\ &= (1 + \Delta x)^2 + (1 + \Delta x)(2 + \Delta y) \\ &= 1 + 2\Delta x + \Delta x^2 + 2 + 2\Delta x + \Delta y + \Delta x\Delta y \\ &= 3 + 4\Delta x + \Delta y + \Delta x^2 + \Delta x\Delta y. \end{aligned} \tag{27}$$

These terms can be grouped naturally. $f(1, 2) = 3$ so the first three terms may be written

$$f(1, 2) + 4\Delta x + \Delta y$$

and these constitute the linear terms or linear approximation to $f(x + \Delta x, y + \Delta y)$. Denote the remaining terms

$$e(\mathbf{r}, \Delta \mathbf{r}) = \Delta x^2 + \Delta x \Delta y.$$

In the one variable case, we compared e to Δx , but now we have both Δx and Δy to contend with. The way around this is to use $\Delta s = |\Delta \mathbf{r}| = \sqrt{\Delta x^2 + \Delta y^2}$. Thus,

$$\frac{e}{\Delta s} = \frac{\Delta x^2 + \Delta x \Delta y}{\Delta s} = \frac{\Delta x}{\Delta s} \Delta x + \frac{\Delta y}{\Delta s} \Delta y,$$

and the quantity on the right approaches zero as $\Delta s \rightarrow 0$. (The reasoning is that since $|\Delta x|, |\Delta y| < \Delta s$, it follows that the fraction $\Delta x/\Delta s$ has absolute value never exceeding 1, while both Δx and Δy must approach 0 as $\Delta s \rightarrow 0$.) It follows that for Δs small, e is small even compared to Δs . For example, let $\Delta x = .003, \Delta y = .004$. Then, $\Delta s = .005$, while $e = (.003)^2 + (.003)(.004) = 0.000021$.

Note that the coefficients of Δx and Δy are just the partial derivatives $f_x(1, 2)$ and $f_y(1, 2)$. This is not very surprising. We can see why it works by calculating

$$f_x(1, 2) = \lim_{\Delta x \rightarrow 0} \frac{f(1 + \Delta x, 2) - f(1, 2)}{\Delta x}$$

which from (27) with $\Delta y = 0$

$$\begin{aligned} &= \lim_{\Delta x \rightarrow 0} \frac{4\Delta x + \Delta x^2}{\Delta x} \\ &= \lim_{\Delta x \rightarrow 0} (4 + \Delta x) = 4. \end{aligned}$$

A similar argument shows that $f_y(1, 2)$ is the coefficient of Δy (which is 1).

In general, suppose we have a function $f : \mathbf{R}^2 \rightarrow \mathbf{R}$. Fix a point in the domain of f with position vector $\mathbf{r} = \langle x, y \rangle$, and consider the change in the function when we change \mathbf{r} by $\Delta \mathbf{r} = \langle \Delta x, \Delta y \rangle$. It may be possible to express f near \mathbf{r} by a linear approximation, i.e., to write

$$f(x + \Delta x, y + \Delta y) = f(x, y) + a\Delta x + b\Delta y + e(\mathbf{r}, \Delta \mathbf{r}) \quad (28)$$

where

$$\lim_{\Delta \mathbf{r} \rightarrow 0} \frac{e(\mathbf{r}, \Delta \mathbf{r})}{|\Delta \mathbf{r}|} = 0.$$

(This last statement says that e is small compared to $\Delta s = |\Delta \mathbf{r}|$ when the latter quantity is small enough.) If this is the case, it is not hard to see, just as in the

example, that

$$a = \frac{\partial f}{\partial x}$$

$$b = \frac{\partial f}{\partial y}.$$

So (28) may be rewritten

$$f(x + \Delta x, y + \Delta y) = f(x, y) + f_x(x, y)\Delta x + f_y(x, y)\Delta y + e(\mathbf{r}, \Delta \mathbf{r}). \quad (29)$$

If this is so, we say that the function is *differentiable* at the point \mathbf{r} in its domain. Equation (29) may be interpreted as follows. The first term on the right $f(x, y)$ is the value of the function at the base point. Added to this is the *linear part of the change in the function*. This has two parts: the partial change $f_x\Delta x$ due only to the change in x and the partial change $f_y\Delta y$ due only to the change in y . Each partial change is appropriately the rate of change for that variable times the change in the variable. Finally, added to this is the discrepancy e resulting from ignoring all but the linear terms.

Equation (29) also has a fairly simple geometric interpretation. Recall from the previous section that the tangent plane (determined by the sectional tangents) at the point $(x_0, y_0, z_0 = f(x_0, y_0))$ in the graph of f has equation

$$z - z_0 = f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0).$$

Except for a change in notation this is exactly what we have for the middle two terms on the right of (29). We have just changed the names of the coordinates of the base point from (x_0, y_0) to (x, y) , and the increments in the variables from $(x - x_0, y - y_0)$ to $(\Delta x, \Delta y)$. Hence, f is differentiable at (x, y) exactly when the tangent plane at $(x, y, f(x, y))$ is a good approximation to the graph of the function, at least if we stay close to the point of tangency. **The Gradient** The *gradient* of

a function $f : \mathbf{R}^2 \rightarrow \mathbf{R}$ is defined to be the vector with components $\langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \rangle$. It is denoted ∇f (pronounced “del f ”) or $\text{grad } f$.

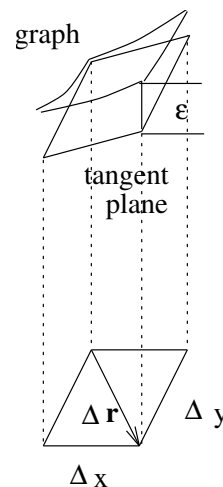
Example Let $f(x, y) = x^2 + xy + 2y^2$. Then $\nabla f = \langle 2x + y, x + 4y \rangle$. It may also be expressed using the unit vectors \mathbf{i} and \mathbf{j} by $\nabla f(x, y) = (2x + y)\mathbf{i} + (x + 4y)\mathbf{j}$.

Notice that the gradient is actually a function of position (x, y) .

The gradient may be used to simplify the expression for the linear approximation. Namely, $f_x\Delta x + f_y\Delta y$ can be rewritten as $\nabla f \cdot \Delta \mathbf{r}$. Hence, we can write the differentiability condition purely in vector notation

$$f(\mathbf{r} + \Delta \mathbf{r}) = f(\mathbf{r}) + \nabla f \cdot \Delta \mathbf{r} + e(\mathbf{r}, \Delta \mathbf{r})$$

where $\frac{e(\mathbf{r}, \Delta \mathbf{r})}{|\Delta \mathbf{r}|} \rightarrow 0$ as $\Delta \mathbf{r} \rightarrow 0$. If you look carefully, this looks quite a bit like the corresponding equation in the single variable case with ∇f playing the role of



the ordinary derivative. For this reason, it makes sense to think of ∇f as a higher dimensional derivative.

Note that much of the discussion in this section could have been done for functions $f : \mathbf{R}^3 \rightarrow \mathbf{R}$ of three variables (or indeed for functions of any number of variables). Thus, for a function of three variables given by $f(\mathbf{r}) = f(x, y, z)$, the gradient $\nabla f = \frac{\partial f}{\partial x} \mathbf{i} + \frac{\partial f}{\partial y} \mathbf{j} + \frac{\partial f}{\partial z} \mathbf{k}$. Also the differentiability condition looks the same in vector notation:

$$f(\mathbf{r} + \Delta \mathbf{r}) = f(\mathbf{r}) + \nabla f \cdot \Delta \mathbf{r} + e(\mathbf{r}, \Delta \mathbf{r})$$

where $e(\mathbf{r}, \Delta \mathbf{r})/|\Delta \mathbf{r}| \rightarrow 0$ as $|\Delta \mathbf{r}| \rightarrow 0$. However, when written out in terms of components, this becomes

$$f(x + \Delta x, y + \Delta y, z + \Delta z) = f(x, y, z) + f_x \Delta x + f_y \Delta y + f_z \Delta z + e.$$

The geometric meaning of all this is much less clear since the number of dimensions is too high for clear visualization. Hence, we usually develop the theory in the two variable case, and then proceed by analogy in higher dimensions. Fortunately, the notation need not change much if we consistently use vectors.

‘O’ and ‘o’ notation In analysis and its applications, one is often interested in the general behavior of functions rather than their precise forms. Thus, we don’t really care how to express the error term $e(\mathbf{r}, \mathbf{r}_0)$ exactly as a formula, but we do know something about how fast it approaches zero. Similarly, the most important thing about the exponential function e^x is not its exact values, but the fact that it gets large very fast as $x \rightarrow \infty$. The term “order of magnitude” is often used in these contexts. There are two common notations used by scientists and mathematicians in this context. We would say that a quantity e is ‘ $O(\Delta s)$ ’ as $\Delta s \rightarrow 0$ if the ratio $e/\Delta s$ stays bounded. That means they have roughly the same order of magnitude. We say that e is ‘ $o(\Delta s)$ ’ if the ratio $e/\Delta s$ goes to zero. That means that e is an order of magnitude (or more) smaller than Δs . ‘ o ’ is stronger than ‘ O ’ in that the former implies the latter, but not necessarily vice versa.

With this terminology, we could express the differentiability condition

$$f(\mathbf{r} + \Delta \mathbf{r}) = f(\mathbf{r}) + \nabla f \cdot \Delta \mathbf{r} + o(|\Delta \mathbf{r}|).$$

You might also see the following

$$f(\mathbf{r} + \Delta \mathbf{r}) = f(\mathbf{r}) + \nabla f \cdot \Delta \mathbf{r} + O(|\Delta \mathbf{r}|^2).$$

This assumes more information about the error than the previous formula. It asserts that the error behaves like the square $|\Delta \mathbf{r}|^2$ (or better) whereas the previous statement is not that explicit. For almost all interesting functions the ‘ $O(|\Delta \mathbf{r}|^2)$ ’ is valid, but mathematicians, always wanting to use the simplest hypotheses, usually develop the subject using the less restrictive ‘ o ’ estimate.

Differential Notation Let $f : \mathbf{R}^2 \rightarrow \mathbf{R}$ denote a function of two variables, and fix a point \mathbf{r} in its domain. As in the single variable case, we write

$$dz = \nabla f \cdot d\mathbf{r} = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy \quad (30)$$

(where we put $\Delta x = dx, \Delta y = dy$, and $\Delta \mathbf{r} = d\mathbf{r}$.) This is the change in z *in the tangent plane* corresponding to a change $d\mathbf{r} = \Delta \mathbf{r}$. It should be distinguished from the change

$$\Delta z = \nabla f \cdot \Delta \mathbf{r} + e.$$

in the function itself. The expression on the right of (30) is often called the *total differential* of the function. As in the one variable case, one can use differentials for “quick and dirty” estimates. What we are doing in essence is assuming all functions are linear, and as long as all changes are small, this is not a bad assumption.

Example 36 The pressure of an ideal gas is given in terms of the volume and Temperature by the relation

$$p = p(v, T) = k \frac{T}{v}$$

where k is a constant. Suppose $v = 10, T = 300$ and both v and T increase by 1%. What is the change in p ? To get an approximate answer to this question, we calculate the total differential

$$dp = \frac{\partial p}{\partial v}dv + \frac{\partial p}{\partial T}dT = -k \frac{T}{v^2}dv + k \frac{1}{v}dT.$$

Putting $v = 10, dv = .1, T = 300, dT = 3$, we get

$$dp = -k \frac{300}{10^2} \cdot 1 + k \frac{1}{10} 3 = 0$$

so to a first approximation, there is no change in p . (The actual values of v and T were not relevant. Can you see why? Hint: Calculate dp/p in general.)

Calculations with differentials work just as well for functions of any number of variables. They amount to a use of the linear approximation. The only difficulty is that one can't easily visualize things in terms of “tangent planes” since the number of dimensions is too large.

Exercises for 3.4.

1. Write out the linear approximation in each case in terms of $\Delta x, \Delta y$, (and if appropriate, Δz) at the indicated point P . Use it to estimate the value of the function at the indicated point Q , and compare the estimate with the ‘true value’ determined from your calculator.

- (a) $f(x, y) = \frac{1}{x+y}$; $P(1, 1)$; $Q(1.02, 0.97)$
- (b) $f(x, y) = \sqrt{x^2 + y^2}$; $P(3, 4)$; $Q(3.02, 3.98)$
- (c) $f(x, y, z) = xyz$; $P(1, 1, 1)$; $Q(0.98, 1.02, 1.02)$
2. Use differentials to approximate
- (a) $(\sqrt{17} + \sqrt{63})^2$
- (b) $(\sqrt{5} - \sqrt{3})^2$
3. The period of a simple pendulum is given by $T = 2\pi\sqrt{\frac{L}{g}}$. Use differentials to estimate the error in the period if the length is 1.1 meter, you take it to be 1.0 meters, and you approximate g by 10.0 rather than $9.8 \frac{m}{s^2}$?
4. Find the gradient vector for the following functions at the given points:
- (a) $f(x, y) = 4x - 7y + 3$, $P(3, -2)$
- (b) $f(x, y) = x^2 - 4y^2 + y - 2$, $P(2, 1)$
- (c) $f(x, y, z) = \sqrt{x^2 + y^2 + z^2}$, $P(2, 2, 2)$
- (d) $f(x, y, z) = 3xy + y^2z - z^3$, $P(1, -1, 8)$
5. Verify the following properties of the gradient. Assume u and v are functions and a and b are constants.
- (a) $\nabla(au + bv) = a\nabla u + b\nabla v$
- (b) $\nabla(uv) = u\nabla v + v\nabla u$
6. Let $f(\mathbf{r}) = \frac{1}{2} \ln(x^2 + y^2)$. Show that $\nabla f = \frac{1}{r} \mathbf{u}_r$ where $r = \sqrt{x^2 + y^2}$ is the radial polar coordinate.
7. Let $f(x, y) = x^3 + y^2x + 3y - 1$. Expand $f(1 + \Delta x, -1 + \Delta y)$ by substituting $1 + \Delta x$ for x and $-1 + \Delta y$ for y . Identify the linear terms and compare with $\nabla f(1, -1) \cdot \Delta \mathbf{r}$. Identify the higher order terms and show that they are $O(\Delta \mathbf{r}^2)$.
8. Let $f : \mathbf{R}^2 \rightarrow \mathbf{R}$ be defined by $f(x, y) = \frac{2xy}{x^2 + y^2}$ for $(x, y) \neq (0, 0)$ and $f(0, 0) = 0$.
- (a) Show that $f_x(0, 0) = f_y(0, 0) = 0$. Hint: What are $f(x, 0)$ and $f(0, y)$?
- (b) Show that f is not continuous at $(0, 0)$. Hint: Express f using polar coordinates and show that the limit as $\mathbf{r} \rightarrow \mathbf{0}$ is not defined.

3.5 The Directional Derivative

Consider a function $f : \mathbf{R}^2 \rightarrow \mathbf{R}$. In the following discussion, refer to the diagram where we have sketched in some of the contour curves of the function. At some point \mathbf{r} in the domain of the function, pick a direction, and draw a ray emanating from the point \mathbf{r} in the indicated direction. We want to consider the rate of change of the function in that direction. This is called the *directional derivative* in the desired direction. You can think of it roughly as the rate at which you cross level curves as you move away from the point in the indicated direction.

Unfortunately, there is no standard notation for the directional derivative. One common notation is as follows. Specify the direction by an appropriate *unit vector* \mathbf{u} . The directional derivative in the direction \mathbf{u} is then denoted

$$\left. \frac{df}{ds} \right|_{\mathbf{u}}.$$

You already know about two cases. If the direction is that of the unit vector \mathbf{i} (parallel to the x -axis), the directional derivative is the partial derivative $f_x(\mathbf{r})$. Similarly, if the direction is that of \mathbf{j} , the directional derivative is the partial derivative $f_y(\mathbf{r})$. It turns out that the directional derivative in any direction can be expressed in terms of the partial derivatives. To see this, let $\Delta \mathbf{r} = \Delta s \mathbf{u}$ be a displacement through distance Δs in the direction \mathbf{u} . Then, if the function is differentiable at \mathbf{r} , we have

$$\begin{aligned} f(\mathbf{r} + \Delta \mathbf{r}) &= f(\mathbf{r}) + \nabla f(\mathbf{r}) \cdot \Delta \mathbf{r} + e \\ &= f(\mathbf{r}) + \nabla f(\mathbf{r}) \cdot (\Delta s \mathbf{u}) + e \end{aligned}$$

so

$$\frac{f(\mathbf{r} + \Delta \mathbf{r}) - f(\mathbf{r})}{\Delta s} = \nabla f(\mathbf{r}) \cdot \mathbf{u} + \frac{e}{\Delta s}.$$

However, by hypothesis, $e/\Delta s \rightarrow 0$, so

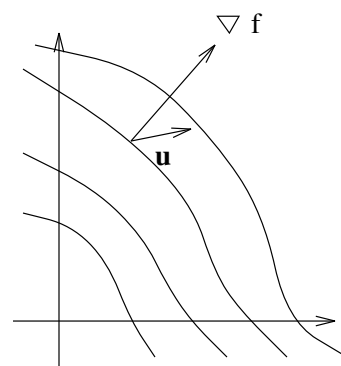
$$\lim_{\Delta s \rightarrow 0} \frac{f(\mathbf{r} + \Delta s \mathbf{u}) - f(\mathbf{r}_0)}{\Delta s} = \nabla f(\mathbf{r}_0) \cdot \mathbf{u}.$$

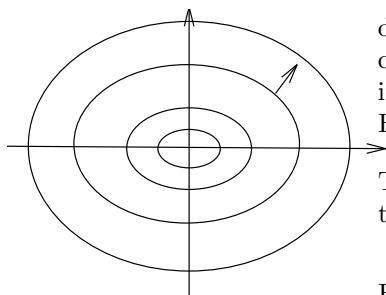
The directional derivative is the limit on the left, so we obtain

$$\left. \frac{df}{ds} \right|_{\mathbf{u}} = \nabla f(\mathbf{r}) \cdot \mathbf{u}.$$

Note that the directional derivative depends on both the point \mathbf{r} at which it is calculated and the direction in which it is calculated. Some of this may be suppressed in the notation if it is otherwise clear, so you may see for example $df/ds = \nabla f \cdot \mathbf{u}$.

Example 37 A climber ascends a mountain where the elevation in feet is given by $z = f(x, y) = 5000 - x^2 - 2y^2$. x and y refer to the coordinates (measured in feet from the summit) of a point on a *flat map* of the mountain. Most of the





discussion which follows refers to calculations involving that map. Suppose the climber finds herself at the point with map coordinates $(20, 10)$ and wishes to move in the direction (on the map) of the unit vector $(\frac{3}{5}, \frac{4}{5})$ (a bit north of northeast). How fast will she descend?

To solve this problem, we calculate the directional derivative in the indicated direction. First,

$$\nabla f = \langle -2x, -4y \rangle = \langle -40, -40 \rangle \quad \text{at } (20, 10).$$

Hence, in the indicated direction

$$\frac{df}{ds} = \langle -40, -40 \rangle \cdot \left\langle \frac{3}{5}, \frac{4}{5} \right\rangle = \frac{-120 - 160}{5} = -56.$$

Thus she descends 56 feet for each foot she moves horizontally (i.e., on the map). (She had better be using a rope!) Notice that the answer is negative which accords with the fact that the climber is descending.

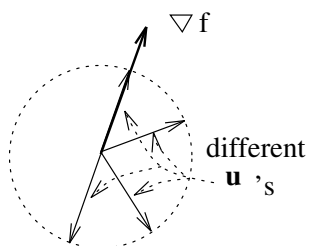
It is worth spending some time thinking about what this means on the actual surface of the mountain, i.e., the graph of the function in \mathbf{R}^3 . The point on the graph would be that with coordinates $(20, 10, 5000 - 20^2 - 2 \cdot 10^2) = (20, 10, 4400)$. *Moving on the mountain*, if the climber moves Δs units horizontally (i.e., on the map), she will move $\sqrt{\Delta s^2 + 56^2 \Delta s^2} = \Delta s \sqrt{3137}$ in space. You should make sure you visualize all of this. In particular, try to understand the relation between the unit vector \mathbf{u} in the plane, and the corresponding displacement vector in space. Can you find a vector in space which is tangent to the graph of the function and which projects on \mathbf{u} ? Can you find a unit vector in the same direction? (The answer to the first question is $\langle \frac{3}{5}, \frac{4}{5}, -56 \rangle$.)

Significance of the Gradient Fix a point \mathbf{r} in the domain of the function f , and consider the directional derivative

$$\left. \frac{df}{ds} \right|_{\mathbf{u}} = \nabla f \cdot \mathbf{u}$$

as a function of \mathbf{u} . Assume $\nabla f \neq \mathbf{0}$ at \mathbf{r} . (Otherwise, the directional derivative is always zero.) The directional derivative is 0 if \mathbf{u} is perpendicular to ∇f . It attains its maximum positive value if \mathbf{u} points in the direction of ∇f , and it attains its minimum (most negative) value if \mathbf{u} points in the direction opposite to ∇f , (i.e., $-\nabla f$.) Finally, if \mathbf{u} points the same way as ∇f , the directional derivative is

$$\frac{df}{ds} = \nabla f \cdot \mathbf{u} = |\nabla f| |\mathbf{u}| = |\nabla f|.$$



The upshot of this is that

1. the direction of the gradient is that in which the directional derivative is as large as possible;

2. the magnitude of the gradient is the directional derivative in that direction

Example 37, revisited Consider the same climber on the same mountain. In which direction should she move (on her map) to go down hill as fast as possible? By the above analysis, this is opposite to the direction of the gradient, i.e., the direction of $-\nabla f(1, 2) = \langle 40, 40 \rangle$. Directions are often given by unit vectors, so we might normalize this to $\mathbf{u} = (1/\sqrt{2})\langle 1, 1 \rangle$. Note that the question of finding a vector on the surface of the mountain pointing down hill is somewhat different. Can you solve that problem? (The answer is $\frac{1}{\sqrt{2}}\langle 1, 1, -80 \rangle$. This is not a unit vector, but you can get a unit vector in the same direction—should you need one—by dividing it by its length.)

Exercises for 3.5.

- For each of the following functions, find the directional derivative in the direction of the given vector:
 - $f(x, y) = x^2 + xy + y^2$, $P(1, -1)$, $\mathbf{v} = \langle 2, 3 \rangle$
 - $f(x, y) = e^y \sin x$, $P(\frac{\pi}{4}, 0)$, $\mathbf{v} = \langle 1, -1 \rangle$
 - $f(x, y, z) = \sqrt{x^2 + y^2 + z^2}$, $P(2, 3, 6)$, $\mathbf{v} = 2\mathbf{i} - 3\mathbf{j} + \mathbf{k}$
 - $f(x, y, z) = 20 - x^2 - y^2 - z^2$, $P(4, 2, 0)$, $\mathbf{v} = \mathbf{k}$
 - $f(x, y, z) = (x - 1)^2 + y^2 + (z + 1)^2$, $P(1, 0, -1)$, $\mathbf{v} = \langle 1, 0, 0 \rangle$
- Find the maximum directional derivative of the function at the given point and find its direction:
 - $f(x, y) = x^2 - 3y^2$, $P(1, -2)$
 - $f(x, y, z) = x^2 + 4y^3 - z$, $P(1, -3, 0)$
 - $f(x, y, z) = \sqrt{xyz}$, $P(-1, 4, -1)$
- A mountain climber stands on a mountain described by the equation $z = 10000(25 - x^2 - y^2)$ where x and y are measured in miles and z is measured in feet. The ground beneath the climber's feet is at the point $(2.5, 3.0, z)$ where z is determined by the above equation. If the climber slips in a northeasterly direction, at what rate will the fall occur? What is the angle of descent? Is there a steeper path from this point?
- An electric charge is uniformly distributed over a thin non-conducting wire of radius a centered at the origin in the x, y -plane. The electrostatic potential due to this charge at a point on the z -axis is given by $V(0, 0, z) = \frac{C}{\sqrt{a^2 + z^2}}$ where C is an appropriate constant. Find $\left. \frac{dV}{dx} \right|_{\mathbf{u}}$ at the point $(0, 0, h)$ in the direction \mathbf{u} towards the origin. Hint. You weren't given enough information to determine ∇V . Why was the information you were given sufficient?

5. The temperature at any given point (x, y, z) in the universe is given by the equation $T = 72 + \frac{1}{2}xyz$, with the origin defined as the corner of Third and Main in Omaha, Nebraska. You are on a hilly spot, $P(-2, -1, -7)$ in Lincoln, and the surrounding topography has equation $z = 3x - y^2$. If you start heading northeast towards Omaha, what initial rate of temperature change will you experience?
6. The temperature inside the cylinder described by $x^2 + y^2 \leq a^2$, $0 \leq z \leq h$ is given by $T = (x^2 + y^2)z$. What is the direction in which the temperature changes as fast as possible at the point $(a/2, a/2, h/2)$?

3.6 Criteria for Differentiability

We need a simple way to tell if a function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ is differentiable. Following the philosophy enunciated previously, we concentrate on the case $n = 2$, and proceed by analogy in higher dimensional cases.

Theorem 3.2 Let $f : \mathbf{R}^2 \rightarrow \mathbf{R}$. Suppose the partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ exist and are continuous functions in the vicinity of the point \mathbf{r} . Then f is differentiable at \mathbf{r} .

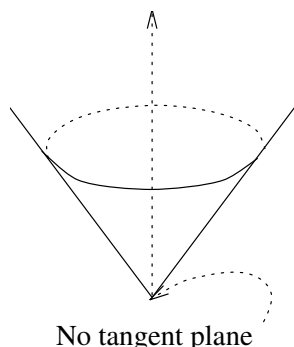
How would you generalize this result to apply to functions of three or more variables?

Examples If we take $f(x, y) = x^2 + 3y^2$, then $f_x(x, y) = 2x$, and $f_y(x, y) = 6y$. These are certainly continuous for all points (x, y) in \mathbf{R}^2 . Hence, the theorem assures us that f will be differentiable at every point in \mathbf{R}^2 . The graph of f is an elliptic paraboloid, and it is very smooth. At each point of the graph, we expect the tangent plane to be a good approximation to the graph.

On the other hand, take $f(x, y) = \sqrt{x^2 + y^2}$. Then $f_x(x, y) = x/\sqrt{x^2 + y^2}$ and $f_y(x, y) = y/\sqrt{x^2 + y^2}$. Both of these fail to be continuous at $(0, 0)$ although the function f is defined there and is even continuous. Hence, the theorem will not insure that f is differentiable at $(0, 0)$. In fact, the graph of f is a right circular cone with its vertex at the origin, and it is clear that there is no well defined tangent plane at the vertex.

There is a hierarchy of degrees of “smoothness” for functions. The lowest level is continuity, and the next level is differentiability. A function can be continuous without being differentiable. We saw an example of this in the function defined by $f(x, y) = \sqrt{x^2 + y^2}$. However, a differentiable function is necessarily continuous. (See the exercises.)

Continuity of partial derivatives provides a still higher level of smoothness. Such



functions are often called \mathcal{C}^1 functions. The theorem tells us that such functions are differentiable. However, if the partial derivatives are not continuous, we cannot necessarily conclude that the function is *not* differentiable. The theorem just does not apply. In fact there are some not too bizarre examples in which the partials are not continuous but where the function is differentiable, i.e., there is a well defined tangent plane which is a good approximation to the graph.

Proof of the Theorem While it is not essential that you understand how the theorem is proved, you might find it enlightening. The proof makes extensive use of the Mean Value Theorem, which you probably saw in your previous Calculus course. (See also *Edwards and Penney*, 3rd Edition, Section 4.3.) The Mean Value Theorem may be stated as follows. Suppose f is a function of a single variable which is defined and continuous for $a \leq x \leq b$ and which is differentiable for $a < x < b$. Then there is a point x_1 with $a < x_1 < b$ such that

$$f(b) - f(a) = f'(x_1)(b - a).$$

If we substitute x for a and Δx for $b - a$, we could also write this

$$f(x + \Delta x) - f(x) = f'(x_1)\Delta x.$$

The quantity on the right looks like the change in the linear approximation except that the derivative is evaluated at x_1 rather than at x . Also the equation is an exact equality rather than an approximation. This form of the Mean Value Theorem is better for our purposes because although in the previous analysis we had $\Delta x > 0$, i.e., $a < b$, the Mean Value Theorem is also true for $\Delta x < 0$. (Just interchange the roles of a and b .) In this form, we would say simply that x_1 lies *between* x and $x + \Delta x$ so we don't have to commit ourselves about the sign of Δx .

To prove the theorem, consider the difference $f(x + \Delta x, y + \Delta y) - f(x, y)$ which we want to relate to $f_x(x, y)\Delta x + f_y(x, y)\Delta y$. We have

$$\begin{aligned} f(x + \Delta x, y + \Delta y) - f(x, y) &= f(x + \Delta x, y + \Delta y) - f(x, y + \Delta y) \\ &\quad + f(x, y + \Delta y) - f(x, y). \end{aligned}$$

Consider the first difference on the right as a function only of the first coordinate (with the second coordinate fixed at $y + \Delta y$.) By the Mean Value Theorem, there is an x_1 between x and $x + \Delta x$ such that

$$f(x + \Delta x, y + \Delta y) - f(x, y + \Delta y) = f_x(x_1, y + \Delta y)\Delta x.$$

Consider the second difference $f(x, y + \Delta y) - f(x, y)$ as a function of the second variable, and apply the Mean Value Theorem again. There is a y_1 between y and $y + \Delta y$ such that

$$f(x, y + \Delta y) - f(x, y) = f_y(x, y_1)\Delta y.$$

Thus,

$$f(x + \Delta x, y + \Delta y) - f(x, y) = f_x(x_1, y + \Delta y)\Delta x + f_y(x, y_1)\Delta y,$$

so

$$\begin{aligned} e &= f(x + \Delta x, y + \Delta y) - f(x, y) - f_x(x, y)\Delta x - f_y(x, y)\Delta y \\ &= f_x(x_1, y + \Delta y)\Delta x + f_y(x, y_1)\Delta y - f_x(x, y)\Delta x - f_y(x, y)\Delta y \\ &= (f_x(x_1, y + \Delta y) - f_x(x, y))\Delta x + (f_y(x, y_1) - f_y(x, y))\Delta y. \end{aligned}$$

Hence,

$$\frac{e}{\Delta s} = (f_x(x_1, y + \Delta y) - f_x(x, y))\frac{\Delta x}{\Delta s} + (f_y(x, y_1) - f_y(x, y))\frac{\Delta y}{\Delta s}. \quad (32)$$

Now let $\Delta s \rightarrow 0$. Since $|\Delta x| \leq \Delta s$, it follows that $\Delta x/\Delta s$ has absolute value at most 1, and a similar argument applies to $\Delta y/\Delta s$. In addition, as $\Delta s \rightarrow 0$, so also $\Delta x \rightarrow 0$ and $\Delta y \rightarrow 0$. Since x_1 is between x and $x + \Delta x$, it follows that $x_1 \rightarrow x$. Similarly, since y_1 lies between y and $y + \Delta y$, it follows that $y_1 \rightarrow y$. Hence, since f_x and f_y are continuous functions,

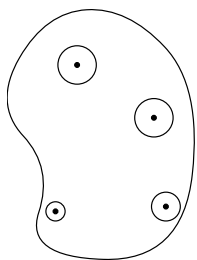
$$\begin{aligned} \lim_{\Delta s \rightarrow 0} f_x(x_1, y + \Delta y) &= f_x(x, y), \\ \lim_{\Delta s \rightarrow 0} f_y(x, y_1) &= f_y(x, y). \end{aligned}$$

However, this implies that the expressions in parentheses on the right of (32) both approach zero. It follows that

$$\lim_{\Delta s \rightarrow 0} \frac{e}{\Delta s} = 0$$

as required.

Domains of Differentiable Functions So far we haven't been very explicit about the domain of a function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ except to say that it is some subset of \mathbf{R}^n . A moment's thought will convince you that to do differential calculus, there will have to be some restriction on possible domains. For example, suppose $n = 2$ so $f(x, y)$ gives a function of two variables. If the domain were the subset of \mathbf{R}^2 which is the locus of the equation $y = x$, it would not make much sense to try to talk about the partial derivative $\partial f / \partial x$ which is supposed to be the derivative of f with y kept constant. If $y = x$, we can't vary x without also varying y . (The same would apply if there was any algebraic relation between x and y and the domain were some curve in \mathbf{R}^2 .) In order to make sense of partial derivatives and related concepts, the domain must be 'fat enough', i.e., the variables must really be independent. However, the derivatives at a point \mathbf{r} depend only on values of the variables near to the point, not on the entire domain of the function. Hence, the domain need only be 'fat' in the immediate vicinity of a point at which we want to take derivatives.

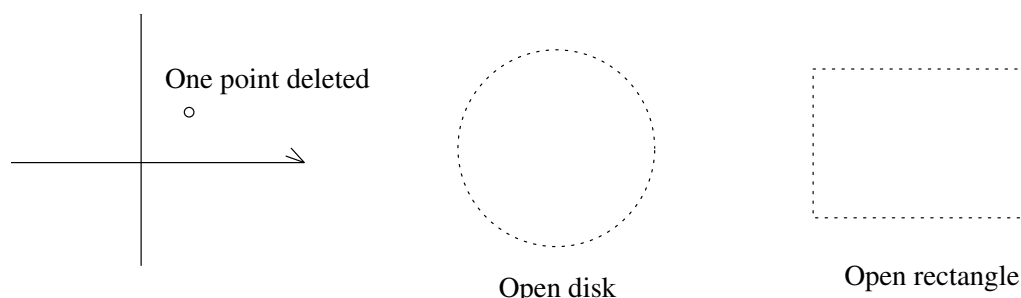


To make this precise, we introduce some new concepts. We concentrate on the case of \mathbf{R}^2 . The generalization to \mathbf{R}^3 and beyond is straightforward. A set D in \mathbf{R}^2 is said to be an *open set* if it has the following property. If a point is in D then there is an entire *disk* of some radius centered at the point which is contained in D . (A disk is a circle including the circumference and what is contained inside it.) We can state this symbolically as follows. If \mathbf{r}_0 is a point in D , then there is a number $\delta > 0$ such that all points \mathbf{r} satisfying $|\mathbf{r} - \mathbf{r}_0| \leq \delta$ are also in D .

Examples The entire plane \mathbf{R}^2 is certainly open.

Also, the set D obtained by leaving out any one point is open.

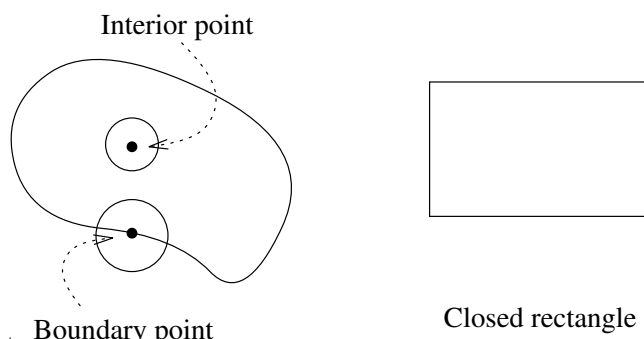
The rectangular region consisting of all points (x, y) such that $a < x < b, c < y < d$ is open, but the region defined instead by $a \leq x \leq b, c \leq y \leq d$ is not open. The points on the perimeter of the rectangle in the latter case are in the set, but they don't have the desired property since any disk centered at such a point will necessarily contain points not in the rectangle.



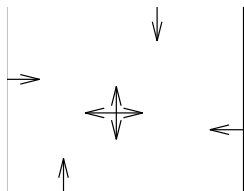
The set of all points inside a circle but not on its circumference is open, but the set of points in a circle or on its circumference is not open. The former is often called an *open disk*, and the latter is called a *closed disk*.

There are many other examples.

There are a couple of related concepts. If D is a subset of \mathbf{R}^2 , a point is said to be on its *boundary* if every disk centered at the point has some points in the set D and some points not in the set D . For example, the perimeter of a rectangle or the circumference of a disk consists of boundary points. A point in D is said to be an *interior point* if it is not on the boundary, i.e., there is some disk centered at the point consisting only of points of D . The interior of a set is always open. A set is open if it consists only of interior points. Finally, we say a set is *closed* if it contains all its boundary points.



Generally, we only want to take derivatives at interior points of the domain of a function because at such points we can move in all possible independent directions, at least if we stay sufficiently close to the point. One way to insure this is to assume that the domain of a function f is always an open set. However, there are times when we want to include the boundary of the set in the domain. At such points, we may not be able to take derivatives. However, we can sometimes do something like that if we are careful. For example, for points on the perimeter of a rectangle we can take “one sided derivatives” by allowing variation in directions which point into the rectangle.



Exercises for 3.6.

1. Let $f : \mathbf{R}^2 \rightarrow \mathbf{R}$. Show that if f is differentiable at \mathbf{r} then f is continuous at \mathbf{r} . Hint: The definition of continuity at \mathbf{r} can be rewritten

$$\lim_{\Delta \mathbf{r} \rightarrow 0} f(\mathbf{r} + \Delta \mathbf{r}) = f(\mathbf{r}).$$

Use the differentiability condition

$$f(\mathbf{r} + \Delta \mathbf{r}) = f(\mathbf{r}) + \nabla f(\mathbf{r}) \cdot \Delta \mathbf{r} + e(\mathbf{r}, \Delta \mathbf{r})$$

to verify the previous statement. What happens to e as $\Delta \mathbf{r} \rightarrow 0$?

2. For each of the following subsets of \mathbf{R}^2 , tell if it is open, closed, or neither.
 - (a) The set of (x, y) satisfying $-1 < x < 1, 0 < y < 2$.
 - (b) The set of (x, y) satisfying $-1 < x \leq 1, 0 < y < 2$.
 - (c) The set of (x, y) satisfying $x \leq y \leq 1, 0 \leq x \leq 2$. (Region is triangular.)
 - (d) The set of all points in \mathbf{R}^2 with the exception of the four points $(\pm 1, \pm 1)$.
 - (e) The set of all points with the exception of the line with equation $y = 2x + 3$.
3. Describe the set of all points in \mathbf{R}^3 with position vector \mathbf{r} satisfying $1 < |\mathbf{r}| < 2$. Is it open or closed?

3.7 The Chain Rule

The chain rule for functions of a single variable tells us how to find the derivative of a function of a function, i.e., a composite function. Thus, if $y = f(x)$ and $x = g(t)$, then

$$\frac{d}{dt}(f(g(t))) = \frac{df}{dx}(x) \frac{dg}{dt}(t) \quad \text{with } x = g(t).$$

The generalization to higher dimensions is quite straightforward, at least if we use vector notation, but its elaboration in terms of components can look pretty involved. Suppose $z = f(\mathbf{r})$ describes a differentiable function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ and $\mathbf{r} = \mathbf{g}(t)$ a vector valued differentiable function $\mathbf{g} : \mathbf{R} \rightarrow \mathbf{R}^n$. (We shall concentrate on the two cases $n = 2$ and $n = 3$, but it all works just as well for any n .) Then $z = f(\mathbf{g}(t))$ describes the composite function $f \circ \mathbf{g} : \mathbf{R} \rightarrow \mathbf{R}$ which is just a scalar function of a single variable. The multidimensional chain rule asserts

$$\frac{d}{dt}(f(\mathbf{g}(t))) = \nabla f(\mathbf{r}) \cdot \frac{d\mathbf{g}}{dt} \quad \text{with } \mathbf{r} = \mathbf{g}(t). \quad (33)$$

Note that the gradient ∇f plays the role that the derivative plays in the single variable case.

Formula (33) looks quite simple in vector form, but it becomes more elaborate if we express things in terms of component functions. Let $h(t) = f(\mathbf{g}(t))$ denote the composite function. Then, for $n = 2$, we have $\nabla f = \langle \partial f / \partial x, \partial f / \partial y \rangle$ and $d\mathbf{g}/dt = \langle dx/dt, dy/dt \rangle$. Thus, the chain rule becomes

$$\frac{dh}{dt} = \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt}. \quad (34)$$

Similarly, for $n = 3$, the chain rule becomes

$$\frac{dh}{dt} = \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt} + \frac{\partial f}{\partial z} \frac{dz}{dt}. \quad (35)$$

Example 38 Let $w = f(x, y, z) = e^{xy+z}$, $x = t$, $y = t^2$, $z = t^3$. (Thus, $\mathbf{g}(t) = \langle t, t^2, t^3 \rangle$.) Then

$$\begin{aligned} \frac{\partial f}{\partial x} &= e^{xy+z} y = e^{2t^3} t^2 \\ \frac{\partial f}{\partial y} &= e^{xy+z} x = e^{2t^3} 2t \\ \frac{\partial f}{\partial z} &= e^{xy+z} = e^{2t^3}, \end{aligned}$$

and

$$\begin{aligned}\frac{dx}{dt} &= 1 \\ \frac{dy}{dt} &= 2t \\ \frac{dz}{dt} &= 3t^2.\end{aligned}$$

Putting these in formula (35) yields

$$\frac{dh}{dt} = e^{2t^3} t^2 + e^{2t^3} t \cdot 2t + e^{2t^3} 3t^2 = e^{2t^3} (6t).$$

There are a couple of things to notice in the example. First, in principle, the *intermediate* variables x, y, z should be expressed in terms of the ultimate independent variable t . Otherwise, the answer might be considered incomplete. Secondly, the derivative could have been calculated by first making the substitutions, and then taking the derivative.

$$\begin{aligned}w = h(t) &= e^{t^2+t^3} = e^{2t^3} \\ \frac{dh}{dt} &= e^{2t^3} (2 \cdot 3t^2) = e^{3t^3} (6t^2).\end{aligned}$$

Here we only needed to use the single variable chain rule (to calculate the derivative of e^u with $u = 2t^3$), and the calculation was much simpler than that using the multidimensional chain rule. This is almost always the case. About the only exception would be that in which we happened to know the partial derivatives $\partial f/\partial x, \partial f/\partial y$, and $\partial f/\partial z$, but we did not know the function f explicitly. In fact, unlike the single variable chain rule, the multidimensional chain rule is a tool for theoretical analysis rather than an aid in calculating derivatives. In that role, it amply justifies itself.

Proof of the Chain Rule

Proof. To prove the chain rule, we start with the differentiability condition for f .

$$f(\mathbf{r} + \Delta\mathbf{r}) = f(\mathbf{r}) + \nabla f(\mathbf{r}) \cdot \Delta\mathbf{r} + e(\mathbf{r}, \Delta\mathbf{r})$$

where $e/|\Delta\mathbf{r}| \rightarrow 0$ as $|\Delta\mathbf{r}| \rightarrow 0$. Hence,

$$\Delta w = f(\mathbf{r} + \Delta\mathbf{r}) - f(\mathbf{r}) = \nabla f(\mathbf{r}) \cdot \Delta\mathbf{r} + e(\mathbf{r}, \Delta\mathbf{r}), \quad (36)$$

and dividing by Δt yields

$$\frac{\Delta w}{\Delta t} = \nabla f(\mathbf{r}) \cdot \frac{\Delta\mathbf{r}}{\Delta t} + \frac{e}{\Delta t}.$$

Now let $\Delta t \rightarrow 0$. On the left, the limit is $dw/dt = dh/dt$. On the right, we have

$$\nabla f \cdot \lim_{\Delta t \rightarrow 0} \frac{\Delta\mathbf{r}}{\Delta t} = \nabla f \cdot \frac{d\mathbf{r}}{dt}$$

which is what we want, so the rest of the argument amounts to showing that the additional term $e/\Delta t$ goes to 0 as $\Delta t \rightarrow 0$.

We need to distinguish two cases. For a given Δt , we may have $\Delta \mathbf{r} = \mathbf{0}$. In that case, $\Delta w = 0$, and it also follows from (36) that $e = 0$. Otherwise, if $\Delta \mathbf{r} \neq \mathbf{0}$, write

$$\frac{e}{|\Delta t|} = \frac{e}{|\Delta \mathbf{r}|} \frac{|\Delta \mathbf{r}|}{|\Delta t|}. \quad (37)$$

Now, let $\Delta t \rightarrow 0$, but first restrict attention just to those Δt for which $\Delta \mathbf{r} = \mathbf{0}$. For those Δt , we have, as noted above, $e = 0$, so we have trivially $e/\Delta t \rightarrow 0$. Next, let $\Delta t \rightarrow 0$, but restrict attention instead just to those Δt for which $\Delta \mathbf{r} \neq \mathbf{0}$. Since $|\Delta \mathbf{r}/\Delta t| \rightarrow |d\mathbf{r}/dt|$, it follows that $|\Delta \mathbf{r}| \rightarrow 0$ in these cases. By assumption $e/|\Delta \mathbf{r}| \rightarrow 0$ generally, so the same is true if we restrict attention to those $|\Delta \mathbf{r}| \neq 0$ obtained from non-zero Δt going to zero. Thus, equation (37) tells us that $e/|\Delta t| \rightarrow 0$ also in the second case.

It should be noted that this is exactly the same as the usual proof of the single variable chain rule except for modifications necessary due to the fact that some of the arguments are vectors rather than scalars. \square

Geometric Interpretation of the Chain Rule For variety, we consider the case of a function $f : \mathbf{R}^3 \rightarrow \mathbf{R}$ composed with a function $\mathbf{g} : \mathbf{R} \rightarrow \mathbf{R}^3$. You might think of the function $f(\mathbf{r})$ giving the temperature w at the point with position vector \mathbf{r} . Then the *level surfaces* of the function would be called isotherms, surfaces of constant temperature. As before $\mathbf{r} = \mathbf{g}(t)$ would be a parametric representation of a curve in \mathbf{R}^3 and we could think of it as the path of a particle moving in space. The derivative $d\mathbf{r}/dt = \mathbf{g}'(t)$ would be the velocity vector at time t . Then the chain rule could be written

$$\frac{dw}{dt} = \nabla f(\mathbf{r}) \cdot \frac{d\mathbf{r}}{dt} \quad \mathbf{r} = \mathbf{g}(t),$$

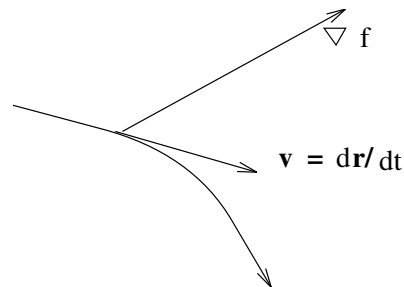
and it would say that the rate of change of temperature w experienced by the particle as it moves through the point with position vector \mathbf{r} would be the gradient of the temperature function f at \mathbf{r} dotted with the velocity vector at that point of the curve. Of course, you could think of the function w as giving any other quantity which might be interesting in the problem you are studying.

Note that the formula for the directional derivative

$$\frac{df}{ds} = \nabla f(\mathbf{r}) \cdot \mathbf{u}$$

is a *special case* of the chain rule. Namely, if the particle moves in such a way that the speed $ds/dt = 1$, then the velocity vector $\mathbf{u} = d\mathbf{r}/dt$ will be a unit vector, and we may identify s with t , i.e., “distance” with “time”.

One important consequence of the chain rule is that at any point \mathbf{r} , the *gradient* $\nabla f(\mathbf{r})$ (provided it is not zero) is *perpendicular to the level surface through \mathbf{r}* . For,



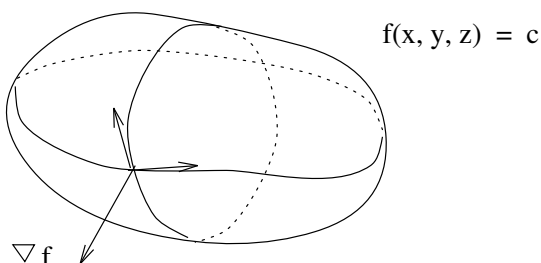
suppose a curve given by $\mathbf{r} = \mathbf{g}(t)$ is contained in the level surface

$$f(\mathbf{r}) = c.$$

Since the derivative of a constant is zero, the chain rule tells us

$$\nabla f \cdot \frac{d\mathbf{r}}{dt} = 0$$

so ∇f is perpendicular to $d\mathbf{r}/dt$. On the other hand, $d\mathbf{r}/dt$ is tangent to the curve, so it is also tangent to the surface. For any reasonable surface, we can manage to get *every possible tangent vector* to the surface by a suitable choice of a curve lying in the surface. Hence, it follows that the gradient is perpendicular to the surface.

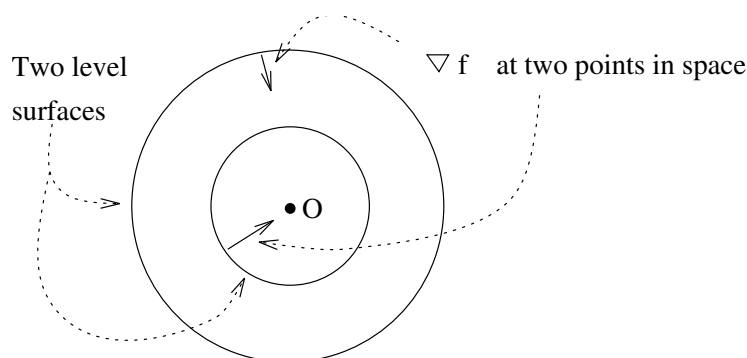


(The realizability of all tangents to level surface is a bit subtle, and we shall study it more closely in a later section. For now, it suffices to say that for reasonable surfaces, there is no problem, and the gradient is always normal (perpendicular) to the level surface.)

Example 39 Let $f(\mathbf{r}) = 1/|\mathbf{r}| = 1/\sqrt{x^2 + y^2 + z^2}$. The level surfaces of this function are spheres centered at the origin. On the other hand,

$$\nabla f = \left\langle \frac{-x}{(x^2 + y^2 + z^2)^{3/2}}, \frac{-y}{(x^2 + y^2 + z^2)^{3/2}}, \frac{-z}{(x^2 + y^2 + z^2)^{3/2}} \right\rangle = -\frac{1}{|\mathbf{r}|^3} \mathbf{r}.$$

This vector points toward the origin, so it is perpendicular to a sphere centered at the origin.



Notice that in this example $|\nabla f| = |\mathbf{r}|/|\mathbf{r}|^3 = 1/|\mathbf{r}|^2$, so it satisfies the *inverse square law*. Except for a constant, the gravitational force due to a point mass at the origin is given by this rule. The function f in this case is the gravitational *potential energy function*. It is true for many forces (e.g., gravitational or electrostatic) that the gradient of the potential energy function is the force.

If we are working in \mathbf{R}^2 rather than \mathbf{R}^3 , then, $f(\mathbf{r}) = f(x, y) = c$ defines a family of *level curves* rather than level surface. As above, the gradient ∇f is generally perpendicular to these level curves.

Other Forms of the Chain Rule The most general chain rule tells us how to find derivatives of composites of functions $\mathbf{R}^n \rightarrow \mathbf{R}^m$ for appropriate combinations of m and n . We consider one special cases here. You will see how to generalize easily to other cases. Let $w = f(x, y)$ describe a scalar valued function of two variables. ($f : \mathbf{R}^2 \rightarrow \mathbf{R}$.) Suppose, in addition, $x = x(s, t)$ and $y = y(s, t)$ describe two functions of two variables s, t . (As we shall see later, this amounts to a single function $\mathbf{R}^2 \rightarrow \mathbf{R}^2$.) Then we may consider the composite function described by

$$w = h(s, t) = f(x(s, t), y(s, t)).$$

The chain rule generates formulas for $\partial h/\partial s$ and $\partial h/\partial t$ as follows. Partial derivatives are ordinary derivatives computed with the assumption that all the variables but one vary. Hence, to compute $\partial h/\partial t$, all we need to do is to use (34), replacing dh/dt by $\partial h/\partial t$, dx/dt by $\partial x/\partial t$, and dy/dt by $\partial y/\partial t$. Similarly, for $\partial h/\partial s$. We get

$$\begin{aligned}\frac{\partial h}{\partial s} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial s} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial s} \\ \frac{\partial h}{\partial t} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial t} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial t}.\end{aligned}\tag{38}$$

Note in these formulas that the differentiation variable on the left (s or t) must agree with the ultimate differentiation variable on the right.

Example 40 Let $f(x, y) = \sqrt{x^2 + y^2}$, $x = r \cos \theta$, $y = r \sin \theta$. Here, the intermediate variables are x, y and the ultimate independent variables are r, θ . We have

$$\begin{aligned}\frac{\partial f}{\partial x} &= \frac{x}{\sqrt{x^2 + y^2}} = \frac{r \cos \theta}{r} = \cos \theta \\ \frac{\partial f}{\partial y} &= \frac{y}{\sqrt{x^2 + y^2}} = \frac{r \sin \theta}{r} = \sin \theta.\end{aligned}$$

Also,

$$\begin{aligned}\frac{\partial x}{\partial r} &= \cos \theta & \frac{\partial y}{\partial r} &= \sin \theta \\ \frac{\partial x}{\partial \theta} &= -r \sin \theta & \frac{\partial y}{\partial \theta} &= r \cos \theta.\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\partial h}{\partial r} &= (\cos \theta)(\cos \theta) + (\sin \theta)(\sin \theta) = 1 \\ \frac{\partial h}{\partial \theta} &= (\cos \theta)(-r \sin \theta) + (\sin \theta)(r \cos \theta) = 0.\end{aligned}$$

Note that one must substitute for x, y in terms of r, θ in the expressions for $\partial f/\partial x$ and $\partial f/\partial y$ or one won't get a complete answer.

The simplicity of the answers is more easily seen if we do the substitution before differentiating.

$$h(r, \theta) = \sqrt{r^2 \cos^2 \theta + r^2 \sin^2 \theta} = r.$$

Hence, $\partial h/\partial r = 1$ and $\partial h/\partial \theta = 0$. Again, this illustrates the point that the multidimensional chain rule is not primarily a computation device. However, we shall see that its utility in theoretical discussions in mathematics *and in applications* more than justifies its use.

A Confusing Point

The most difficult use of the chain rule is in situations like the following. Suppose $w = f(x, y, z)$, $z = z(x, y)$, and we want the partial derivatives of $w = h(x, y) = f(x, y, z(x, y))$ with respect to x and y . The correct formulas are

$$\begin{aligned}\frac{\partial h}{\partial x} &= \frac{\partial f}{\partial x} + \frac{\partial f}{\partial z} \frac{\partial z}{\partial x} \\ \frac{\partial h}{\partial y} &= \frac{\partial f}{\partial y} + \frac{\partial f}{\partial z} \frac{\partial z}{\partial y}.\end{aligned}$$

One way to see the truth of these formulas is as follows. Suppose we introduce new variables s and t with $x = x(s, t) = s$, $y = y(s, t) = t$, and $z = z(s, t)$. Let $w = h(s, t) = f(x(s, t), y(s, t), z(s, t)) = f(s, t, z(s, t))$. Then, according to the chain rule

$$\begin{aligned}\frac{\partial h}{\partial s} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial s} + \frac{\partial f}{\partial z} \frac{\partial z}{\partial s} \\ &= \frac{\partial f}{\partial x}(1) + \frac{\partial f}{\partial z} \frac{\partial z}{\partial s}.\end{aligned}$$

A similar argument works for $\frac{\partial h}{\partial t}$. We may now obtain the previous formulas by identifying x with s and y with t .

The source of the problem is confusing the *variables* with functions. Thus, writing $z = z(x, y)$ lacks precision since the name ' z ' should not be used both for the variable and the function expressing it in terms of other variables. However, this is a common 'abuse of notation' in applications, since it allows us to concentrate on the physical interpretation of the variables which might otherwise be obscured by a more precise mathematical formalism.

To see how confusing all this can be, let's consider a thermodynamic application. The entropy s is a function $s = s(p, v, T)$ of pressure p , volume v , and temperature T . Also, the pressure may be assumed to be a function $p = p(v, T)$. Thus, ultimately, we may express the entropy $s = s(v, T)$. Note the several abuses of notation. $s(p, v, T)$ and $s(v, T)$ refer of course to *different* functions. With this notation, the above formulas give

$$\begin{aligned}\frac{\partial s}{\partial v} &= \frac{\partial s}{\partial v} + \frac{\partial s}{\partial p} \frac{\partial p}{\partial v} \\ \frac{\partial s}{\partial T} &= \frac{\partial s}{\partial T} + \frac{\partial s}{\partial p} \frac{\partial p}{\partial T}\end{aligned}$$

which suggests that we may cancel the common terms on both sides to conclude that the remaining term is zero. This is not correct, since, as just mentioned, the two functions ' s ' begin differentiated are different. To clarify this, one should write the formulas

$$\begin{aligned}\left(\frac{\partial s}{\partial v}\right)_T &= \left(\frac{\partial s}{\partial v}\right)_{p,T} + \left(\frac{\partial s}{\partial p}\right)_{v,T} \left(\frac{\partial p}{\partial v}\right)_T \\ \left(\frac{\partial s}{\partial T}\right)_v &= \left(\frac{\partial s}{\partial T}\right)_{v,T} + \left(\frac{\partial s}{\partial p}\right)_{v,T} \left(\frac{\partial p}{\partial T}\right)_v.\end{aligned}$$

Here, the additional subscripts tell us which variables are kept constant, so by implication we may see which variables the given quantity is supposed to be a function of. **Gradient in Polar Coordinates** As you have seen, it is sometimes

useful to resolve vectors in the plane in terms of the polar unit vectors \mathbf{u}_r and \mathbf{u}_θ . We want to do this for the gradient of a function f given initially by $w = f(\mathbf{r}) = f(x, y)$. Suppose

$$\nabla f = A_r \mathbf{u}_r + A_\theta \mathbf{u}_\theta. \quad (39)$$

For a particle moving on a curve in the plane, the chain rule tells us

$$\frac{dw}{dt} = \nabla f \cdot \frac{d\mathbf{r}}{dt}. \quad (40)$$

On the other hand,

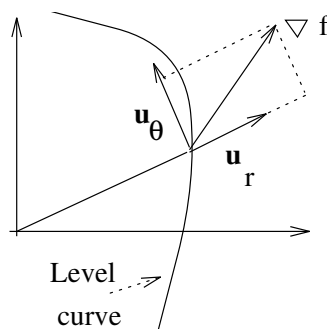
$$\frac{d\mathbf{r}}{dt} = \frac{dr}{dt} \mathbf{u}_r + r \frac{d\theta}{dt} \mathbf{u}_\theta. \quad (41)$$

Putting (39) and (41) in (40) yields

$$\frac{dw}{dt} = (A_r \mathbf{u}_r + A_\theta \mathbf{u}_\theta) \cdot \left(\frac{dr}{dt} \mathbf{u}_r + r \frac{d\theta}{dt} \mathbf{u}_\theta\right) = A_r \frac{dr}{dt} + A_\theta r \frac{d\theta}{dt}. \quad (42)$$

On the other hand, we can write $w = g(r, \theta) = f(x, y) = f(r \cos \theta, r \sin \theta)$, so applying the chain rule directly to g , gives

$$\frac{dw}{dt} = \frac{\partial g}{\partial r} \frac{dr}{dt} + \frac{\partial g}{\partial \theta} \frac{d\theta}{dt}. \quad (43)$$



We now argue that since the curve could be anything, so could dr/dt and $d\theta/dt$. Hence, the coefficients of these two quantities in (42) and (43) must be the same. Hence, $A_r = \partial g/\partial r$ and $A_\theta = \partial g/\partial \theta$, i.e., $A_\theta = (1/r)\partial g/\partial \theta$. It follows that

$$\nabla f = \frac{\partial g}{\partial r} \mathbf{u}_r + \frac{1}{r} \frac{\partial g}{\partial \theta} \mathbf{u}_\theta$$

where g is the function expressing the desired quantity in polar coordinates.

Example 39, revisited Let $f(x, y) = 1/\sqrt{x^2 + y^2} = 1/r = g(r, \theta)$. Notice that this is what we would get if we restricted Example 39 to the x, y -plane by setting $z = 0$. Then $\partial g/\partial r = -(1/r^2)$ and $\partial g/\partial \theta = 0$. Hence,

$$\nabla f = -\frac{1}{r^2} \mathbf{u}_r.$$

This is the same answer we got previously except that we are restricting attention to the x, y -plane.

A similar analysis can be done in space if one uses the appropriate generalization of polar coordinates, in this case what are called *spherical coordinates*. We shall return to this later in the course.

Exercises for 3.7.

In the following problems, use the appropriate form of the chain rule. The rule you need may not have been stated explicitly in the text, but you ought to be able to figure out what it is. The notation is typical of what is used in practice and is not always completely precise.

- Find dw/dt by first using the chain rule, then by determining $w(t)$ explicitly before differentiating. Check that the answers are the same by expressing them both entirely in terms of t .
 - $w = x^2 + y^2$, $x = t$, $y = -2t$.
 - $w = \frac{1}{x+y}$, $x = \cos t$, $y = \sin t$.
 - $w = \ln(x^2 + y^2 + z^2)$, $x = 2 - 3t$, $y = \sqrt{t}$, $z = t$.
- Use the chain rule to find $\partial h/\partial x$, $\partial h/\partial y$, and $\partial h/\partial z$.
 - $h = e^{-2u-3v+w}$, $u = xz$, $v = yz$, $w = xy$.
 - $h = \sqrt{u^2 + v^2 + w^2}$, $u = x + y + z$, $v = x^2$, $w = y^2$.
- Suppose $w = w(x, y)$ and $y = y(x)$. Hence, ultimately $w = w(x)$. With this abuse of notation, derive a correct formula for $\frac{dw}{dx}$. (Hint: You might try writing the above more precisely $w = f(x, y)$, $y = g(x)$, and $w = h(x) = f(x, g(x))$.)

4. In each case, use the chain rule to find $\partial h/\partial x$ and $\partial h/\partial y$ in terms of x and y for the given composite function $h(x, y)$. Then express h explicitly in terms of x and y , and find those partials again. Check that you get the same thing.
- (a) $h = u^2 + v^2 + x^2 + y^2$, $u = 2x - 3y$, $v = 2x + 3y$.
- (b) $h = uvwxy$, $u = x^2 + y^2$, $v = 5x - 6y$, $w = xy$.
5. In each case find a normal vector to the indicated level set at the indicated point.
- (a) $x^2 + 3y^2 = 7$ at $(2, -1)$.
- (b) $x^2 + y^2 - z^2 = 1$ at $(2, 1, 2)$.
- (c) $x^3 - x + y^2 - z = 0$ at $(2, -3, 15)$.
6. Find a normal vector to the level curve defined by $F(x, y) = f(x) - y = 0$ at a general point. Show that it is perpendicular to a tangent vector to the graph $y = f(x)$. Hint: What is the product of the slopes of the tangent line and the normal line?
7. The temperature on a heated plate is given by the formula $T = T(x, y) = x^2 + xy + y^2$. A psychologist induces a bug to follow the circular path given by $\mathbf{r} = 3 \cos 2t \mathbf{i} + 3 \sin 2t \mathbf{j}$. Find the rate of change of temperature experienced by the bug at $t = \pi/2$.
8. In a meteorological theory, it is assumed that pressure is a function only of z , $p = p(z)$. A rocket with remote sensing equipment is launched in a parabolic path. It is intuitively clear that at the top of the path, it will report that the rate of change $dp/dt = 0$. Verify this conclusion mathematically. Could you draw the same conclusion if $p = p(x, z)$ and the path is in the y, z -plane? the x, z -plane?
9. In each case express ∇f in polar coordinates in terms of \mathbf{u}_r and \mathbf{u}_θ .
- (a) $f(x, y) = x$.
- (b) $f(x, y) = y$.
- (c) $f(x, y) = x^2 + y^2 = r^2$.
10. Suppose $f(x, y) = \tan^{-1} \frac{y}{x}$ for $x, y > 0$. Calculate ∇f in rectangular coordinates. Also, express $f(x, y) = g(r, \theta)$ and use the formula $\nabla f = (\partial g/\partial r)\mathbf{u}_r + (1/r)(\partial g/\partial \theta)\mathbf{u}_\theta$. Can you convince yourself the two answers are the same?

3.8 Tangents to Level Sets and Implicit Differentiation

We saw in the previous section that for a function $f : \mathbf{R}^2 \rightarrow \mathbf{R}$ of two variables, the gradient $\nabla f(\mathbf{r}_0)$ is perpendicular to the level curve $f(\mathbf{r}) = c$ through \mathbf{r}_0 , and similarly for functions $f : \mathbf{R}^3 \rightarrow \mathbf{R}$ except that the locus is a level *surface* instead.

Example 41 Consider the locus in \mathbf{R}^2 of the equation

$$x^3 + 3xy^2 + y = 15$$

at the point $(1, 2)$. The normal vector is $\nabla f(1, 2)$ for $f(x, y) = x^3 + 3xy^2 + y$. That is, the normal vector is

$$\langle 3x^2 + 3y^2, 6xy + 1 \rangle = \langle 15, 13 \rangle.$$

The tangent line will be characterized by the relation

$$\nabla f(\mathbf{r}_0) \cdot (\mathbf{r} - \mathbf{r}_0) = 0$$

which in this case becomes

$$\langle 15, 13 \rangle \cdot \langle x - 1, y - 2 \rangle = 15(x - 1) + 13(y - 2) = 0.$$

Simplifying, this gives

$$15x + 13y = 41.$$

Example 42 Consider the hyperboloid of one sheet with equation

$$x^2 + 2y^2 - z^2 = 2.$$

at $(1, 1, 1)$. The normal vector is $\nabla f(1, 1, 1)$ where $f(x, y, z) = x^2 + 2y^2 - z^2$. Thus, it is

$$\langle 2x, 4y, -2z \rangle = \langle 2, 4, -2 \rangle.$$

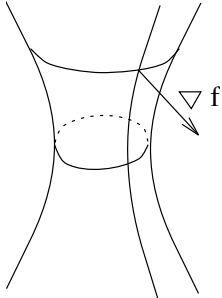
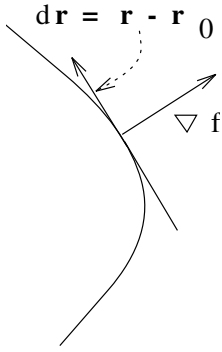
The tangent plane will be characterized by the equation $\nabla f(\mathbf{r}_0) \cdot (\mathbf{r} - \mathbf{r}_0) = 0$ which in this case becomes

$$2(x - 1) + 4(y - 1) - 2(z - 1) = 0$$

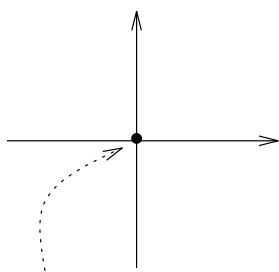
or

$$2x + 4y - 2z = 2.$$

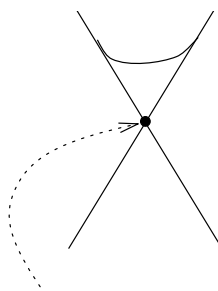
There is a subtle point involved here. What justification do we have for believing that an equation of the form $f(x, y) = c$ defines what we can honestly call a curve in \mathbf{R}^2 , and what justification is there for calling the line above a tangent? Similar questions can be asked in \mathbf{R}^3 about the locus of $f(x, y, z) = c$ and the corresponding plane. In fact, there are simple examples of where this would not be a reasonable



use of language. For example, the locus in \mathbf{R}^2 of $x^2 + y^2 = 0$ is the *point* $(0, 0)$, and it certainly is not a curve in any ordinary sense. Similarly, we saw previously that the locus in \mathbf{R}^3 of $x^2 + y^2 - z^2 = 0$ is a (double) cone, which looks like a surface all right, but it does not have a well defined tangent plane at the origin. If you look carefully at both these examples, you will see what went wrong; in each case the gradient ∇f vanishes at the point under consideration. It turns out that if \mathbf{r}_0 is a point satisfying the equation $f(\mathbf{r}) = c$ and $\nabla f(\mathbf{r}_0) \neq 0$, then the level set looks like what it should look like and the locus of $\nabla f(\mathbf{r}_0) \cdot (\mathbf{r} - \mathbf{r}_0) = 0$ makes sense as a tangent (line or plane) to that level set.

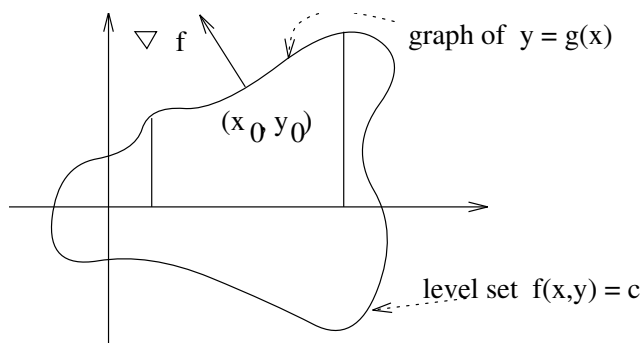


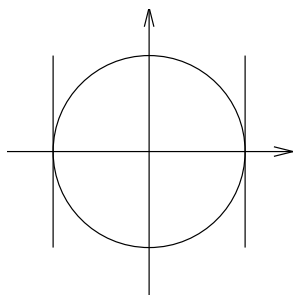
Graph of $x^2 + y^2 = 0$ is a point.



Tangent plane not well defined

The basis for all of this is a deep theorem called the *implicit function theorem*. We won't try to prove this theorem here, or even to state it very precisely, but we indicate roughly what it has to do with the above discussion. Consider first the case of a level curve $f(x, y) = c$ in \mathbf{R}^2 . Let (x_0, y_0) be a point on that curve, where $\nabla f(x_0, y_0) \neq \mathbf{0}$. The implicit function theorem says that (subject to reasonable smoothness assumptions on f) the curve is identical with the graph of some function g of one variable, at least if we stay close enough to the point (x_0, y_0) . Moreover, the tangent line to the graph is the same as the line obtained from the equation $\nabla f(\mathbf{r}_0) \cdot (\mathbf{r} - \mathbf{r}_0) = 0$.





Example 43 The locus of $f(x, y) = x^2 + y^2 = 1$ is a circle of radius 1 centered at the origin. The gradient $\nabla f = \langle 2x, 2y \rangle$ does not vanish at any point on the circle. If we fix a point (x_0, y_0) with $y_0 > 0$ (top semicircle), then in the vicinity of that point, the circle can be identified with the graph of the function $y = g(x) = \sqrt{1 - x^2}$. If $y_0 < 0$, we have to use $y = g(x) = -\sqrt{1 - x^2}$ instead. It is not hard to check that the tangent line is the same whichever method we use to find it.

There is one problem. Namely, at the points $(1, 0)$ and $(-1, 0)$ the tangent lines are vertical, so neither can be obtained as a tangent to the graph of a function given by $y = g(x)$. (The slopes would be infinite.) Instead, we have to reverse the roles of x and y and use the graph of $x = g(y) = \sqrt{1 - y^2}$ near the point $(1, 0)$ or $x = g(y) = -\sqrt{1 - y^2}$ near the point $(-1, 0)$.

Similar remarks apply to level surfaces $f(x, y, z) = c$ in \mathbf{R}^3 except that the conclusion is that in the neighborhood of a point at which ∇f does not vanish the level set can be identified with the graph of a function g of two variables. In particular, it really looks like a surface, and the equations of the tangent plane work out right. (If the normal vector ∇f points in a horizontal direction, i.e., $f_z = 0$, you may have to try $x = g(y, z)$ or $y = g(x, z)$ rather than $z = g(x, y)$.)

Having noted that level sets can be thought of (at least locally) as graphs of functions, we should note that the reverse is also true. For example, suppose $f : \mathbf{R}^2 \rightarrow \mathbf{R}$ is a function of two variables. Its graph is the locus of $z = f(x, y)$. If we define $F(x, y, z) = z - f(x, y)$, the graph is the zero level set

$$F(x, y, z) = z - f(x, y) = 0.$$

This remark allows us to treat the theory of tangent planes to graphs as a special case of the theory of tangent planes to level sets. The normal vector is $\nabla F = \langle -f_x, -f_y, 1 \rangle$ and the equation of the tangent plane is $\langle -f_x(\mathbf{r}_0), -f_y(\mathbf{r}_0), 1 \rangle \cdot \langle x - x_0, y - y_0, z - z_0 \rangle = 0$.

Example 44 Consider the graph of the function $z = xy$ to be the level set of the function

$$F(x, y, z) = z - xy = 0.$$

At $(0, 0, 0)$, $\nabla F = \langle -y, -x, 1 \rangle = \langle 0, 0, 1 \rangle$. Hence, the equation of the tangent plane is

$$0(x - 0) + 0(y - 0) - 1(z - 0) = 0$$

or

$$z = 0.$$

Thus, the tangent plane to the graph (which is a hyperbolic paraboloid) at $(0, 0, 0)$ is the x, y -plane. In particular, note that it actually intersects the surface in *two lines*, the x and y axes.

Implicit Differentiation The above discussion casts some light on the problem of “implicit differentiation”. You recall that this is a method for finding dy/dx in

a situation in which y may not be given explicitly in terms of x , but rather it is assumed there is a functional relation $y = y(x)$ which is consistent with a relation

$$f(x, y) = c$$

between y and x .

Example 45 Suppose $x^2 + y^2 = 1$ and find dy/dx . To solve this we differentiate the relation using the usual rules to obtain

$$2x + 2y \frac{dy}{dx} = 0$$

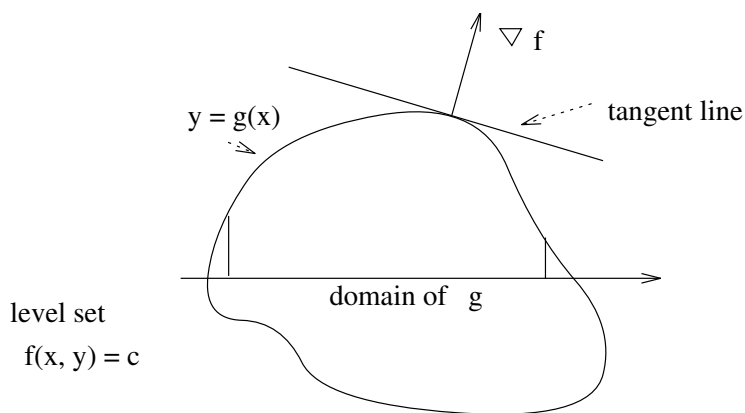
which can be solved

$$\frac{dy}{dx} = -\frac{x}{y}.$$

Note that the answer depends on both x and y , so for example it would be different for $y > 0$ (the top semi-circle) and $y < 0$ (the bottom semi-circle). This is consistent with the fact that in order to pick out a *unique* functional relationship $y = y(x)$, we must specify where on the circle we are. In addition, the method does not make sense if $y = 0$, i.e., at the points $(\pm 1, 0)$. As we saw above, it would be more appropriate to express $x = x(y)$ as a function of y in the vicinity of those points.

The process of implicit differentiation can be explained in terms of the tangent line to a level curve as follows. First, rewrite the equation of the tangent to $f(x, y) = c$ in differential form

$$\nabla f \cdot d\mathbf{r} = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy = 0. \quad (44)$$



This is just a change of notation where we drop the subscript 0 in describing the point on the curve and we use $d\mathbf{r}$ rather than $\mathbf{r} - \mathbf{r}_0$ for the displacement along

the tangent line. (We usually think of $d\mathbf{r}$ as being very small so we don't have to distinguish between the displacement in the tangent direction and the displacement along the curve, at least if we are willing to tolerate a very small error.) (44) can be solved for dy by writing

$$dy = -\frac{\partial f/\partial x}{\partial f/\partial y} dx$$

provided the denominator $f_y \neq 0$. On the other hand, the implicit function theorem tells us that near the point of tangency we may assume that the graph is the graph of a function given by $y = g(x)$, and the tangent line to that graph would be expressed in differential notation

$$dy = g'(x)dx = \frac{dy}{dx}dx.$$

Since it is the same tangent line in either case, we conclude

$$\frac{dy}{dx} = -\frac{\partial f/\partial x}{\partial f/\partial y}.$$

(Note that the assumption $f_y \neq 0$ at the point of tangency plays an additional role here. That condition insures that the normal vector ∇f won't point horizontally, i.e., that the tangent line is not vertical. Hence, it makes sense to consider the tangent as a tangent line to the graph of a function $y = g(x)$.)

Example 45, revisited For

$$f(x, y) = x^2 + y^2 = 1,$$

we have $f_x(x, y) = 2x$, $f_y(x, y) = 2y$, so (45) tells us $dy/dx = -x/y$, just as before.

Similar reasoning applies to level surfaces in \mathbf{R}^3 . Let

$$f(x, y, z) = c$$

define such a level surface, and write the equation of the tangent plane in differential notation

$$\nabla f \cdot d\mathbf{r} = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial z}dz = 0. \quad (46)$$

Solving for dz yields

$$dz = -\frac{\partial f/\partial x}{\partial f/\partial z}dx - \frac{\partial f/\partial y}{\partial f/\partial z}dy$$

provided the denominator $f_z \neq 0$. On the other hand, at such a point, the implicit function theorem allows us to identify the surface near the point with the graph of a function $z = g(x, y)$. Moreover, we may write the equation of the tangent plane to the graph in differential form as

$$dz = \frac{\partial g}{\partial x}dx + \frac{\partial g}{\partial y}dy.$$

Since it is the same tangent plane in either case, we conclude that

$$\begin{aligned}\frac{\partial g}{\partial x} &= -\frac{\partial f/\partial x}{\partial f/\partial z} \\ \frac{\partial g}{\partial y} &= -\frac{\partial f/\partial y}{\partial f/\partial z}.\end{aligned}$$

Example 46 Suppose $xyz + xz^2 + z^3 = 1$. Assuming $z = g(x, y)$, find $\partial z/\partial x$ and $\partial z/\partial y$ in general and also for $x = 1, y = -1, z = 1$. To solve this, we put $f(x, y, z) = xyz + xz^2 + z^3$ and set its total differential to zero

$$df = f_x dx + f_y dy + f_z dz = (yz + z^2)dx + (xz)dy + (xy + 2xz + 3z^2)dz = 0.$$

This can be rewritten

$$dz = -\frac{yz + z^2}{xy + 2xz + 3z^2}dx - \frac{xz}{xy + 2xz + 3z^2}dy$$

so

$$\begin{aligned}\frac{\partial z}{\partial x} &= -\frac{yz + z^2}{xy + 2xz + 3z^2} \\ \frac{\partial z}{\partial y} &= -\frac{xz}{xy + 2xz + 3z^2}\end{aligned}$$

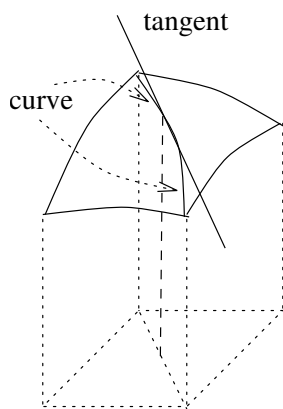
provided the denominator does not vanish. At $(1, -1, 1)$ we have $\partial z/\partial x = 0/4 = 0$ and $\partial z/\partial y = 1/4$.

There is another way this problem could be done which parallels the approach you learned in your previous course. For example, to find $\partial z/\partial x$, just apply the “operator” $\partial/\partial x$ to the equation $xyz + xz^2 + z^3 = 1$ under the assumption that $z = g(x, y)$ but x, y are independent. We get

$$yz + xy\frac{\partial z}{\partial x} + z^2 + 2xz\frac{\partial z}{\partial x} + 3z^2\frac{\partial z}{\partial x} = 0.$$

This equation can be solved for $\partial z/\partial x$ and we get the same answer as above. (Check it!)

A Theoretical Point Consider the level surface $f(\mathbf{r}) = c$ at a point where the gradient ∇f does not vanish. In showing that the gradient is perpendicular to the level surface (in the previous section), we claimed that every possible tangent vector to the surface at the point of tangency is in fact tangent to some curve lying in the surface. A moment’s thought shows there are some problems with this assertion. First, how do we know that the level set looks like a surface, (as opposed say to a point) and if it does, what do we mean by its tangent plane? These issues are settled by using the implicit function theorem. For as long as ∇f does not vanish at the point, that theorem allows us to view the surface in the vicinity of the given



point as the graph of some function. However, the graph of a function certainly has the right properties for our concept of “surface”. Moreover, there is no problem with tangent vectors to a graph, and any such vector determines a *plane section perpendicular to the domain of the function* which intersects the graph in a curve with the desired tangent. (See the diagram.)

Exercises for 3.8.

- Find the equation $\nabla f(\mathbf{r}_0) \cdot (\mathbf{r} - \mathbf{r}_0) = 0$ for each of the following level surfaces.
 - $2x^2 + 3y + z^3 = 5$, $P(3, -7, 2)$.
 - $x^4 + 3xy^2z + xy^2z^2 + z^3 = 0$, $P(0, -2, 0)$
- Show that every tangent plane of the cone $z^2 = x^2 + y^2$ passes through the origin. (See also Exercise (4) in Section 3.)
- Find all point(s) where the tangent plane to the surface $z = x^2 + 4xy + 4y^2 - 12x + 4y$ is horizontal.
- In each of the following, find $\partial z / \partial x$ and $\partial z / \partial y$ in terms of x, y , and z if you assume $z = f(x, y)$. Do not try to solve explicitly for z .
 - $x^{\frac{5}{2}} + y^{\frac{5}{2}} + z^{\frac{5}{2}} = 1$
 - $2e^x z + 4e^y z - 6e^x y = 17$
 - $xyz = 2x + 2y + 2z$
 - $4x^2 y^2 z + 5xyz^3 - 2x^3 y = -2$
- Suppose that $f(x, y, z) = c$ expresses a relation among the three variables x, y , and z . We may suppose that this defines any one of the variables as a function of the other two. Show that

$$\left(\frac{\partial x}{\partial y} \right)_z \left(\frac{\partial y}{\partial z} \right)_x \left(\frac{\partial z}{\partial x} \right)_y = -1.$$

Hint: You should be able to express each of the indicated derivatives in terms of f_x, f_y , and f_z .

- Each of the equations $g(x, y, z) = c$ and $h(x, y, z) = d$ defines a surface in \mathbf{R}^3 . The intersection of the two surfaces is usually a curve in \mathbf{R}^3 . At any given point on this curve, ∇g is perpendicular to the first surface and ∇h is perpendicular to the second surface, so both are perpendicular to the curve.
 - Use this information to find a vector *tangent* to the curve at the point. (b) Apply this analysis to find a tangent vector to the intersection of the cylinder $x^2 + y^2 = 25$ with the plane $x + y - z = 0$ at the point $(3, 4, 5)$.

3.9 Higher Partial Derivatives

Just as in the single variable case, you can continue taking derivatives in the multidimensional case. However, because there is more than one independent variable, the situation is a bit more complicated.

Example 47 Let $f(x, y) = x^2y + y^2$. Then

$$\frac{\partial f}{\partial x} = 2xy \quad \frac{\partial f}{\partial y} = x^2 + 2y.$$

Each of these can be differentiated with respect to either x or y and the results are denoted

$$\begin{aligned} \frac{\partial^2 f}{\partial x^2} &= \frac{\partial}{\partial x} \left(\frac{\partial f}{\partial x} \right) = 2y & \frac{\partial^2 f}{\partial x \partial y} &= \frac{\partial}{\partial x} \left(\frac{\partial f}{\partial y} \right) = 2x \\ \frac{\partial^2 f}{\partial y \partial x} &= \frac{\partial}{\partial y} \left(\frac{\partial f}{\partial x} \right) = 2x & \frac{\partial^2 f}{\partial y^2} &= \frac{\partial}{\partial y} \left(\frac{\partial f}{\partial y} \right) = 2. \end{aligned}$$

All these are called *second order partial derivatives*. Note that the order in which the operations are performed is from right to left; the operation closer to the function is performed first. $\frac{\partial^2 f}{\partial x \partial y}$ and $\frac{\partial^2 f}{\partial y \partial x}$ are called *mixed partials*. Other notation for partial derivatives is f_{xx} , f_{xy} , f_{yx} , and f_{yy} . However, just to make life difficult for you, the order is different. For example, f_{xy} means first differentiate with respect to x and then differentiate the result with respect to y .

Fortunately, it doesn't usually make any difference which order you do the operations. For example, in Example 47, we have

$$\begin{aligned} \frac{\partial^2 f}{\partial x \partial y} &= \frac{\partial}{\partial x} \frac{\partial f}{\partial y} = 2x \\ \frac{\partial^2 f}{\partial y \partial x} &= \frac{\partial}{\partial y} \frac{\partial f}{\partial x} = 2x \end{aligned}$$

so the mixed partials are equal. The following theorem gives us conditions under which we can be sure such is the case. We state it for functions of two variables, but its analogue holds for functions of any number of variables.

Theorem 3.3 Let $z = f(x, y)$ denote a function of two variables defined on some open set in \mathbf{R}^2 . Assume the partial derivatives f_x , f_y , and f_{xy} are defined and f_{xy} is continuous on that set. Then f_{yx} exists and $f_{xy}(x, y) = f_{yx}(x, y)$.

A function with continuous second order partial derivatives is usually called \mathcal{C}^2 . This is more stringent than the condition of being \mathcal{C}^1 (having continuous first order partials). It will almost always be true that functions you have to deal with in applications are \mathcal{C}^2 except possibly for an isolated set of points.

Clearly, we can continue this game ad infinitum. There are 8 possible 3rd order derivatives for a function of two variables:

$$f_{xxx}, f_{yxx}, f_{xyx}, f_{xxy}, f_{xyy}, f_{yxy}, f_{yyx}, f_{yyy}.$$

(These could also be denoted $\partial^3 f / \partial x^3$, $\partial^3 f / \partial x^2 \partial y$, etc.) However, for sufficiently smooth functions, the 2nd, 3rd, and 4th are the same, as are the 5th, 6th, and 7th.

How many second order partial derivatives are there for a function of 3 variables x, y , and z ? How many are the same for \mathcal{C}^2 functions?

Proof of the Theorem We include here a proof of Theorem 3.3 because some of you might be curious to see how it is done. However, you will be excused if you choose to skip the proof.

The proof is based on the Mean Value Theorem (as was the proof of Theorem 3.1).

We have

$$f_{yx}(x, y) = \lim_{\Delta x \rightarrow 0} \frac{f_y(x + \Delta x, y) - f_y(x, y)}{\Delta x}.$$

However,

$$\begin{aligned} f_y(x, y) &= \lim_{\Delta y \rightarrow 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y} \\ f_y(x + \Delta x, y) &= \lim_{\Delta y \rightarrow 0} \frac{f(x + \Delta x, y + \Delta y) - f(x + \Delta x, y)}{\Delta y}. \end{aligned}$$

Hence,

$$\begin{aligned} f_{yx} &= \lim_{\Delta y \rightarrow 0} \lim_{\Delta x \rightarrow 0} \left[\frac{f(x + \Delta x, y + \Delta y) - f(x + \Delta x, y)}{\Delta x \Delta y} - \frac{f(x, y + \Delta y) - f(x, y)}{\Delta x \Delta y} \right] \\ &= \lim_{\Delta y \rightarrow 0} \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x, y + \Delta y) - f(x + \Delta x, y) - f(x, y + \Delta y) + f(x, y)}{\Delta x \Delta y}. \end{aligned} \tag{47}$$

Call the expression in the numerator Δ . Now, put

$$g(x) = f(x, y + \Delta y) - f(x, y),$$

so the dependence on y is suppressed. Note that

$$g(x + \Delta x) - g(x) = f(x + \Delta x, y + \Delta y) - f(x + \Delta x, y) - f(x, y + \Delta y) + f(x, y) = \Delta.$$

By the Mean Value Theorem,

$$\Delta = g(x + \Delta x) - g(x) = g'(x_1) \Delta x$$

for some x_1 between x and $x + \Delta x$. Remembering what g is, we get

$$\Delta = (f_x(x_1, y + \Delta y) - f_x(x_1, y)) \Delta x.$$

(The differentiation in g' was with respect to x .) Now apply the Mean Value Theorem again to get

$$\Delta = (f_{xy}(x_1, y_1)\Delta y)\Delta x$$

for some y_1 between y and $y + \Delta y$. Note that $(x_1, y_1) \rightarrow (x, y)$ as $\Delta x, \Delta y \rightarrow 0$. Referring again to (47), we have

$$f_{yx}(x, y) = \lim_{\Delta y \rightarrow 0} \lim_{\Delta x \rightarrow 0} \frac{\Delta}{\Delta x \Delta y} = \lim_{\Delta y \rightarrow 0} \lim_{\Delta x \rightarrow 0} f_{xy}(x_1, y_1) = f_{xy}(x, y).$$

The last equality follows from the hypothesis that f_{xy} is continuous.

An Example It is easy to give an example of a function for which the mixed partials are not equal. Let

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

This formula does not make sense for $(x, y) = (0, 0)$ but it is not hard to see that $\lim_{(x, y) \rightarrow (0, 0)} f(x, y) = 0$. Hence, we can extend the definition of the function by defining $f(0, 0) = 0$, and the resulting function is continuous.

The mixed partial derivatives of this function at $(0, 0)$ are not equal. To see this, note first that

$$f_x(x, y) = \frac{y(x^4 + 4x^2y^2 - y^4)}{(x^2 + y^2)^2}$$

as long as $(x, y) \neq (0, 0)$. (That is a messy but routine differentiation which you can work out for yourself.) To determine $f_x(0, 0)$, note that for $y = 0$, we get $f(x, 0) = 0$. Hence, it follows that $f_x(x, 0) = 0$ for every x including $x = 0$. We can now calculate $f_{xy}(0, 0)$ as

$$f_{xy}(0, 0) = \lim_{\Delta y \rightarrow 0} \frac{f_x(0, 0 + \Delta y) - f_x(0, 0)}{\Delta y} = \lim_{\Delta y \rightarrow 0} \frac{1}{\Delta y} \frac{\Delta y (-(\Delta y)^4)}{(\Delta y^2)^2} = -1.$$

Calculating $f_{yx}(0, 0)$ (in the other order) proceeds along the same lines. However, since $f(y, x) = -f(x, y)$, you can see that reversing the roles of x and y will yield $+1$ instead, and indeed $f_{yx}(0, 0) = +1$. Hence, the two mixed partials are not equal.

An Application to Thermodynamics You may have been studying thermodynamics in your chemistry course. In thermodynamics, one studies the relationships among variables called *pressure*, *volume*, and *temperature*. These are usually denoted p , v , and T . They are assumed to satisfy some equation of state

$$f(p, v, T) = 0.$$

For example, $pv - kT = 0$ is the law which is supposed to hold for an ideal gas, but there are other more complicated laws such as the van der Waals equation. In any case, that means that we can pick two of the three variables as the independent variables and the remaining variable depends on them. However, there is no

preferred choice for independent variables and one switches from one to another, depending on the circumstances. In addition, one introduces other quantities such as the internal energy (u), the entropy (s), the enthalpy, etc. which are all functions of the other variables. (In certain circumstances, one may use these other variables as independent variables.)

It is possible to state two of the basic laws of thermodynamics in terms of entropy (s) and internal energy (u) as follows.

$$T \left(\frac{\partial s}{\partial T} \right)_v = \left(\frac{\partial u}{\partial T} \right)_v \quad \text{First Law}$$

$$T \left(\frac{\partial s}{\partial v} \right)_T = \left(\frac{\partial u}{\partial v} \right)_T + p \quad \text{Second Law.}$$

From these relations, it is possible to derive many others. For example, from the first law, we get by differentiating with respect to v ,

$$T \frac{\partial^2 s}{\partial v \partial T} = \frac{\partial^2 u}{\partial v \partial T}.$$

From the second law, we get by differentiating with respect to T

$$\left(\frac{\partial s}{\partial v} \right)_T + T \frac{\partial^2 s}{\partial T \partial v} = \frac{\partial^2 u}{\partial T \partial v} + \left(\frac{\partial p}{\partial T} \right)_v.$$

Subtracting the first from the second and using the equality of the mixed partials yields

$$\left(\frac{\partial s}{\partial v} \right)_T = \left(\frac{\partial p}{\partial T} \right)_v.$$

This is one of four relations called *Maxwell's relations*.

Exercises for 3.9.

1. Calculate f_{xy} and f_{yx} and show that they are equal.
 - (a) $f(x, y) = x^2 \cos y$.
 - (b) $f(x, y) = x \ln(x + 2y)$.
 - (c) $f(x, y) = x^3 - 5x^2y + 7xy^2 - y^3$.
2. Suppose $f(x, y)$ denotes a twice differentiable function with equal mixed partials. Show that if $f(x, y) = -f(y, x)$ then $f_{xy}(x, y) = 0$ at any point on the line $y = x$. Why does this show that the function considered in the Example in the section cannot have equal mixed partials at the origin?
3. For the function $f(x, y) = e^{x+y}$, show that all partials $\frac{\partial^{m+n} f}{\partial y^m \partial x^n}$ are the same. What are they?

4. The vertical displacement $y(x, t)$ of a point on a vibrating string (as a function of position x and time t) is governed by the *wave equation*

$$\frac{\partial^2 y}{\partial t^2} = c^2 \frac{\partial^2 y}{\partial x^2}.$$

(This is an approximation which is only accurate for small displacements.)
Show that there are solutions of the form:

(a) $y = \sin(kx + at).$

(b) $y = \sin kx \cos at.$

What are k and a ?

5. Laplace's Equation

$$\nabla^2 u = 0$$

governs many physical phenomena, e.g., the potential energy of a gravitational field.

Which of the following functions satisfy Laplace's Equation? Use $\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$ for a function $u(x, y)$ of two variables.

(a) $u = \ln(x + y)$

(b) $u = 2x - 3y + 4x^2 - 4y^2 + xy$

(c) $u = \sqrt{x^2 + y^2}$

6. Suppose $w = f(x, y)$. Then, in polar coordinates, $w = f(r \cos \theta, r \sin \theta) = h(r, \theta)$. Show that

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \frac{\partial^2 h}{\partial r^2} + \frac{1}{r} \frac{\partial h}{\partial r} + \frac{1}{r^2} \frac{\partial^2 h}{\partial \theta^2}.$$

7. Given the thermodynamic relation

$$du = T ds - p dv$$

derive the Maxwell relation

$$\left(\frac{\partial T}{\partial v} \right)_s = - \left(\frac{\partial p}{\partial s} \right)_v.$$

Chapter 4

Multiple Integrals

4.1 Introduction

We now turn to the subject of integration for functions of more than one variable. As before, the case of functions of two variables is a good starting point.

We start with a discussion of how integration commonly arises in applications, and to be concrete we shall discuss the concept of *center of mass*. (Your physics book should have a discussion of the significance of this concept.) Given a finite system of particles with masses m_1, m_2, \dots, m_n with position vectors at $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$, the center of mass of the system is defined to be

$$\mathbf{r}_{cm} = \frac{1}{M}(m_1\mathbf{r}_1 + m_2\mathbf{r}_2 + \cdots + m_n\mathbf{r}_n) = \frac{1}{M} \sum_{i=1}^n m_i \mathbf{r}_i$$

where

$$M = m_1 + m_2 + \cdots + m_n = \sum_{i=1}^n m_i$$

is the total mass. The center of mass is important in dynamics because it moves like a single particle of mass M which is acted upon by the sum of the external forces on the individual particles.

The principle we want to illustrate is that for each concept definable for finite sets of points, there is an appropriate generalization for continuous distributions in which the finite sum is replaced by an *integral*.

Example 48 Let a mass M be uniformly distributed along a thin rod of length L . We shall set up and evaluate an integral representing the center of mass of this system.



We treat the rod as if it were a purely 1-dimensional distribution, thus ignoring its extension in the other directions. We introduce a coordinate x to represent distance along the rod, so $0 \leq x \leq L$. Since the mass is uniformly distributed, the mass density per unit length will be

$$\delta = \frac{M}{L}.$$

The link between the finite and continuous is obtained by imagining that the rod is partitioned into small segments by picking division points

$$0 = x_0 < x_1 < x_2 < \cdots < x_{i-1} < x_i < \cdots < x_n = L.$$

The easiest way to do this would be to have them equally spaced, but to be completely general we don't assume that. Thus, the mass contained in the i th segment is $\Delta m_i = \delta \Delta x_i$ where $\Delta x_i = x_i - x_{i-1}$. We now replace each segment Δx_i by a *point mass* of the same mass Δm_i placed at the right hand endpoint x_i . (As we shall see, it is not critical where the point mass is positioned as long as it is somewhere in the segment. It could be instead at the left endpoint x_{i-1} or at any point \tilde{x}_i satisfying $x_{i-1} \leq \tilde{x}_i \leq x_i$.) The x -coordinate of the center of mass of this finite discrete system is

$$\frac{1}{M} \sum_{i=1}^n (\delta \Delta x_i) x_i = \frac{1}{M} \sum_{i=1}^n \delta x_i \Delta x_i.$$

To find—actually, to define—the x -coordinate of the center of mass of the original continuous distribution, we let $n \rightarrow \infty$ and all $\Delta x_i \rightarrow 0$ and take the limit. The result is the integral

$$x_{cm} = \frac{1}{M} \int_0^L \delta x \, dx.$$

This can be evaluated in the usual way

$$\frac{1}{M} \int_0^L \delta x \, dx = \frac{1}{M} \delta x \left. \frac{x^2}{2} \right|_0^L = \frac{1}{M} \frac{M}{L} \frac{L^2}{2} = \frac{L}{2}.$$

Thus $x_{cm} = L/2$ just as you would expect.

Note that had we used some other \tilde{x}_i in the above construction, we would have gotten instead the sum

$$\sum_{i=1}^n \delta \tilde{x}_i \Delta x_i,$$

but that wouldn't have made any difference in the limit. As you may remember from your previous study of integral calculus, all such sums are called *Riemann*

sums and, in the limit, it doesn't matter which \tilde{x}_i you choose in the i th segment. All such sums approach the same limit, the definite integral.

In the above example, we could have had a non-uniformly distributed mass. In that case, the linear density $\delta(x)$ would have been a non-constant function of position x . In that case, a typical Riemann sum would look like

$$\sum_{i=1}^n \delta(\tilde{x}_i) \tilde{x}_i \Delta x_i$$

where $x_{i-1} \leq \tilde{x}_i \leq x_i$, and in the limit it would approach the definite integral

$$\int_0^L \delta(x)x \, dx.$$

The x -coordinate of the center of mass would be

$$\frac{1}{M} \int_0^L \delta(x)x \, dx,$$

but in this case, using similar reasoning, we would calculate the mass as an integral

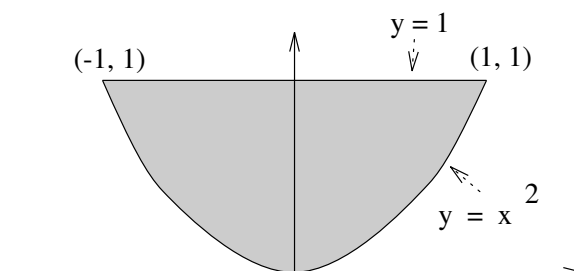
$$M = \int_0^L \delta(x) \, dx.$$

In such problems there are always two parts to the process. The first step is *conceptual*; it involves setting up an integral by visualizing it as a limit of finite sums. The second step is to evaluate the integral by using antiderivatives. In the example, $x^2/2$ doesn't have anything to do with any sums, it is just an antiderivative or indefinite integral of x . The link with limits of sums is provided by the Fundamental Theorem of Calculus.

Sometimes we can't evaluate the integral through the use of antiderivatives. For example, $\delta(x)x$ might be an expression for which there is no antiderivative expressible in terms of known functions. Then we have to go back to the idea of the definite integral as the limit of finite sums and try to approximate it that way. Refinements of this idea such as the trapezoidal rule or Simpson's rule are often helpful in that case.

Let's now consider a two dimensional example.

Example 49 Suppose a mass M is uniformly distributed over a thin sheet of some shape. By ignoring the thickness of the sheet, we may assume the mass is distributed over a region in a plane, and after choosing coordinates x, y in that plane, we may describe the region by the equations of its bounding curves. To be explicit, suppose the mass is distributed over the region D contained between the line $y = 1$ on top and the parabola $y = x^2$ underneath. We want to find the center of mass (x_{cm}, y_{cm}) of the distribution.



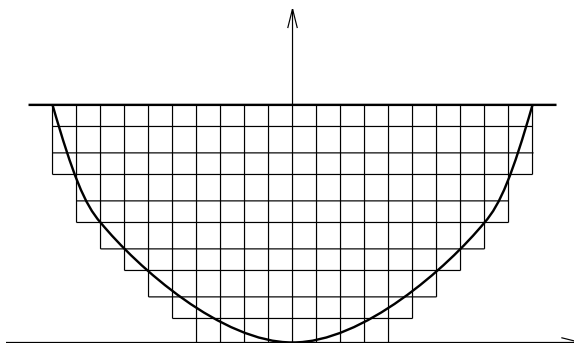
It is clear that $x_{cm} = 0$. (Why?) Hence, we concentrate on finding y_{cm} . We try to proceed as in Example 48. The first *conceptual* step will be to visualize it as the limit of finite sums.

Before beginning, note that the mass density per unit area will be $\delta = M/A$ where A is the area of the region D , but to find the area, we need to evaluate an integral.

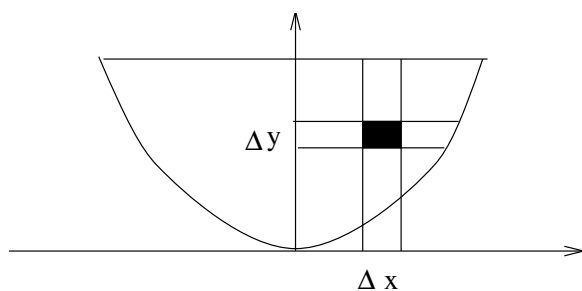
$$A = \int_{-1}^1 (1 - x^2) dx = x - \frac{x^3}{3} \Big|_{-1}^1 = \frac{4}{3}.$$

$$\text{Thus, } \delta = \frac{M}{4/3} = \frac{3M}{4}.$$

Imagine next that the region D is *dissected* into small rectangles by a grid as in the diagram. (There is a problem on the bottom edge where we won't have rectangles, but ignore that for the moment.)



Of course, to actually do this, we will have to number the rectangles some way, so we need subscripts and we have to keep track of the bookkeeping. However, for the moment let's ignore all that. Consider a typical rectangle with sides $\Delta x, \Delta y$, and area $\Delta A = \Delta x \Delta y$. The mass inside that rectangle will be $\Delta m = \delta \Delta A$.



Now imagine that each small rectangle is replaced by a particle of the same mass Δm placed at some point (x, y) inside the rectangle. (If the rectangles are all small enough, it won't matter much how the points (x, y) are chosen. For example, you could always choose the upper right corner, or the lower left corner, or the center, or anything else that takes your fancy.) The y -coordinate of the center of mass of this system will be

$$y_{cm} = \frac{1}{M} \sum_{\text{all rectangles}} y \Delta m = \frac{1}{M} \sum_{\text{all rectangles}} y \delta \Delta A.$$

Now let the number of rectangles approach ∞ while the size of each rectangle approaches zero. The sum approaches a *limit* which is denoted

$$\iint_D y \delta dA$$

and which is called a *double integral*. (Two integral signs are used to remind us of the dimensionality of the domain.) The y -coordinate of the center of mass of the continuous distribution is

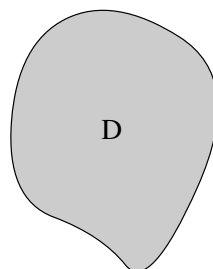
$$y_{cm} = \frac{1}{M} \iint_D y \delta dA.$$

This completes the first part of the problem: setting up the integral. However, we have no analogue *as yet* of the second step: evaluating the integral using antiderivatives and the Fundamental Theorem of Calculus. Of course, we could always approximate it by an appropriate sum, and the smaller we take the rectangles and the more of them there are, the better the approximation will be.

The General Concept of a Double Integral If the mass density per unit area were a function of position $\delta(x, y)$, then the mass in one of the small rectangles would be approximately $\delta(x, y) \Delta A$ where as above (x, y) is any point in that rectangle. In the limit, the y -coordinate of the center of mass would be

$$y_{cm} = \frac{1}{M} \iint_D y \delta(x, y) dA.$$

More generally, let D be a subset of \mathbf{R}^2 and let $f(x, y)$ denote a function defined on D .

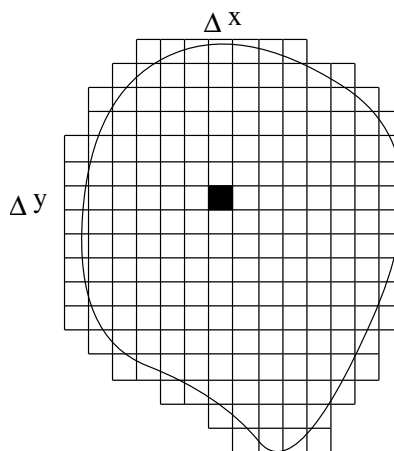


As above, dissect the region into small rectangles. For each such rectangle, let its sides have dimensions Δx and Δy , so its area is $\Delta A = \Delta x \Delta y$, and choose a point (x, y) in the rectangle. Form the sum

$$\sum_{\text{all rectangles}} f(x, y) \Delta A$$

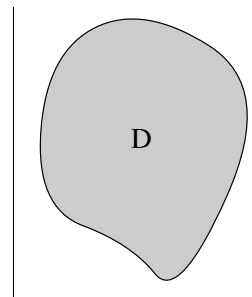
and consider what happens as the size of each rectangle approaches zero and the number of rectangles approaches ∞ . If the resulting sums approach a definite limit, we call that limit the *double integral* of the function over the region D and we denote it

$$\iint_D f(x, y) dA.$$

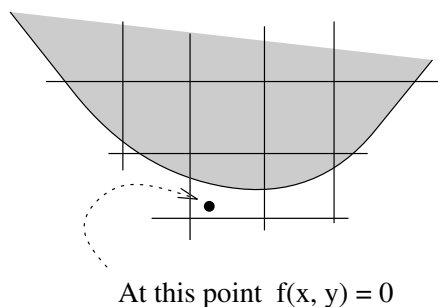
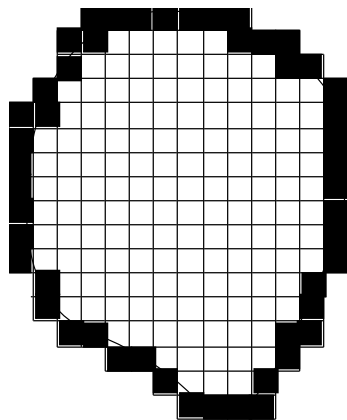


There are some problems with this definition that should be discussed briefly. First of all, in principle the set D could be quite arbitrary, but in that case the limit may not exist and in any case it may be impossible to evaluate. Usually, in useful situations, the region D is something quite reasonable. For example, it might be

bounded by a finite set of smooth curves as in Example 49. Secondly, the region D had better be bounded; that is, it should be possible to enclose it in a sufficiently large rectangle. If that were not the case, it would not be possible to dissect the region into *finitely many* small rectangles. (What one does for unbounded regions will be discussed later.)



Another issue was raised briefly above. On the boundary of the region D , the dissection may not yield rectangles. It turns out that in all reasonable cases, this does not matter. For suppose the dissection into small rectangles is obtained by imposing a grid on an enclosing rectangle R containing the region D . Consider the rectangles in the grid which overlap the region D but don't lie entirely within it. If we allow some or all of these rectangles in the sum, we run the risk of "overestimating the" sum, but if we omit them all, we run the risk of "underestimating" it. However, in all reasonable cases, it won't matter which we do, since the total area of these questionable rectangles will be small compared to the area of the region D , and it will approach zero in the limit. That is because, in the limit, we would obtain the "area" of the boundary of D , and since the boundary would ordinarily be a finite collection of smooth curves, that "area" would be zero. One way to deal with this question of partially overlapping rectangles is as follows. The contribution from such a rectangle would be $f(x, y)\Delta A$ where (x, y) is some point in the rectangle. If the point is in the region D , we include the term in the sum. On the other hand, we could have chosen for that rectangle a point (x, y) not in the region D , so $f(x, y)$ might not even be defined. In that case, just redefine it to be zero, so the term $f(x, y)\Delta A$ would be zero in any case. In essence, this amounts to defining a new function $f^*(x, y)$ which agrees with the function f inside the region D and is zero outside the region, and considering sums for this function.



Finally, we come to the nitty gritty of how we go about adding up the contributions from the individual small rectangles. Getting this straight is essential either

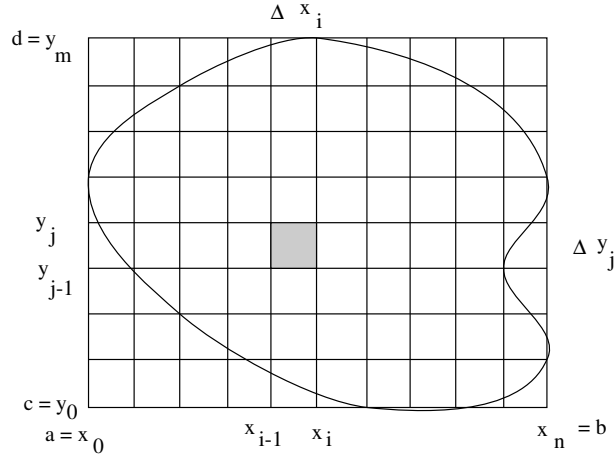
for developing a precise, rigorous theory of the double integral, or for actually approximating it numerically, say using a computer program. Here is how it is done. Assume the region D is contained in a (large) rectangle described by two inequalities $a \leq x \leq b$ and $c \leq y \leq d$. We form a grid on this rectangle by choosing division points

$$a = x_0 < x_1 < x_2 < \cdots < x_{i-1} < x_i < \cdots < x_m = b$$

along the x axis, and

$$c = y_0 < y_1 < y_2 < \cdots < y_{j-1} < y_j < \cdots < y_n = d$$

along the y -axis. Hence, each rectangle is characterized by a pair of indices (i, j) where $1 \leq i \leq m, 1 \leq j \leq n$. There are a total mn rectangles. The division points can be chosen in any convenient manner. They might be equally spaced, but they need not be. Put $\Delta x_i = x_i - x_{i-1}$, $\Delta y_j = y_j - y_{j-1}$ and $\Delta A_{ij} = \Delta x_i \Delta y_j$.



In the (i, j) -rectangle, we choose a point $(\tilde{x}_{ij}, \tilde{y}_{ij})$. As mentioned above, there are a variety of ways we could choose such a point. The contribution from this rectangle will be $f(\tilde{x}_{ij}, \tilde{y}_{ij})\Delta A_{ij}$ except that we will agree to set $f(\tilde{x}_{ij}, \tilde{y}_{ij}) = 0$ if the point is not in the region D . Finally, we add up the contributions. There are clearly many ways we could do this. For example, we could form

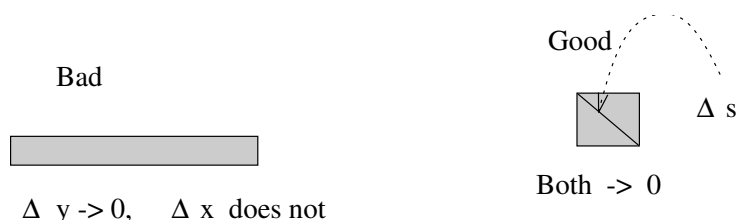
$$\sum_{j=1}^n \sum_{i=1}^m f(\tilde{x}_{ij}, \tilde{y}_{ij}) \Delta A_{ij}$$

which amounts to summing across “horizontal strips” first and then summing up the contributions from these strips. Alternately, we could form

$$\sum_{i=1}^m \sum_{j=1}^n f(\tilde{x}_{ij}, \tilde{y}_{ij}) \Delta A_{ij}$$

which sums first along “vertical strips”. There are even more complicated ways of adding up, which might be a good idea in special circumstances, but the bookkeeping would be more complicated to describe.

To complete the process, it is necessary to take a limit, but this also requires some care. If we just let the number mn of rectangles go to ∞ , we could encounter problems. For example, if $n \rightarrow \infty$, but m stays fixed, we would expect some difficulty. In that case, the area ΔA_{ij} of individual rectangles would approach zero, but their width would stay constant. The way around this is to insist that not only should the number of rectangles go to ∞ , but also the largest possible *diagonal* of any rectangle in a dissection should approach zero.



Perhaps it will make the process seem a bit more concrete if we give a (Pascal) computer program to approximate the integral in Example 49

$$\iint_D y \delta \, dA$$

where D is the region described by $x^2 \leq y \leq 1$, $-1 \leq x \leq 1$. As above, the variable δ represents the density. The horizontal interval $-1 \leq x \leq 1$ is divided into m equal subintervals, each of length $Dx = 2/m$. Similarly, the y -range is divided into m equal subintervals, each of length $Dy = 1/m$. (Note that this means that $n = m$.) Thus, the region D is covered by a collection of subrectangles, each of area $DA = Dx Dy$, but some of these subrectangles overlap the bottom edge of the region. The integrand is evaluated in the i, j -rectangle at the upper right hand corner (x, y) . The sums are done first along vertical strips, but in such a way that subrectangles below the bottom edge of the region don't contribute to the sum. (Examine the program to see if the subrectangles overlapping the bottom edge contribute to the sum.)

```

program integrate;
var m,i,j : integer;
    x, y, Dx, Dy, DA, delta, sum : real;
begin
    write('Give density');
    readln(delta);
    write('Give number of subintervals');

```

```

readln(m);
Dx := 2/m;
Dy := 1/m;
DA := Dx*Dy;
x := -1.0;
sum := 0.0;
for i := 1 to m do
  begin
    x := x + Dx;
    y := 0;
    for j := 1 to m do
      begin
        y := y + Dy;
        if x*x <= y then
          sum := sum + delta*y*DA;
        end;
      end;
    end;
  writeln('Approximation = ', sum);
end.

```

Here are some results from an equivalent program (written in the programming language C) where $\delta = 1.0$.

m	Approximation
10	0.898000
50	0.818863
100	0.809920
200	0.804981
1000	0.800981

As we shall see in the next section, the exact answer, determined by calculus, is 0.8. If you examine the program carefully, you will note that the approximating sums overestimate the answer because the overlapping rectangles on the bottom parabolic edge are more often included than excluded.

Other Notations There are a variety of other notations you may see for double integrals. First, you may see

$$\int_D f(x, y) dA$$

with just one integral sign to stand for the summation process. In that case, you have to look for other clues to the dimensionality of the domain of integration. You may also see

$$\iint_D f(x, y) dx dy.$$

The idea here is that each small element of area is a rectangle with sides dx and dy , so the area is $dA = dx dy$. However, we shall see later that there may be

advantages to dissecting the region into things other than rectangles (as when using polar coordinates, for example). Hence, it is better generally to stick with the ‘ dA ’ notation. In the next section, we shall introduce the notion of an *iterated integral*. This is a related but logically distinct concept with a slightly different notation, and the two concepts should not be confused.

What Is an Integral? If you ignore the subtlety involved in taking a limit, then an integral of the form

$$\int_a^b f(x) dx \quad \text{or} \quad \iint_D f(x, y) dA$$

should be thought of basically as a *sum*. In the first case, you take an element of length dx , weight it by the factor $f(x)$ to get $f(x) dx$ and then add up the results. Similarly, in the second case, the element of area dA is weighted by the factor $f(x, y)$ to get $f(x, y) dA$ and the results are added up to get the double integral. These are the first of many examples we shall see of such *sums*. Of course, the concepts have many important applications, but no one of these applications tells you what the integral ‘really is’. Unfortunately, students are often misled by their first introduction to integrals where the *sum* $\int_a^b f(x) dx$ is interpreted as the area between the x -axis and the graph of $y = f(x)$. This is one of many interpretations of such an integral (i.e., sum) and should not be assigned special significance. Similarly, all the other integrals we shall discuss are (except for technicalities concerning limits) sums; they are not areas or volumes or any special interpretation.

Exercises for 4.1.

1. Enter the program in the text into a computer, compile it and run it. Try different values of m and see what happens. (If you prefer, you may write an equivalent program in some other programming language.)
2. Modify the program in the text to approximate the double integral $\iint_D (x + y) dA$ where D is the region above the line $y = x$ and below the parabola $y = \sqrt{x}$. Run your program for various values of m .
3. (Optional) The program in the text is not very efficient. Try to improve it so as to cut down the number of multiplications. Also, see if you can think of better ways to approximate the integral. For example, instead of using the value of the integrand at the upper right hand corner of each subrectangle, you might use the average value of the integrand at the four corners of the subrectangle. See if that gives greater accuracy for the same value of m .

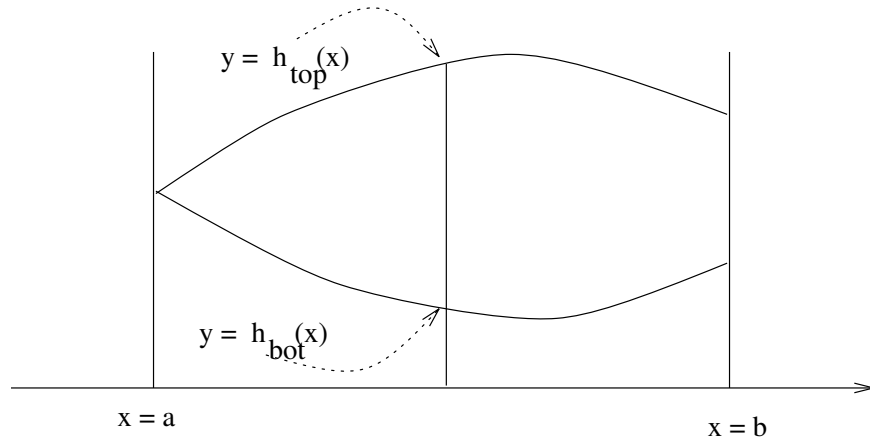
4.2 Iterated Integrals

Let $z = f(x, y) = f(\mathbf{r})$ denote a function of two variables defined on some region D in \mathbf{R}^2 . In the previous section we discussed the definition of the double integral

$$\iint_D f(x, y) dA.$$

We discuss next how you can use antiderivatives to calculate such a double integral.

Suppose first that the region D is *bounded vertically by graphs of functions*. More precisely, suppose the region is bounded on the left and right by vertical lines $x = a, x = b$ and between those lines it is bounded below by the graph of a function $y = h_{\text{bot}}(x)$ and above by the graph of another function $y = h_{\text{top}}(x)$.



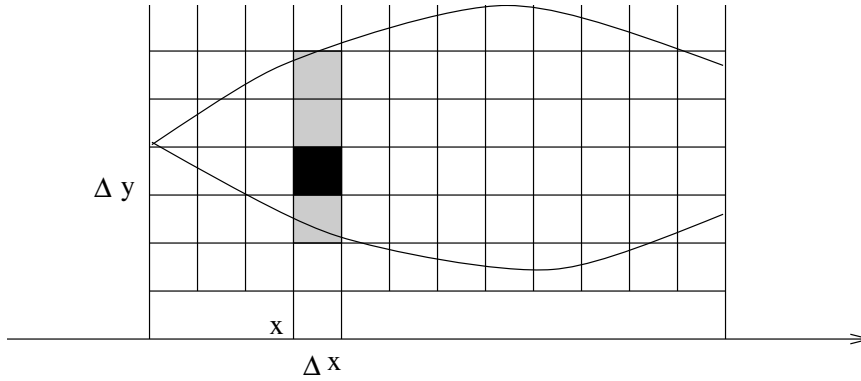
A region bounded vertically by graphs

The region in Example 2 in the previous section was just such a region: $a = -1, b = 1, h_{\text{bot}}(x) = x^2, h_{\text{top}}(x) = 1$. Many, but not all regions have such descriptions.

As in the previous section, imagine the region dissected into many small rectangles by imposing a mesh, and indicate an approximating sum for the double integral $\iint_D f(x, y) dA$ schematically by

$$\sum_{\text{all rectangles}} f(x, y) \Delta x \Delta y$$

where we have used $\Delta A = \Delta x \Delta y$.



We pointed out in the previous section that there are two rather obvious ways to add up the terms in the sum (as well as many not very obvious ways to do it.) The way which suggests itself here is to add up along each vertical strip and then add up the contributions from the individual strips:

$$\sum_{\text{all rectangles}} f(x, y) \Delta x \Delta y = \sum_{\text{all strips at } x} \sum_{\text{strip}} f(x, y) \Delta x \Delta y.$$

For any given strip, we can assume we are using the same value of x , (say the value of x at the right side of all the rectangles in that strip), and we can factor out the common factor Δx to obtain

$$\sum_{\text{all strips}} \left(\sum_{\text{strip at } x} f(x, y) \Delta y \right) \Delta x. \quad (48)$$

Now let the number of rectangles go to ∞ while the size of each rectangle shrinks to 0. Concentrating on what happens in a vertical strip, we see that

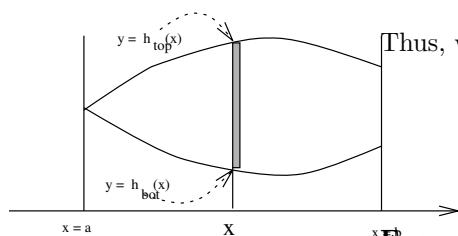
$$\sum_{\text{strip at } x} f(x, y) \Delta y \longrightarrow G(x) = \int_{y=h_{\text{bot}}(x)}^{y=h_{\text{top}}(x)} f(x, y) dy.$$

Note that the integral on the right is a ‘partial integral’. That is, x is temporarily kept constant and we integrate with respect to y . Note also that the limits depend on x since, in the sum on the left, we only include rectangles that lie within the region, i.e., those for which $h_{\text{bot}}(x) \leq y \leq h_{\text{top}}(x)$. Thus the integral is altogether a function $G(x)$ of x . If we replace the internal sum in expression (48) by the partial integral $G(x)$ (which it is approximately equal to), we get

$$\sum_{\text{all strips}} G(x) \Delta x$$

which in the limit approaches

$$\int_a^b G(x) dx = \int_a^b \left(\int_{y=h_{\text{bot}}(x)}^{y=h_{\text{top}}(x)} f(x, y) dy \right) dx.$$



Thus, we obtain finally the formula

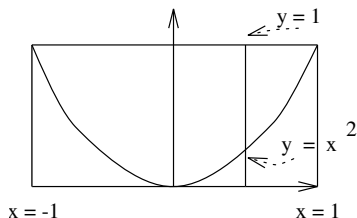
$$\iint_D f(x, y) dA = \int_a^b \left(\int_{y=h_{bot}(x)}^{y=h_{top}(x)} f(x, y) dy \right) dx. \quad (49)$$

Example 50, (Example 2, Section 1) We had

$$y_{cm} = \frac{1}{M} \iint_D y \delta dA$$

where $\delta = 3M/4$. We use the above method to calculate the integral, where $a = -1, b = 1, h_{bot}(x) = x^2$, and $h_{top}(x) = 1$.

$$\begin{aligned} \iint_D y \delta dA &= \int_{-1}^1 \left(\int_{y=x^2}^{y=1} y \delta dy \right) dx \\ &= \int_{-1}^1 \left(\delta \frac{y^2}{2} \Big|_{x^2}^1 \right) dx \\ &= \int_{-1}^1 \delta \left(\frac{1}{2} - \frac{x^4}{2} \right) dx \\ &= \delta \int_{-1}^1 \frac{1}{2} (1 - x^4) dx \\ &= \frac{\delta}{2} \left[x - \frac{x^5}{5} \Big|_{-1}^1 \right] \\ &= \frac{\delta}{2} \left[1 - \frac{1}{5} - \left(-1 - \frac{-1}{5} \right) \right] \\ &= \frac{8\delta}{10}. \end{aligned}$$



(Note that if $\delta = 1$, this gives 0.8 as suggested by numerical approximation in Section 1.) We may now determine the y -coordinate of the center of mass

$$y_{cm} = \frac{1}{M} \delta \frac{8}{10} = \frac{1}{M} \frac{3M}{4} \frac{8}{10} = \frac{3}{5}.$$

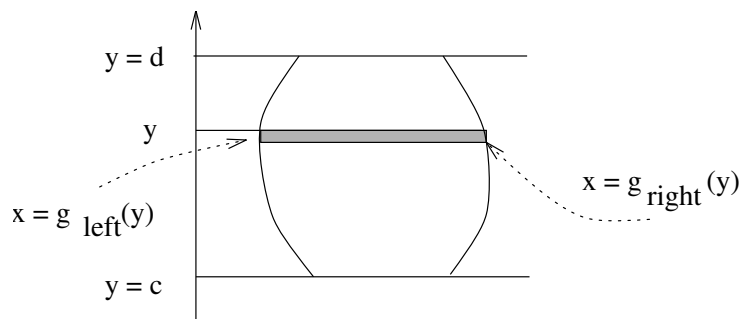
There are a couple of remarks that should be made about formula (49). As noted earlier, the double integral on the left is defined as the limit of sums obtained by dissecting the region. The integral on the right is called an *iterated integral* and it represents something different. It is an ‘integral of an integral’ where the inner integral is a ‘partial integral’ depending on x . However, both integrations are with respect to a *single variable*, so both can be evaluated by means of anti-derivatives and the Fundamental Theorem, as we did in the example.

The above derivation of formula (49) is not a rigorous argument. It is not possible to separate the process of taking the inner limit (counting vertically) from the

outer limit (counting horizontally). A correct proof is actually quite difficult, and it is usually done in a course in real analysis. The correctness of the formula for reasonable functions is called *Fubini's Theorem*. You should remember the intuition behind the argument, because it will help you understand how to turn a double integral into an iterated integral.

If we reverse the roles of x and y in the above analysis, we obtain a formula for regions D which are *bounded horizontally by graphs*. Such a region is bounded below by a line $y = c$, above by a line $y = d$, and between those lines it is bounded on the left by the graph of a function $x = g_{\text{left}}(y)$ and on the right by the graph of another function $x = g_{\text{right}}(y)$. For such a region, the double integral is evaluated by summing first along horizontal strips (y constant, x changing) and then summing vertically the contributions from the strips (y changing). The analogous formula is

$$\iint_D f(x, y) dA = \int_c^d \left(\int_{x=g_{\text{left}}(y)}^{x=g_{\text{right}}(y)} f(x, y) dx \right) dy. \quad (50)$$

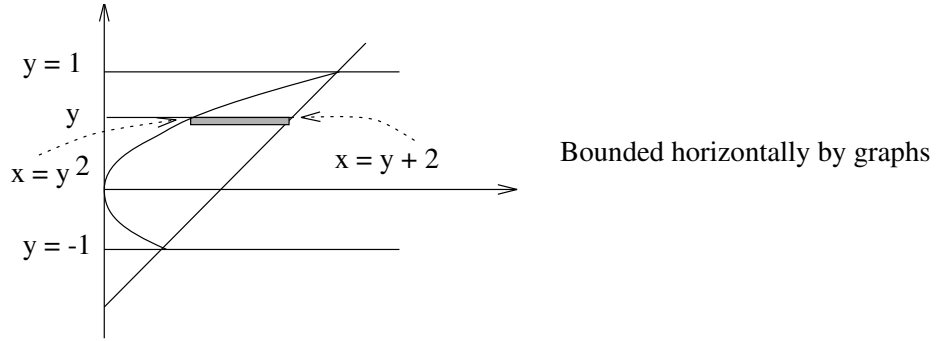


Note that the integration is in the direction in which the variables *increase*: x from left to right, and y from bottom to top.

Example 51 Let $f(x, y) = x^2 + y^2$, and let D be the region between the parabola $x = y^2$ on the left and the line $x = y + 2$ on the right. These curves intersect when

$$\begin{aligned} y^2 &= y + 2 \\ \text{or} \quad y^2 - y - 2 &= 0 \\ \text{or} \quad (y - 2)(y + 1) &= 0 \\ \text{so} \quad y &= 2 \quad \text{or} \quad y = -1. \end{aligned}$$

Hence, D also lies between $y = -1$ below and $y = 2$ above.

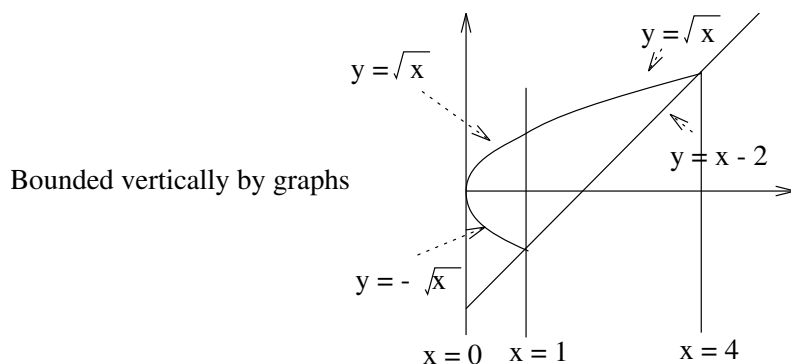


Thus

$$\begin{aligned}
 \iint_D x^2 + y^2 \, dA &= \int_{-1}^2 \left(\int_{x=y^2}^{x=y+2} x^2 + y^2 \, dx \right) dy \\
 &= \int_{-1}^2 \left(\frac{x^3}{3} + xy^2 \Big|_{y^2}^{y+2} \right) dy \\
 &= \int_{-1}^2 \left(\frac{(y+2)^3}{3} + (y+2)y^2 - \frac{y^6}{3} - y^4 \right) dy \\
 &= \left(\frac{(y+2)^4}{12} + \frac{y^4}{4} + \frac{2y^3}{3} - \frac{y^7}{21} - \frac{y^5}{5} \right) \Big|_{-1}^2 \\
 &= \frac{256}{12} + \frac{16}{4} + \frac{16}{3} - \frac{128}{21} - \frac{32}{5} - \frac{1}{12} - \frac{1}{4} + \frac{2}{3} - \frac{1}{21} - \frac{1}{5} \\
 &= \frac{639}{35} \approx 18.26.
 \end{aligned}$$

Note that the region is also bounded vertically by graphs, so in principle the integral could be evaluated by the previous method using formula (49). There is a serious problem in trying this, however. The top graph is that of $y = h_{top}(x) = \sqrt{x}$, but the bottom graph is described by two different formulas depending on what x is. It is a parabola to the left of the point $(1, -1)$ and a line to the right of that point, so

$$h_{bot}(x) = \begin{cases} -\sqrt{x} & 0 \leq x \leq 1 \\ y - 2 & 1 \leq x \leq 4. \end{cases}$$



(The x -values at the relevant points are determined from the corresponding y -values which were calculated above.) That means to do the calculation effectively using vertical strips we must in effect decompose the region D into two subregions meeting along the line $x = 1$ and treat each one separately. Then

$$\iint_D x^2 + y^2 dA = \int_0^1 \left(\int_{y=-\sqrt{x}}^{y=\sqrt{x}} x^2 + y^2 dy \right) dx + \int_1^4 \left(\int_{y=x-2}^{y=\sqrt{x}} x^2 + y^2 dy \right) dx.$$

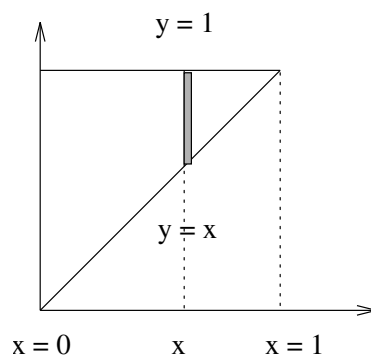
You should work out the two iterated integrals on the right just to check that their sum gives the same answer as above.

As the previous example indicates, even if in principle you could treat a region as either bounded vertically by graphs or as bounded horizontally by graphs, the choice can make a big difference in how easy it is to calculate. In some cases, it may be impossible to do the iterated integrals by antiderivatives in one order but fairly easy in the other order.

Example 52 Consider the iterated integral

$$\int_0^1 \int_x^1 \frac{\sin y}{y} dy dx.$$

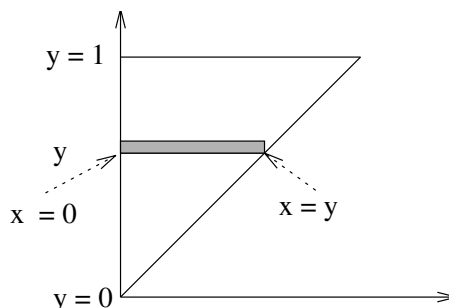
This is the iterated integral obtained from the double integral of the function $f(x, y) = \sin y/y$ for the triangular region D contained between the vertical lines $x = 0, x = 1$, the line $y = x$ below, and the line $y = 1$ above.



The indefinite integral (anti-derivative)

$$\int \frac{\sin y}{y} dy$$

cannot be expressed in terms of known elementary functions. (Try to integrate it or look in an integral table if you don't believe that.) Hence, the iterated integral cannot be evaluated by anti-derivatives. However, the triangular region may be described just as well by bounding it horizontally by graphs: it lies between $y = 0$ and $y = 1$ and for each y between



$x = 0$ and $x = y$. Thus, the double integral can be evaluated from the iterated integral

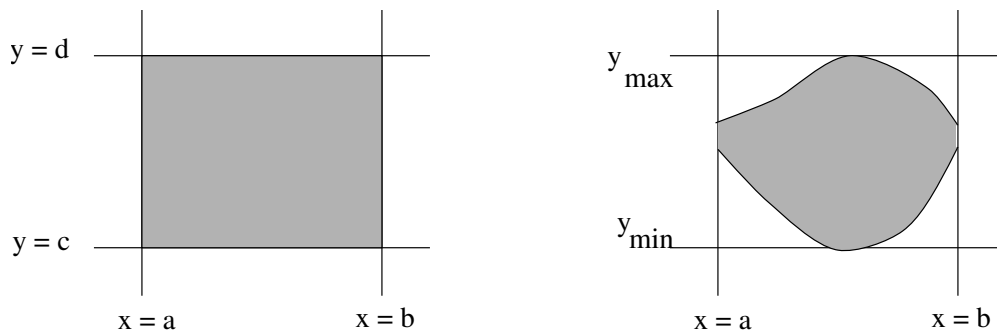
$$\begin{aligned} \int_0^1 \int_0^y \frac{\sin y}{y} dx dy &= \int_0^1 \frac{\sin y}{y} x \Big|_0^y dy \\ &= \int_0^1 \frac{\sin y}{y} y dy = \int_0^1 \sin y dy \\ &= -\cos y \Big|_0^1 = 1 - \cos 1. \end{aligned}$$

Note that in order to set up the iterated integral in the other order, *we had to draw a diagram* and work directly from that. *There are no algebraic rules which will allow you to switch orders without using a diagram.*

The simplest kind of region is a rectangle, described, say, by inequalities $a \leq x \leq b, c \leq y \leq d$. In this case the iterated integrals look like

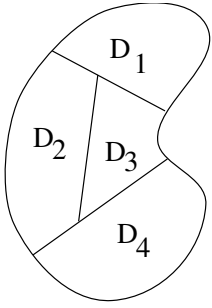
$$\int_a^b \int_c^d f(x, y) dy dx = \int_c^d \int_a^b f(x, y) dx dy$$

and it should not make much difference which you choose. You should immediately recognize a rectangular region from the fact that the internal limits are both constants. In the general case they will depend on one of the variables. Since the geometry can be somewhat complicated, it is easy to put constant limits where they are not appropriate. Suppose for example we have a region bounded vertically by graphs. The appropriate way to think of it is that we temporarily fix one value of x (between the given x -limits), and *for that x* add up along a vertical strip (y varying) of width dx . Hence, the limits for that strip will generally depend on x . Unfortunately, students often oversimplify and take for limits the minimum and maximum values of y for the region as a whole. If you do that, you have in effect replaced the desired region by a minimal bounding rectangle. (See the diagram.) You can recognize that you have done that when the limits tell you the region is a rectangle, but you know it is not.



In the previous examples, we dealt with regions which were bounded by graphs, either vertically or horizontally. There are many examples of regions which are neither. For such a region, we employ the ‘divide and conquer’ strategy. That is, we try to decompose the region into subregions that are bounded by graphs. In so doing we use the following *additivity* rule. If D can be decomposed into subsets D_1, D_2, \dots, D_k where at worst any two subsets D_i and D_j share a common boundary which is a smooth curve, then

$$\iint_D f dA = \iint_{D_1} f dA + \iint_{D_2} f dA + \cdots + \iint_{D_k} f dA.$$



(This rule certainly makes sense intuitively, but is a little tricky to prove.)

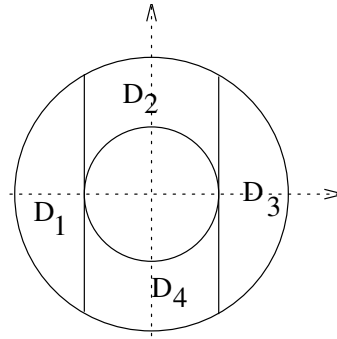
Example 53 Consider

$$\iint_D \frac{1}{x^2 + y^2} dA$$

for D the region between the circle $x^2 + y^2 = 1$ and the circle $x^2 + y^2 = 4$. D can be decomposed into 4 regions

$$D = D_1 \cup D_2 \cup D_3 \cup D_4$$

as indicated in the diagram.



Each of these regions is bounded by graphs, and the double integrals on them may be evaluated by iterated integrals. Thus, we have

$$\begin{aligned} \iint_{D_1} \frac{1}{x^2 + y^2} dA &= \int_{-2}^{-1} \int_{-\sqrt{4-x^2}}^{\sqrt{4-x^2}} \frac{1}{x^2 + y^2} dy dx \\ \iint_{D_2} \frac{1}{x^2 + y^2} dA &= \int_{-1}^1 \int_{\sqrt{1-x^2}}^{\sqrt{4-x^2}} \frac{1}{x^2 + y^2} dy dx \\ \iint_{D_3} \frac{1}{x^2 + y^2} dA &= \int_1^2 \int_{-\sqrt{4-x^2}}^{\sqrt{4-x^2}} \frac{1}{x^2 + y^2} dy dx \\ \iint_{D_4} \frac{1}{x^2 + y^2} dA &= \int_{-1}^1 \int_{-\sqrt{4-x^2}}^{\sqrt{1-x^2}} \frac{1}{x^2 + y^2} dy dx. \end{aligned}$$

Because of the symmetric nature of the integrand $1/(x^2 + y^2)$, the integrals for D_1 and D_3 are the same as are those for D_2 and D_4 . Hence, only two of the four integrals need to be computed. However, these integrals are not easy to do. For

example, for the region D_2 ,

$$\begin{aligned} \int_{\sqrt{1-x^2}}^{\sqrt{4-x^2}} \frac{1}{x^2+y^2} dy &= \frac{1}{x} \tan^{-1}\left(\frac{y}{x}\right) \Big|_{\sqrt{1-x^2}}^{\sqrt{4-x^2}} \\ &= \frac{1}{x} \left(\tan^{-1}\left(\frac{\sqrt{4-x^2}}{x}\right) - \tan^{-1}\left(\frac{\sqrt{1-x^2}}{x}\right) \right). \end{aligned}$$

Hence,

$$\iint_{D_2} \frac{1}{x^2+y^2} dA = \int_{-1}^1 \frac{1}{x} \left(\tan^{-1}\left(\frac{\sqrt{4-x^2}}{x}\right) - \tan^{-1}\left(\frac{\sqrt{1-x^2}}{x}\right) \right) dx.$$

I asked Mathematica to do this for me, but it could not give me an exact answer. It claimed that a good numerical approximation was 1.16264.

We shall see in the next section how to do this problem using polar coordinates. It is much easier that way.

Things to Integrate The choice of $f(x, y)$ in

$$\iint_D f(x, y) dA$$

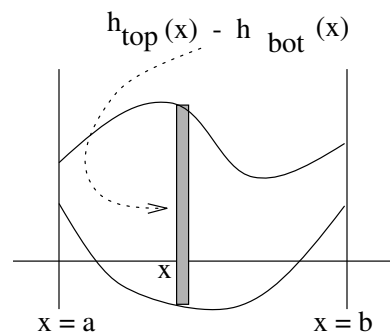
depends strongly on what problem we want to solve. We saw that $f(x, y) = y\delta$ was appropriate for finding the y -coordinate of the center of mass. You will later learn about moments of inertia where the appropriate f might be $f(x, y) = (x^2 + y^2)\delta$. In electrostatics, $f(x, y)$ might represent the charge density, and the integral would be the total charge in the domain D . For the special case, $f(x, y) = 1$, the integral

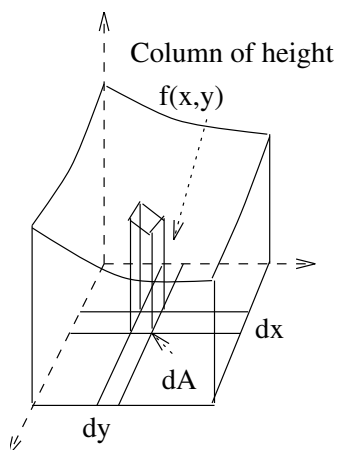
$$\iint_D 1 dA$$

is just the *area of the region D* . (The integral just amounts to adding up the contributions from small elements of area ' dA ', so the sum is just the total area.) If the region D is bounded vertically by graphs $y = h_{bot}(x)$ and $y = h_{top}(x)$, and extends horizontally from $x = a$ to $x = b$, we get

$$\begin{aligned} A &= \iint_D 1 dA = \int_a^b \int_{h_{bot}(x)}^{h_{top}(x)} dy dx \\ &= \int_a^b y \Big|_{h_{bot}(x)}^{h_{top}(x)} dx \\ &= \int_a^b (h_{top}(x) - h_{bot}(x)) dx. \end{aligned}$$

You may recognize the last expression as the appropriate formula for the area between two curves that you learned in your single variable integral calculus course.





There is one fairly simple geometric interpretation which always makes sense. $f(x, y)$ is the *height* at the point (x, y) of the graph of the function. We can think of $f(x, y)dA$ as the *volume* is a column of height $f(x, y)$ and base area dA . Hence, the double integral $\iint_D f dA$ represents the *volume* under the graph of the function. However, just as in the case of area under a graph in single variable calculus, this volume must be considered a *signed* quantity. Contributions from the part of the graph under the x, y -plane are considered negative, and the integral is the algebraic sum of the contributions from above and below the x, y -plane.

Exercises for 4.2.

- Evaluate the following iterated integrals with constant limits: (a) $\int_1^2 \int_0^3 (5x - 3y) dy dx$ (b) $\int_0^2 \int_0^2 (xy - 7x^2y^2 + 2xy^3 - y^5) dy dx$ (c) $\int_0^\pi \int_0^{\frac{\pi}{2}} (\cos x \sin y) dy dx$
- Show, for the answers to Problem 1, that the order of integration is unimportant, i.e.

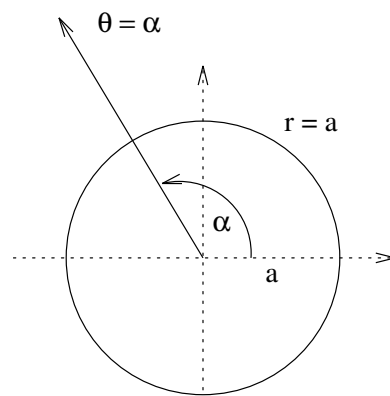
$$\int_a^b \int_c^d f(x, y) dy dx = \int_c^d \int_a^b f(x, y) dx dy.$$
- Evaluate the following iterated integrals: (a) $\int_0^2 \int_0^x (2x - 5y) dy dx$. (b) $\int_0^1 \int_{x^2}^x (2y - x^2) dy dx$. (c) $\int_0^1 \int_0^{x^2} e^{y/x} dy dx$. (d) $\int_0^\pi \int_0^y \cos(y^2) dx dy$.
- For each of the following, sketch the region of integration, then use your drawing to reverse the order of integration, then evaluate the new iterated integral. (a) $\int_0^2 \int_0^{x^2} 4xy dy dx$. (b) $\int_0^1 \int_{3x}^{4-x^2} 1 dy dx$. Hint: it isn't easier in the other order in this case. (c) $\int_0^\pi \int_y^\pi \frac{\sin x}{x} dx dy$. (d) $\int_0^1 \int_{\sqrt{y}}^1 e^{x^3} dx dy$.
- Use double integration to find the area between each of the following sets of curves.
 - $x = y^2, y = x^2$.
 - $y = x, y = 2x^2 - 3$.
- Use double integration to find the volume under the surface $z = f(x, y)$ and between the given curves.
 - $z = x + y, x = 0, x = 2, y = 0, y = 2$.
 - $z = 2 - x + 3y, x = -1, x = 1, y = 0, y = 1$.
 - $z = x^2 - 3 \cos y, x = 0, x = 2, y = \frac{\pi}{2}, y = \pi$.
 - $z = e^x + e^y, x = 0, x = 1, y = 0, y = x$.
 - $z = 4 - x^2 - y^2, y = x, y = x^2$.

7. Use appropriate double integrals to determine the volume of each of the following solids. You may take advantage of symmetry.
- (a) A sphere of radius a .
 - (b) A cylinder with base of radius a , and height h .
 - (c) An ellipsoid with semimajor axes a , b , and c .

4.3 Double Integrals in Polar Coordinates

Some double integrals become much simpler if one uses polar coordinates. If there is circular symmetry of some sort present in the underlying problem, you should always consider polar coordinates as a possible alternative.

Graphing in Polar Coordinates You should learn to recognize certain curves when expressed in polar coordinates. For example, as mentioned in Chapter I, Section 2, the equation $r = a$ describes a circle of radius a centered at the origin, and the equation $\theta = \alpha$ describes a ray starting at the origin and extending to ∞ . The ray makes angle α with the positive x -axis. Note that, depending on the value of α , the ray could be in any of the four quadrants.



Here are some more complicated graphs.

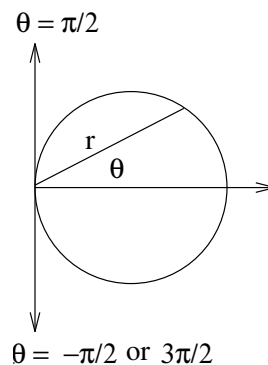
Example 54 The equation $r = 2a \cos \theta$ describes a circle of radius a centered at the point $(a, 0)$. To see this transform to rectangular coordinates, using first $\cos \theta = x/r$, and then $r^2 = x^2 + y^2$, so

$$r = 2a \frac{x}{r} \quad \text{or} \quad r^2 = 2ax \quad \text{or} \quad x^2 + y^2 = 2ax.$$

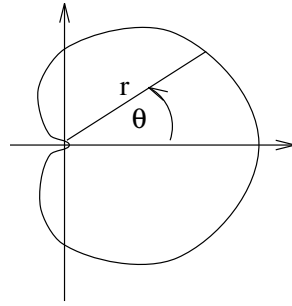
Now transpose and complete the square to obtain

$$x^2 - 2ax + a^2 + y^2 = a^2 \quad \text{or} \quad (x - a)^2 + y^2 = a^2.$$

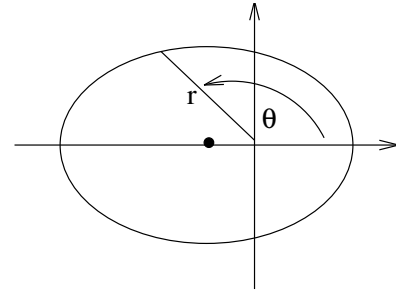
Note that the equation $r = 2a \cos \theta$ has no locus for $\pi/2 < \theta < 3\pi/2$ since in that range $\cos \theta < 0$, and we are not allowing r to be negative. If we were to allow r to be negative, we would retrace the circle. (Can you see why?) This example shows the importance of thinking carefully about what the symbols mean. When we get to integration in polar coordinates you will see that an unthinking use of formulas in a case like this can lead either to double the correct answer or zero because some part of a figure is considered twice, possibly with the same sign or possibly with opposite signs.



Example 55 The equation $r = a(1 + \cos \theta)$ has as locus a curve called a *cardioid*. See the picture. Perhaps you can see the reason for the name. Here, the appropriate range of θ would be $0 \leq \theta \leq 2\pi$.



Cardioid



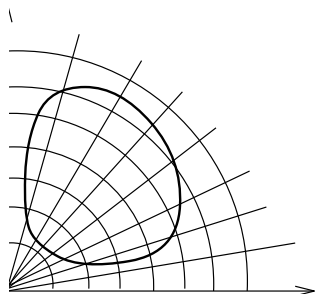
Ellipse

Example 56 The equation $r = \frac{2}{3 + \cos \theta}$ has as locus an ellipse with one focus at the origin. You can verify this by changing to rectangular coordinates as above.

$$\begin{aligned}
 r &= \frac{2}{3 + \frac{x}{r}} \\
 3r + x &= 2 && \text{by cross multiplying} \\
 3r &= 2 - x \\
 9r^2 &= (2 - x)^2 && \text{square—this might add some points} \\
 9(x^2 + y^2) &= 4 - 4x + x^2 \\
 8x^2 + 4x + 9y^2 &= 4 \\
 8\left(x^2 + \frac{1}{2}x + \frac{1}{16}\right) + 9y^2 &= \frac{9}{2} && \text{completing the square} \\
 8\left(x + \frac{1}{4}\right)^2 + 9y^2 &= \frac{9}{2} \\
 \frac{(x + 1/4)^2}{9/16} + \frac{y^2}{1/2} &= 1.
 \end{aligned}$$

The locus of the last equation is an ellipse centered at the point $(-1/4, 0)$, with semi-major axis $3/4$ and semi-minor axis $1/\sqrt{2}$. That the origin is one focus of the ellipse requires going into the details of the analytic geometry of ellipses which we won't explore here, but it is true. Note that in the step where we squared both sides, we might have added to the locus points where $r > 0$ but $2 - x < 0$. You should convince yourself that no point on the ellipse has this property.

Integration in Polar Coordinates To evaluate the integral $\iint_D f(x, y) dA$ in polar coordinates, we must use a dissection of the region which is appropriate for polar coordinates. In rectangular coordinates, the dissection into rectangles is obtained from a network of vertical lines, $x = \text{constant}$, and horizontal lines, $y = \text{constant}$. In polar coordinates the corresponding network would consist of concentric circles, $r = \text{constant}$, and rays from the origin, $\theta = \text{constant}$.



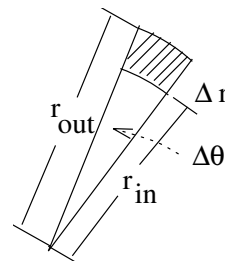
As indicated in the diagram, a typical element of area produced by such a network is bounded by two nearby circular arcs separated radially by Δr and two nearby rays separated angularly by $\Delta\theta$. Such an element of area is called a *polar rectangle*. To determine its area we argue as follows. The area of a circular wedge of radius r and subtending an angle $\Delta\theta$ at the center of the circle is $\frac{1}{2}r^2\Delta\theta$. (If you are not familiar with this formula, try to reason it out. The point is that the area of the wedge is to the area of the circle as $\Delta\theta$, the subtended angle, is to 2π .) Suppose now that the polar rectangle has inner radius r_{in} and outer radius r_{out} . Then the area of the polar rectangle is

$$\Delta A = \frac{1}{2}(r_{out}^2\Delta\theta - r_{in}^2\Delta\theta) = \frac{1}{2}(r_{out} + r_{in})(r_{out} - r_{in})\Delta\theta = \frac{r_{out} + r_{in}}{2}(\Delta r)\Delta\theta.$$

Now put $r = \frac{r_{out} + r_{in}}{2}$ (the average radius) to get

$$\Delta A = r\Delta r\Delta\theta = \Delta r(r\Delta\theta). \quad (51)$$

This formula is an exact equality, but note that if we use any other value of r falling within the range $r_{in} \leq r \leq r_{out}$, it will only make a slight difference in the answer if the polar rectangle is small enough.



Given a dissection of the region D into polar rectangles, we can form as before the sum

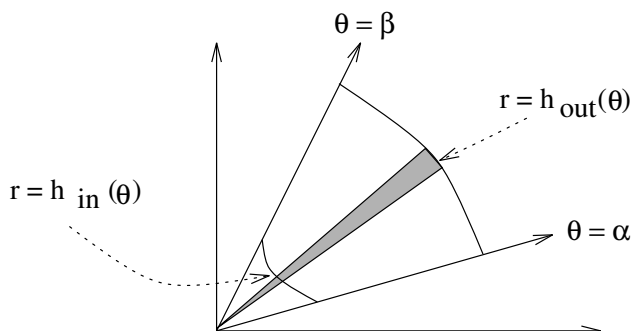
$$\sum_{\substack{\text{all polar} \\ \text{rectangles}}} f(x, y)\Delta A$$

where for each polar rectangle, (x, y) is a point somewhere inside the rectangle. Again, it doesn't matter much which point is chosen, so we can assume $r = \sqrt{x^2 + y^2}$ has the same value as the r in the formula $\Delta A = r\Delta r\Delta\theta$. If we now put $x = r\cos\theta$, $y = r\sin\theta$, the sum takes the form

$$\sum_{\substack{\text{all polar} \\ \text{rectangles}}} f(r\cos\theta, r\sin\theta)r\Delta r\Delta\theta. \quad (52)$$

It is fairly clear (and can even be proved with some effort) that the limit of this sum as the number of polar rectangles approaches ∞ (and as the size of each shrinks to zero) is the double integral $\iint_D f dA$.

Suppose now that the region is *bounded radially by graphs*. By this we mean that it lies between two rays $\theta = \alpha$ and $\theta = \beta$ (with $\alpha < \beta$), and for each θ it lies between two polar graphs: $r = h_{in}(\theta)$ on the inside and $r = h_{out}(\theta)$ on the outside.



For such a region, we can do the sum by adding first the contributions from those polar rectangles within a given *thin wedge*—say it is at position θ and has angular width $\Delta\theta$ —and then adding up the contributions from the wedges.

$$\begin{aligned} \sum_{\text{all polar rectangles}} f(r \cos \theta, r \sin \theta) r \Delta r \Delta \theta &= \sum_{\text{all wedges}} \sum_{\text{in a wedge}} f(r \cos \theta, r \sin \theta) r \Delta r \Delta \theta \\ &= \sum_{\text{all wedges}} \left(\sum_{\text{in a wedge}} f(r \cos \theta, r \sin \theta) r \Delta r \right) \Delta \theta. \end{aligned}$$

In the limit, the expression in parentheses

$$\sum_{\text{in a wedge}} f(r \cos \theta, r \sin \theta) r \Delta r \longrightarrow \int_{r=h_{in}(\theta)}^{r=h_{out}(\theta)} f(r \cos \theta, r \sin \theta) r \, dr = G(\theta).$$

As indicated, this is a partial integral depending on θ . Putting this back in (53), we obtain the approximation

$$\sum_{\text{all wedges}} G(\theta) \Delta \theta$$

which in the limit approaches

$$\int_{\alpha}^{\beta} G(\theta) \, d\theta.$$

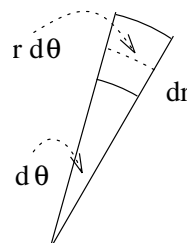
Thus, we get finally

$$\iint_D f \, dA = \int_{\alpha}^{\beta} \int_{r=h_{in}(\theta)}^{r=h_{out}(\theta)} f(r \cos \theta, r \sin \theta) r \, dr \, d\theta. \quad (54)$$

Note that the derivation of formula (54) suggests the symbolic rule

$$dA = r dr d\theta = dr (r d\theta).$$

r is to be thought of as a *correction factor* for changing the (false) ‘area’ $dr d\theta$ to the correct area dA . One way to think of this is that the polar rectangle is almost a true rectangle of dimensions dr by $rd\theta$.

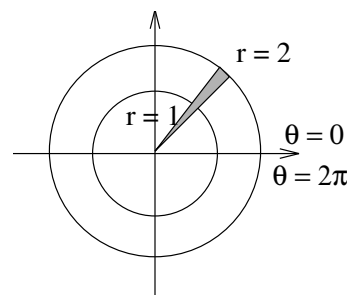


Example 57 We find $\iint_D 1/(x^2 + y^2) dA$ for D the region between the circles $x^2 + y^2 = 1$ and $x^2 + y^2 = 4$. Here,

$$f(x, y) = \frac{1}{x^2 + y^2} = \frac{1}{r^2},$$

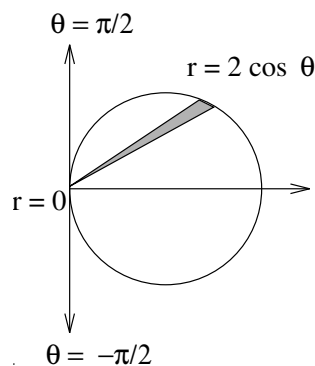
and D lies between the inner graph $r = 1$ and the outer graph $r = 2$. It also lies between two rays, but they happen to be the same ray described in two different ways: $\theta = \alpha = 0$ and $\theta = \beta = 2\pi$. Hence, the double integral is calculated by

$$\begin{aligned} \iint_D f dA &= \int_0^{2\pi} \int_{r=1}^{r=2} \frac{1}{r^2} r dr d\theta \\ &= \int_0^{2\pi} \int_{r=1}^{r=2} \frac{1}{r} dr d\theta \\ &= \int_0^{2\pi} (\ln r)|_1^2 d\theta \\ &= \int_0^{2\pi} \ln 2 d\theta = \ln 2 \theta|_0^{2\pi} \\ &= 2\pi \ln 2. \end{aligned}$$



Example 58 We find the *volume* under the cone $z = \sqrt{x^2 + y^2}$ and over the circular disk inside $r = 2 \cos \theta$. That is given by the double integral $\iint_D \sqrt{x^2 + y^2} dA$. In this case, $f(x, y) = r$, and we can describe the region as bounded by the rays $\theta = -\pi/2$ and $\theta = \pi/2$ and, for each θ , as lying between the inner graph $r = 0$ (i.e., the origin) and the outer graph $r = 2 \cos \theta$. The integral is

$$\begin{aligned} \int_{-\pi/2}^{\pi/2} \int_0^{2 \cos \theta} r r dr d\theta &= \int_{-\pi/2}^{\pi/2} \left. \frac{r^3}{3} \right|_0^{2 \cos \theta} d\theta \\ &= \int_{-\pi/2}^{\pi/2} \frac{8}{3} \cos^3 \theta d\theta \\ &= \frac{8}{3} \frac{4}{3} = \frac{32}{9}. \end{aligned}$$

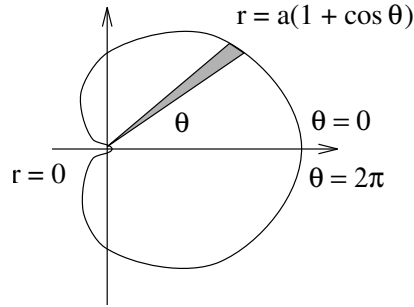


(The last integration was done by Mathematica.)

Example 59 We find the center of mass of the region inside the cardioid $r = a(1 + \cos \theta)$ assuming constant density δ . We may as well assume the density is

$\delta = 1$. (Can you see why?) Then the mass is the same as the area, which is

$$\begin{aligned}
 \iint_D 1 \, dA &= \int_0^{2\pi} \int_0^{a(1+\cos \theta)} 1 \, r \, dr \, d\theta \\
 &= \int_0^{2\pi} \left. \frac{r^2}{2} \right|_0^{a(1+\cos \theta)} d\theta \\
 &= \frac{a^2}{2} \int_0^{2\pi} (1 + 2\cos \theta + \cos^2 \theta) d\theta \\
 &= \frac{a^2}{2} (2\pi + 0 + \pi) = \frac{3\pi a^2}{2}.
 \end{aligned}$$



We used here the formulas

$$\begin{aligned}
 \int_0^{2\pi} \cos \theta \, d\theta &= 0 \\
 \int_0^{2\pi} \cos^2 \theta \, d\theta &= \pi
 \end{aligned}$$

both of which can be derived without integrating anything very complicated. (Do you know the appropriate tricks?)

The y -coordinate of the center of mass is 0 (by symmetry), and the x -coordinate is

$$\begin{aligned}
 \frac{1}{A} \iint_D x \, dA &= \frac{1}{A} \int_0^{2\pi} \int_0^{a(1+\cos \theta)} r \cos \theta \, r \, dr \, d\theta \\
 &= \frac{1}{A} \int_0^{2\pi} \left. \frac{r^3}{3} \right|_0^{a(1+\cos \theta)} \cos \theta \, d\theta \\
 &= \frac{a^3}{3A} \int_0^{2\pi} (\cos \theta + 3\cos^2 \theta + 3\cos^3 \theta + \cos^4 \theta) d\theta \\
 &= \frac{a^3}{3A} (0 + 3\pi + 0 + \frac{3\pi}{4}) = \frac{15\pi a^3}{12A} = \frac{15\pi a^3}{18\pi a^2} = \frac{5a}{6}.
 \end{aligned}$$

This used the additional rule

$$\int_0^{2\pi} \cos^4 \theta \, d\theta = \frac{3\pi}{4}$$

for which I know no simple derivation. You can hack it out, or preferably use a table or Mathematica.

It is always true that the center of mass of a distribution of *constant density* does not depend on the density. (δ appears as a factor in two places which cancel.) In that case, the center of mass is a purely geometric property of the region, and it is called the *centroid*. Occasionally, in polar coordinates one deals with a region which is *bounded angularly by graphs*. In that case, you would integrate first—the inner integral—with respect to θ and then with respect to r . The student is encouraged to work out what the corresponding iterated integral would look like.

Exercises for 4.3.

In the following problems, remember that by our conventions $r \geq 0$. Thus you should ignore any points where the equation would yield $r < 0$.

1. Sketch and identify the following graphs:

- (a) $r = 2$.
- (b) $\theta = \frac{3\pi}{2}$.
- (c) $r = 1 - 2 \sin \theta$.
- (d) $r = 3 \cos 2\theta$.
- (e) $r^2 = -2 \cos 2\theta$.
- (f) $r = \frac{3}{2 + \sin \theta}$.

2. Sketch the following regions and find their areas by double integration:

- (a) the disc inside the circle $r = 3$.
- (b) the disc inside the circle $r = 3 \sin \theta$.
- (c) the area outside the circle $r = 1$ and inside the circle $r = 2 \cos \theta$.
- (d) the area within the cardioid $r = 1 + \cos \theta$ and the circle $r = 1$.
- (e) the area within the lemniscate $r^2 = -3 \sin 2\theta$.

3. Find the volume under the given surface and above the region bounded by the given curve(s):

- (a) $z = x^2 + y^2$, $r = 2$.
- (b) $z = x^2 + y^2$, $r = -2 \cos \theta$.
- (c) $z = \sqrt{x^2 + y^2}$, $r = 4 \sin \theta$.

4. Calculate the following double integrals by switching to polar coordinates: (a)

$$\int_{-1}^1 \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} (x^2+y^2) dx dy \quad (b) \quad \int_0^2 \int_0^{\sqrt{4-x^2}} \frac{1}{\sqrt{x^2+y^2}} dy dx \quad (c) \quad \int_0^1 \int_x^{\sqrt{2-x^2}} e^{x^2+y^2} dy dx$$

(d) $\int_0^1 \int_0^x x dy dx$. (This is a bit silly to do in polar coordinates, but try it for the practice. Use the fact that $x = 1$ implies $r = \sec \theta$.)

5. Find the volume of the following objects.

(a) The region inside the sphere $x^2 + y^2 + z^2 = 4a^2$ and above the plane $z = a$.

(b) An ‘ice cream cone’ bounded above by the sphere of radius a centered at the origin and below by right circular cone $z = \sqrt{x^2 + y^2}$.

(c) The region below the paraboloid $z = 4 - x^2 - y^2$ and above the x, y -plane.

(d) The solid torus formed by revolving the vertical disk within $(x-b)^2 + z^2 = a^2$ about the z -axis. (Assume $a < b$.)

4.4 Triple Integrals

The theory of integration in space proceeds much as the theory in the plane. Suppose E is a bounded subset of \mathbf{R}^3 , and suppose $w = f(x, y, z) = f(\mathbf{r})$ describes a function defined on E . (‘ E is bounded’ means that E is contained in some *rectangular box*, $a \leq x \leq b, c \leq y \leq d, e \leq z \leq f$.) Usually, E will be a quite reasonable looking set; in particular its boundary will consist of a finite set of smooth surfaces. To define the integral, we proceed by analogy with the 2-dimensional case. First partition the region E into a collection of small *rectangular boxes or cells* through a lattice of closely spaced parallel planes. We employ three such families of planes, each perpendicular to one of the coordinate axes. A typical cell will be a box with dimensions $\Delta x, \Delta y$, and Δz and volume

$$\Delta V = \Delta x \Delta y \Delta z.$$

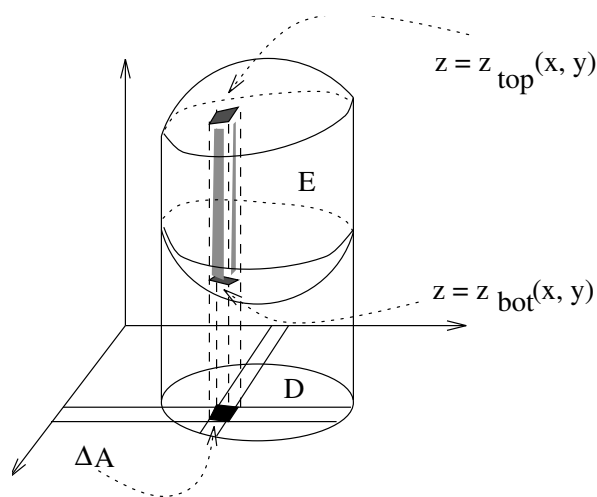
Choose a point (x, y, z) in such a cell, and form $f(x, y, z) \Delta V$, which will be the contribution from that cell. Now form the sum

$$\sum_{\text{all cells}} f(x, y, z) \Delta V$$

and let the number of cells go to ∞ while the size of each goes to zero. If the region and function are reasonably smooth, the sums will approach a definite limit which is called a *triple integral* and which is denoted

$$\iiint_E f(x, y, z) dV \quad \text{or} \quad \int_E f(x, y, z) dV.$$

Iterated Integrals in \mathbf{R}^3 There are many ways in which one could add up the terms in the above sum. We start with one that is appropriate if E is *bounded by graphs in the z -direction*. That is, E lies above the graph of a function $z = z_{\text{bot}}(x, y)$ and below the graph of another function $z = z_{\text{top}}(x, y)$. In addition, we assume that E is bounded on its sides by a ‘cylindrical surface’ consisting of perpendiculars to a closed curve γ in the x, y -plane. The region D in the x, y -plane bounded by γ is then the *projection* of E in the x, y -plane. The ‘cylinder’ is only cylindrical in a very general sense, and it may even happen that part or all of it consists of curves rather than surfaces. For example, consider the region between the cone $z = \sqrt{x^2 + y^2}$ (below) and the plane $z = 1$ (above). The ‘cylinder’ in this case is just a circle.



Suppose the region is dissected into cells as described above. There will be a corresponding dissection of the projected region D into small rectangles. Consider one such rectangle with area ΔA positioned at (x, y) in D , and consider the contribution from the cells forming the *column* lying over that rectangle. In the limit, this will approach an integral

$$\int_{z=z_{\text{bot}}(x,y)}^{z=z_{\text{top}}(x,y)} f(x, y, z) dz \Delta A = G(x, y) \Delta A.$$

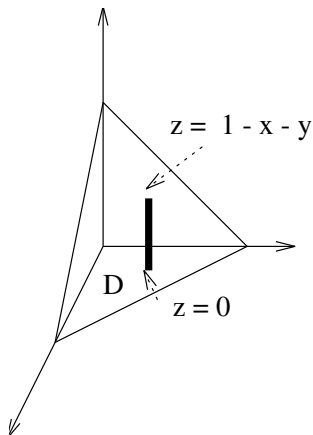
Here, we used the fact that all the cells in the column share a common base with area ΔA , so the volume of any such cell is $\Delta V = \Delta z \Delta A$ where Δz is its height. $G(x, y)$ is the *partial integral* obtained by integrating with respect to z , keeping x, y constant, and then evaluating at limits depending on x, y . Putting this in the sum, we have schematically

$$\sum_{\substack{\text{rectangles} \\ \text{in } D}} \sum_{\substack{\text{cells in} \\ \text{column}}} f(x, y, z) \Delta V \rightarrow \sum_{\substack{\text{rectangles} \\ \text{in } D}} G(x, y) \Delta A \rightarrow \iint_D G(x, y) dA.$$

Recalling what $G(x, y)$ is, we get the following formula for the triple integral.

$$\iint_E f(x, y, z) dV = \iint_D \left(\int_{z=z_{\text{bot}}(x, y)}^{z=z_{\text{top}}(x, y)} f(x, y, z) dz \right) dA$$

This in effect reduces the calculation of triple integrals to that of double integrals. The double integral can be done in any order that is appropriate. (It may even be done in polar coordinates!)



Example 60 We shall find the centroid (center of mass for constant density) of the solid region E in the first octant which lies beneath the plane $x + y + z = 1$. The solid E has four faces. It is an example of a *tetrahedron*. If we take the density to be 1, the mass will equal the volume. A tetrahedron is a special case of pyramid, and you should recall from high school that the volume of a pyramid is $1/3$ its height times the area of its base. In this case, we get $M = V = (1/3) \times (1) \times (1/2) = 1/6$. By symmetry it is clear that the three coordinates of the centroid are equal, so we need only find

$$x_{cm} = \frac{1}{V} \iiint_E x dV.$$

To evaluate the triple integral, note that E is z -bounded by graphs since it lies between $z = z_{\text{bot}}(x, y) = 0$ and $z = z_{\text{top}}(x, y) = 1 - x - y$. The projection of E in the x, y -plane is the triangular region D in the first quadrant, bounded by the line $x + y = 1$. Hence,

$$\begin{aligned} \iiint_E x dV &= \iint_D \int_{z=0}^{z=1-x-y} x dz dA \\ &= \iint_D x \int_{z=0}^{z=1-x-y} dz dA \\ &= \iint_D [x(1-x-y)] dA = \iint_D (x - x^2 - xy) dA. \end{aligned}$$

The problem has now been reduced to a double integral. It is best to treat this as a separate problem, redrawing if necessary a diagram of the region D . We can view D as bounded in the y -direction by the graphs $y = 0$ and $y = 1 - x$ with $0 \leq x \leq 1$. Thus,

$$\begin{aligned} \iint_D (x - x^2 - xy) dA &= \int_0^1 \int_{y=0}^{y=1-x} (x - x^2 - xy) dy dx \\ &= \int_0^1 (xy - x^2y - xy^2/2) \Big|_{y=0}^{y=1-x} dx \\ &= \int_0^1 (x(1-x) - x^2(1-x) - x(1-x)^2/2) dx \\ &= \int_0^1 (x^3/2 - x^2 + x/2) dx \\ &= x^4/8 - x^3/3 + x^2/4 \Big|_0^1 = \frac{1}{24}. \end{aligned}$$

It follows that

$$x_{cm} = \frac{1}{V} \iiint_E x \, dV = \frac{1/24}{1/6} = \frac{1}{4}.$$

Hence the centroid is at $(1/4, 1/4, 1/4)$.

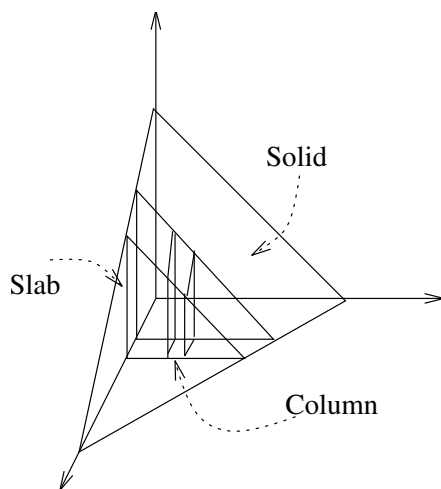
Note that if we had suppressed the evaluations temporarily, the triple integral above would appear as the following triply iterated integral

$$\begin{aligned} \iiint_E x \, dV &= \iint_D \int_{z=0}^{z=1-x-y} x \, dz \, dA \\ &= \int_0^1 \int_{y=0}^{y=1-x} \int_{z=0}^{z=1-x-y} x \, dz \, dy \, dx. \end{aligned}$$

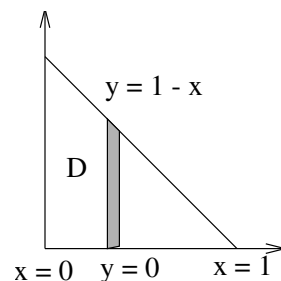
You should try to visualize the dissection of the solid region associated with each step in the iterated integral.

$$\underbrace{\int_0^1 \underbrace{\int_{y=0}^{y=1-x} \underbrace{\int_{z=0}^{z=1-x-y} x \, dz}_{\text{column}} \, dy}_{\text{slab}} \, dx}_{\text{solid}}.$$

First, we sum in the z -direction to include all cells in a *column*. Next, we sum in the y -direction to include all cells in a row of columns to form a *slab*. Finally, we sum in the x -direction to put these slabs together to form the entire solid region.



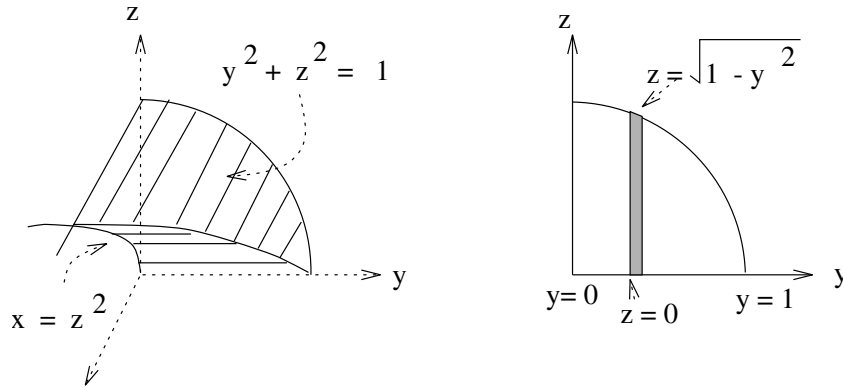
The above example was done in the order $dz \, dy \, dx$. There are in fact *six possible orders of integration in \mathbf{R}^3* . Which is appropriate depends on how the solid region and its projections in the coordinate planes are bounded by graphs.



Example 61 We find the volume in the first octant of the solid E bounded by the graphs $y^2 + z^2 = 1$ and $x = z^2$. The former surface is part of a right circular cylinder perpendicular to the y, z -plane. The latter is a cylinder (in the general sense) perpendicular to the x, z -plane. E is z -bounded by graphs, but it is not easy to visualize its projection in the x, y -plane. In this case, it would be better to project instead in the y, z -plane or the x, z -plane. Let's project in the y, z -plane, so E will be viewed as bounded in the x -direction by the graph $x = 0$ (behind) and $x = z^2$ (in front). The projection D of E in the y, z -plane is the quarter disc inside the circle $y^2 + z^2 = 1$.

$$\begin{aligned} V &= \iiint_E 1 \, dV = \iint_D \int_{x=0}^{x=z^2} 1 \, dx \, dA \\ &= \iint_D x \Big|_{x=0}^{x=z^2} \, dA \\ &= \iint_D z^2 \, dA. \end{aligned}$$

We now calculate the double integral in the y, z -plane by viewing D as bounded in the z -direction by the graphs $z = 0$ and $z = \sqrt{1 - y^2}$ with $0 \leq y \leq 1$.



$$\begin{aligned} \iint_D z^2 \, dA &= \int_0^1 \int_{z=0}^{z=\sqrt{1-y^2}} z^2 \, dz \, dy \\ &= \int_0^1 z^3/3 \Big|_{z=0}^{z=\sqrt{1-y^2}} \, dy \\ &= \frac{1}{3} \int_0^1 (1-y^2)^{3/2} \, dy = \frac{\pi}{16}. \end{aligned}$$

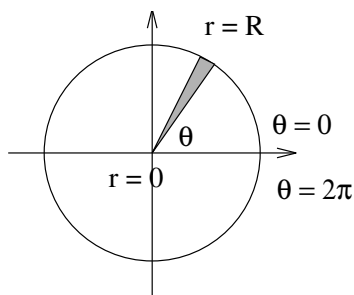
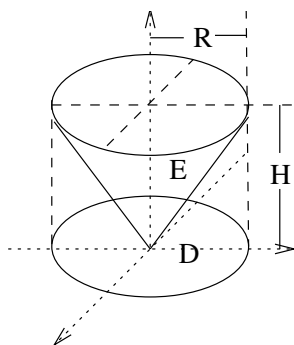
The last step was done by Mathematica.

You should try to do the same triple integral by viewing it as bounded in the y -direction by the graphs $y = 0$ and $y = \sqrt{1 - z^2}$ and projecting in the x, z -plane.

Sometimes it may be appropriate to do the double integral in polar coordinates.

Example 62 We shall find $\iiint_E z \, dV$ where E is the solid cone contained between the cone $z = \frac{H}{R} \sqrt{x^2 + y^2}$ and the plane $z = H$. This is a cone of height H and radius R . Its projection in the x, y plane is the region D inside the circle $x^2 + y^2 = R^2$.

$$\begin{aligned} \iiint_E z \, dV &= \iint_D \int_{z=(H/R)\sqrt{x^2+y^2}}^{z=H} z \, dz \, dA \\ &= \iint_D \left. \frac{z^2}{2} \right|_{z=(H/R)\sqrt{x^2+y^2}}^{z=H} dA \\ &= \frac{1}{2} \iint_D (H^2 - (H/R)^2(x^2 + y^2)) \, dA. \end{aligned}$$



We could of course do the double integral in rectangular coordinates. (The region D in the x, y -plane is bounded in the y -direction by $y = -\sqrt{R^2 - x^2}$ and $y = \sqrt{R^2 - x^2}$ with $-R \leq x \leq R$.) You should try to do it that way. It makes more sense, however, to use polar coordinates.

$$\begin{aligned} \frac{1}{2} \iint_D (H^2 - (H/R)^2(x^2 + y^2)) \, dA &= \frac{1}{2} \int_0^{2\pi} \int_{r=0}^{r=R} (H^2 - (H/R)^2 r^2) r \, dr \, d\theta \\ &= \frac{1}{2} \int_0^{2\pi} \left(H^2 r^2/2 - (H/R)^2 r^4/4 \right) \Big|_{r=0}^{r=R} d\theta \\ &= \frac{1}{2} \int_0^{2\pi} (H^2 R^2/2 - (H^2/R^2) R^4/4) d\theta \\ &= \frac{1}{2} \int_0^{2\pi} H^2 R^2/4 d\theta = \frac{1}{2} \frac{H^2 R^2}{4} 2\pi = \frac{\pi H^2 R^2}{4}. \end{aligned}$$

Note that you can do Example 61 this way if you are willing to introduce polar coordinates in the y, z -plane.

Exercises for 4.4.

1. Calculate the triple integral $\iiint_E f(x, y, z) dV$ for each of the following:
 - (a) $f(x, y, z) = 2x + 3y - z^2$, over the rectangular box with $0 \leq x \leq 4$, $0 \leq y \leq 2$, and $0 \leq z \leq 1$.
 - (b) $f(x, y, z) = yz \cos z$, over the cube with opposite vertices at the origin and (π, π, π) .
 - (c) $f(x, y, z) = xyz$, over the region inside the cone $z^2 = x^2 + y^2$ and between the planes $z = 0$ and $z = 9$.
 - (d) $f(x, y, z) = 2z - 4y^2$, over the region between the cylinders $z = x^2$ and $z = x^3$ and between the planes $y = -1$ and $y = 2$.
2. Consider the tetrahedron bounded by the plane $2x + 3y + z = 6$ and the coordinate planes. There are six possible orders of integration for the triple integral representing its volume. Write out iterated integrals for each of these and evaluate two of them.
3. Sketch the solid bounded by the given surfaces and find its volume by triple integration. If you see an easier way to find the same volume, use that to check your answer.
 - (a) $3x - 2y + 4z = -2$, $x = 0$, $y = 0$, $z = 0$.
 - (b) $y = x^2 + z^2$, $y = 4$.
 - (c) $x^2 + y^2 + z^2 = 1$, $x^2 + y^2 + z^2 = 4$.
 - (d) $z = 2x^2 + 3y^2$, $z = 5 - 3x^2 - 2y^2$. (Note that the two surfaces intersect in a circle of radius 1.)
4. Use a triple integral to find the volume of the solid in the first octant bounded by the cylinders $x^2 + y^2 = 1$ and $x^2 + z^2 = 1$. Do this both in the order ' $dz dy dx$ ' and in the order ' $dx dz dy$ '. (The latter order involves a complication you might not expect.)
5. Find the centroid for each of the following objects. (Take $\delta = 1$.)
 - (a) A solid hemisphere of radius a .
 - (b) A solid right cone of radius a and height h .
 - (c) The solid bounded by the paraboloid $z = x^2 + y^2$ and the plane $z = h > 0$.

4.5 Cylindrical Coordinates

In the previous section, we saw that we could switch to polar coordinates in the x, y -plane when doing a triple integral. This really amounts to introducing a new coordinate system in space called *cylindrical coordinates*. The cylindrical coordinates of a point in space with rectangular coordinates (x, y, z) are (r, θ, z) where (r, θ) are the polar coordinates of the projection (x, y) of the point in the x, y -plane. Just as with polar coordinates, we insist that $r \geq 0$, and usually θ is restricted to some range of size 2π . Moreover, the same formulas

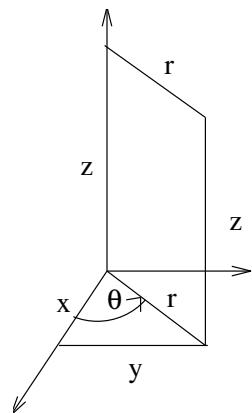
$$x = r \cos \theta$$

$$y = r \sin \theta$$

and

$$r = \sqrt{x^2 + y^2}$$

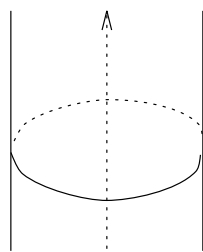
$$\tan \theta = \frac{y}{x} \quad \text{if } x \neq 0.$$



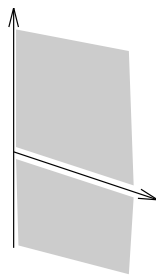
The geometric interpretations of r and θ in space are a bit different. r is the perpendicular distance of the point to the z axis (as well as being the distance of its projection to the origin). θ is the angle that the plane determined by the point and the z -axis makes with the positive x, z -plane.

You should learn to recognize certain important surfaces when described in cylindrical coordinates.

Example 63 $r = a$ describes an (infinite) cylinder of radius a centered on the z -axis. If we let a vary, we obtain an infinite family of concentric cylinders. We can treat the case $a = 0$ (the z -axis) as a degenerate cylinder of radius 0.



$r = a$ Cylinder

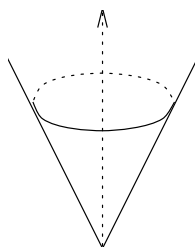


$\theta = \alpha$ Half plane

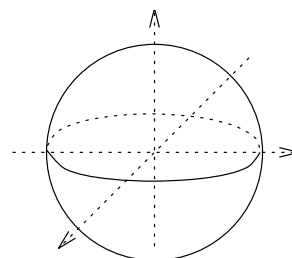
Example 64 $\theta = \alpha$ describes a *half plane* making angle α with the positive x, z -plane. Note that in this half plane, r can assume any non-negative value and z can assume any value.

Example 65 $z = mr$ describes an (infinite) cone centered on the z -axis with vertex at the origin. For a fixed value of θ , we obtain a ray in this cone which starts at the origin and extends to ∞ . This ray makes angle $\tan^{-1} m$ with the z -axis, and if we let θ vary, the ray rotates around the z -axis generating the cone. Note also that if $m > 0$, the angle with the z -axis is acute and the cone lies above the x, y -plane. If $m < 0$, the angle is obtuse, and the cone lies below the x, y -plane. The case $m = 0$ yields the x, y -plane ($z = 0$) which may be considered a special ‘cone’.

Note that in rectangular coordinates, $z = mr$ becomes $z = m\sqrt{x^2 + y^2}$.



$z = m r$ Cone



Sphere
 $r^2 + z^2 = a^2$

Example 66 $r^2 + z^2 = a^2$ describes a sphere of radius a centered at the origin. The easiest way to see this is to put $r^2 = x^2 + y^2$ whence the equation becomes $x^2 + y^2 + z^2 = a^2$. The top hemisphere of the sphere would be described by $z = \sqrt{a^2 - r^2}$ and the bottom hemisphere by $z = -\sqrt{a^2 - r^2}$.

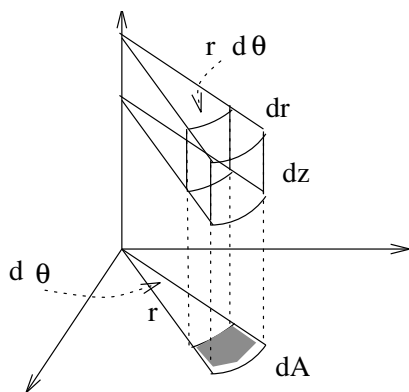
Integrals in Cylindrical Coordinates Suppose E is a solid region in \mathbf{R}^3 , and it is bounded in the z direction by graphs. If we use cylindrical coordinates directly (rather than switching to polar coordinates after the z integration), the triple integral would take the form

$$\iiint_E f(x, y, z) dV = \iint_D \int_{z=z_{\text{bot}}(r, \theta)}^{z=z_{\text{top}}(x, y)} f(r \cos \theta, r \sin \theta, z) dz dA$$

where the upper and lower graphs are expressed in cylindrical coordinates and the double integral over the region D should be done in polar coordinates. Symbolically, we have for $dA = r dr d\theta$, so we may also write

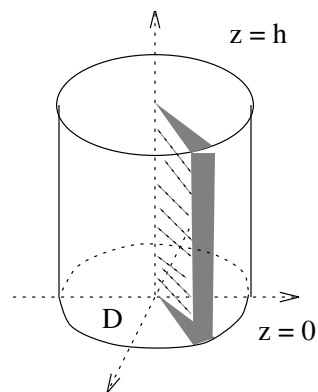
$$dV = dA dz = r dr d\theta dz = r dz dr d\theta$$

for the element of volume in cylindrical coordinates. Implicit in this is a dissection of the region into *cylindrical cells* as indicated in the diagram.



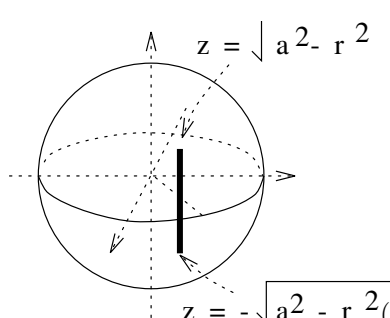
Example 67 We calculate $\iiint_E x^2 dV$ for E the solid region contained within the cylinder $x^2 + y^2 = a^2$ and between the planes $z = 0$ and $z = h$. Here $f(x, y, z) = x^2 = r^2 \cos^2 \theta$, and E is bounded in the z direction between $z = 0$ and $z = h$. The projection of D in the x, y -plane is the disc inside the circle $x^2 + y^2 = a^2$ (i.e., $r = a$). Thus,

$$\begin{aligned}
 \iiint_E x^2 dV &= \iint_D \int_{z=0}^{z=h} r^2 \cos^2 \theta dz dA \\
 &= \iint_D r^2 \cos^2 \theta z \Big|_{z=0}^{z=h} dA \\
 &= h \iint_D r^2 \cos^2 \theta dA \\
 &= h \int_0^{2\pi} \int_{r=0}^{r=a} r^2 \cos^2 \theta r dr d\theta \\
 &= h \int_0^{2\pi} \int_{r=0}^{r=a} r^3 \cos^2 \theta dr d\theta \\
 &= h \int_0^{2\pi} r^4/4 \Big|_{r=0}^{r=a} \cos^2 \theta d\theta \\
 &= h \frac{a^4}{4} \int_0^{2\pi} \cos^2 \theta d\theta = h \frac{a^4}{4} \pi = \frac{\pi a^4 h}{4}.
 \end{aligned}$$



Example 68 We shall find $\iiint_E (x^2 + y^2) dV$ for E a solid sphere of radius a centered at the origin. Here $f(x, y, z) = x^2 + y^2 = r^2$. Moreover, the surface of the sphere has equation $r^2 + z^2 = a^2$ in cylindrical coordinates, so the solid sphere may be viewed as lying between the graphs $z = -\sqrt{a^2 - r^2}$ below and $z = \sqrt{a^2 - r^2}$ above. The projection D in the x, y -plane is the disc inside the circle $r = a$. Hence, the

integral is



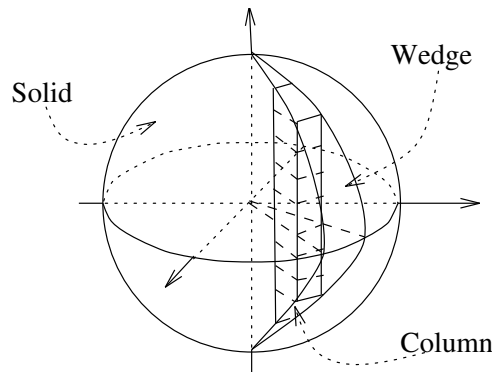
$$\begin{aligned}
 \iiint_E (x^2 + y^2) dV &= \iint_D \int_{z=-\sqrt{a^2-r^2}}^{z=\sqrt{a^2-r^2}} r^2 dz dA \\
 &= \iint_D r^2 \left(z \Big|_{z=-\sqrt{a^2-r^2}}^{z=\sqrt{a^2-r^2}} \right) dA \\
 &= \iint_D r^2 (2\sqrt{a^2-r^2}) dA \\
 &= \int_0^{2\pi} \int_{r=0}^{r=a} r^2 (2\sqrt{a^2-r^2}) r dr d\theta \\
 &= \int_0^{2\pi} \int_{r=0}^{r=a} r^3 (2\sqrt{a^2-r^2}) dr d\theta = 2\pi \frac{4a^5}{15} = \frac{8\pi a^5}{15}.
 \end{aligned}$$

(The last step was done by Mathematica.)

There is no need to first reduce to a double integral. We could have written out the triply iterated integral directly.

$$\underbrace{\int_0^{2\pi} \underbrace{\int_{r=0}^{r=a} \underbrace{\int_{z=-\sqrt{a^2-r^2}}^{z=\sqrt{a^2-r^2}} r^2 r dz}_{\text{column}} dr}_{\text{wedge}} d\theta}_{\text{solid}}.$$

The order of integration suggests a dissection of the sphere. The first integration with respect to z (r, θ fixed) adds up the contributions from cells in a *column*. The second integration with respect to r (θ fixed) adds up the contributions from columns forming a *wedge*. The final integration with respect to θ adds up the contributions from the wedges to form the solid sphere.

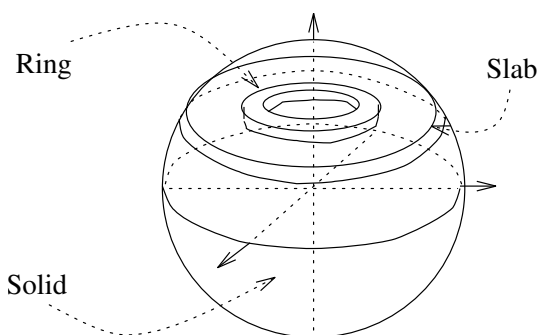


It is sometimes worthwhile doing the summation in some other order. For example,

consider the order

$$\underbrace{\int_{z=-a}^{z=a} \underbrace{\int_{r=0}^{r=\sqrt{a^2-z^2}} \underbrace{\int_0^{2\pi} r^2 r \, d\theta}_{\text{ring}} \, dr}_{\text{slab}} \, dz}_{\text{solid}}.$$

The first integration with respect to θ (r, z fixed) adds up the contribution from all cells in a *ring* at height z above the x, y -plane and distance r from the z -axis. The next integration with respect to r (z fixed) adds up the contributions from all rings at height z which form a *circular slab* of radius $\sqrt{a^2 - z^2}$. Finally, the last integration adds up the contributions from all the slabs as z ranges from $z = -a$ to $z = a$.



You should try the integration in this order to see if it is easier. You should also try to visualize the dissection for the order $dr, dz, d\theta$.

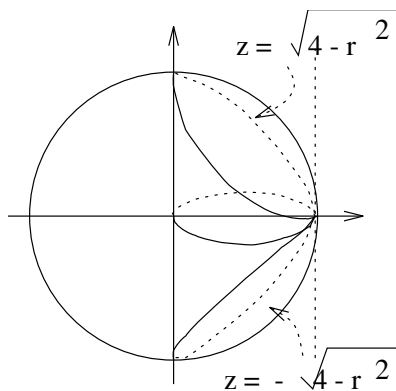
Example 69 We shall find the volume of the solid region E inside both the sphere $x^2 + y^2 + z^2 = 4$ and the cylinder $r = 2 \cos \theta$. Recall that the second equation describes a circle in the x, y -plane of radius 1 and centered at $(1, 0)$. However, in space, it describes the cylinder *perpendicular* to that circle. The appropriate range

for θ is $-\pi/2 \leq \theta \leq \pi/2$. The volume is

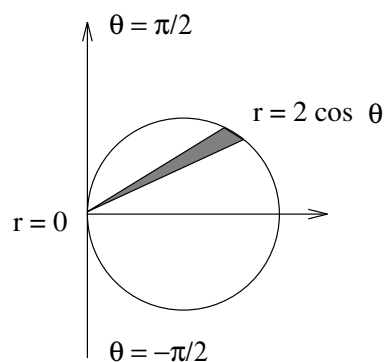
$$\begin{aligned}
 \iiint_E 1 \, dV &= \int_{-\pi/2}^{\pi/2} \int_{r=0}^{r=2 \cos \theta} \int_{z=-\sqrt{4-r^2}}^{z=\sqrt{4-r^2}} 1 \, dz \, r \, dr \, d\theta \\
 &= \int_{-\pi/2}^{\pi/2} \int_{r=0}^{r=2 \cos \theta} z \Big|_{z=-\sqrt{4-r^2}}^{z=\sqrt{4-r^2}} r \, dr \, d\theta \\
 &= \int_{-\pi/2}^{\pi/2} \int_{r=0}^{r=2 \cos \theta} 2\sqrt{4-r^2} \, r \, dr \, d\theta \\
 &= \int_{-\pi/2}^{\pi/2} \left(-\frac{(4-r^2)^{3/2}}{3/2} \right) \Big|_{r=0}^{r=2 \cos \theta} d\theta \\
 &= \frac{2}{3} \int_{-\pi/2}^{\pi/2} (8 - 8|\sin \theta|^3) \, d\theta.
 \end{aligned}$$

Here we used $1 - \cos^2 \theta = \sin^2 \theta$ and the fact that $\sqrt{\sin^2 \theta} = |\sin \theta|$, not $\sin \theta$. This would cause a problem in integration over the range $-\pi/2 \leq \theta \leq \pi/2$, so we get around it by integrating over $0 \leq \theta \leq \pi/2$ and doubling the result. We get

$$\frac{16}{3} 2 \int_0^{\pi/2} (1 - \sin^3 \theta) \, d\theta = \frac{32}{3} \left(\frac{\pi}{2} - \frac{2}{3} \right) = \frac{16(3\pi - 4)}{9}.$$



Off center cylinder intersecting sphere



Projected region in plane

Moments of Inertia In the Example 68, we calculated $\iiint_E r^2 \, dV$ where r is the perpendicular distance to the z -axis. This is a special case of the concept of *moment of inertia*. In physics, the moment of inertia of a finite set of points about an axis L is defined to be

$$I_L = \sum_i m_i r_i^2$$

where m_i is the mass of the i th particle and r_i is its perpendicular distance to the axis. The generalization for a mass distribution of density δ is

$$I_L = \iiint_E r^2 dm = \int_E r^2 \delta dV.$$

Here r is the distance of a point inside the solid region E to the axis L . We often choose the coordinate system so the z -axis lies along L . The density $\delta = \delta(x, y, z)$ can generally be variable. Moments of inertia are very important in the study of rotational motion of rigid bodies.

Example 68, (revisited) For a mass of constant density δ distributed over a solid sphere of radius a , the moment of inertia about a diameter (which we can take to be the z -axis) is

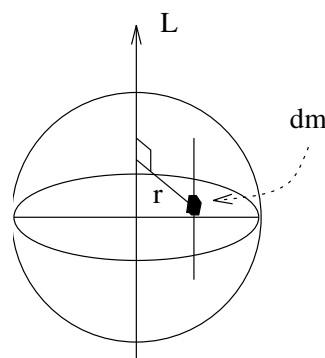
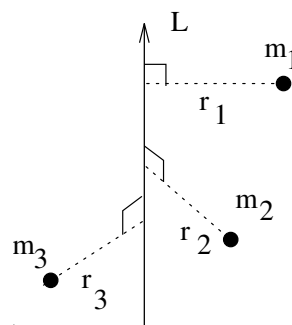
$$I_z = \iiint_E r^2 \delta dV = \delta \iiint_E r^2 dV = \frac{8\pi a^5 \delta}{15}.$$

However, the total mass in the sphere will be

$$M = V\delta = \frac{4\pi a^3}{3}\delta$$

so the moment of inertia may be rewritten

$$I_z = \frac{2}{5}Ma^2.$$



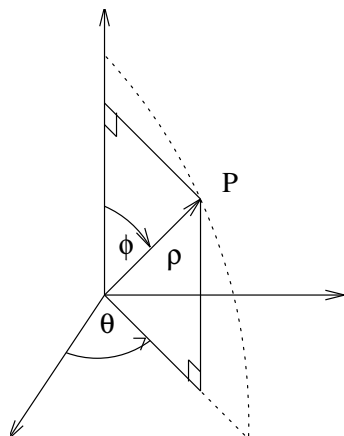
Exercises for 4.5.

- Find cylindrical coordinates for the point with the given rectangular coordinates:
 - $P(0, 0, 5)$.
 - $P(-1, 3, 2)$.
 - $P(3, 2, 0)$.
 - $P(-1, 1, -1)$.
 - $P(0, 4, -7)$.
- Identify the following graphs from their equations in cylindrical coordinates:
 - $r = 3$.
 - $\theta = \frac{\pi}{2}$.
 - $r = 2 \cos \theta$.
 - $\sin \theta + \cos \theta = 0$.
 - $z = 5 + r^2$.
- Convert the following equations to cylindrical coordinates:
 - $x^2 + y^2 + z^2 = 4$.
 - $z = x^2 + y^2$.
 - $3x + 2y - z = 0$.
 - $y = 4$.

4. A sphere of radius a centered at the origin has a cylindrical hole of radius $a/2$, centered on the z -axis drilled in it. Describe the solid region left by inequalities involving cylindrical coordinates.
5. Find the following volumes by using triple integration in cylindrical coordinates. The region
 - (a) inside both the sphere $x^2 + y^2 + z^2 = 9$ and the cylinder $x^2 + y^2 = 1$,
 - (b) inside a sphere of radius a ,
 - (c) between the paraboloid $z = 3 - x^2 - y^2$ and the plane $z = 0$,
 - (d) above the paraboloid $z = r^2$ and under the plane $z = x$,
 - (e) inside both the paraboloids $z = 9 - x^2 - y^2$ and $z = r^2$,
 - (f) inside the right circular cone $z = (a/2)r$ and under the plane $z = a$.
6. Find the centroid of a solid hemisphere of radius a using cylindrical coordinates.
7. Find the volume above the x, y -plane, inside the cylinder $r = 2 \sin \theta$ and under the plane $z = y$.
8. Find the volume bounded by the planes $z = y$, $z = 0$, $z = x$, $z = -x$, $y = 1$. First find the volume in rectangular coordinates. Then find the volume in cylindrical coordinates. (This is just for practice. Ordinarily it would be silly to use cylindrical coordinates for such a problem.)

4.6 Spherical Coordinates

Cylindrical coordinates are one way to generalize polar coordinates to space, but there is another way that is more useful in problems exhibiting spherical symmetry. We suppose as usual that a rectangular coordinate system has been chosen. The *spherical coordinates* (ρ, ϕ, θ) of a point P in space are defined as follows. ρ is the distance $|\overrightarrow{OP}|$ of the point to the origin. It is always non-negative, and it should be distinguished from the cylindrical coordinate r which is the distance from the z -axis. The *azimuthal angle* ϕ is the angle between \overrightarrow{OP} and the positive z -axis. ϕ is always assumed to lie between 0 and π . Finally, the *longitudinal angle* θ is the same as the cylindrical coordinate θ . θ is assumed to range over some interval of size 2π , e.g., $0 \leq \theta < 2\pi$. Note the reason for the range chosen for ϕ . Fix ρ and θ . If $\phi = 0$, the point is on the positive z -axis, and as ϕ increases, the point swings down toward the negative z -axis, but it stays in the half plane determined by that value of θ . For $\phi = \pi$, the point is on the negative z -axis, but if we allow ϕ to increase further, the point swings into the *opposite* half plane with longitudinal angle $\theta + \pi$. Such points can be obtained just as well by swinging down from the positive z -axis in the opposite half plane determined by $\theta + \pi$.



The following relationships hold between spherical coordinates, cylindrical coordinates, and spherical coordinates. Refer to the diagram

$$r = \rho \sin \phi$$

$$z = \rho \cos \phi$$

so

$$x = \rho \sin \phi \cos \theta$$

$$y = \rho \sin \phi \sin \theta$$

$$z = \rho \cos \phi$$

and

$$\rho = \sqrt{r^2 + z^2} = \sqrt{x^2 + y^2 + z^2}$$

$$\tan \phi = \frac{r}{z} \quad \text{if } z \neq 0.$$

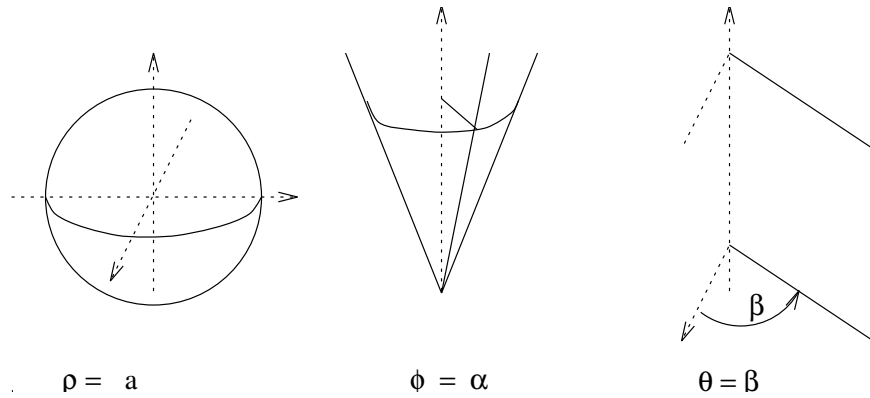
One may think of the spherical coordinates (ρ, ϕ) as polar coordinates in the *half plane* determined by fixing θ . However, because of the restrictions on ϕ , this is not quite the same as polar coordinates in the x, y -plane.

You should learn to recognize certain important surfaces when described in spherical coordinates.

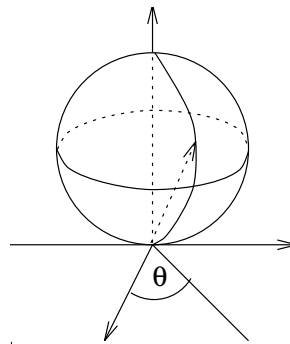
Example 70 $\rho = a$ describes a sphere of radius a centered at the origin.

Example 71 $\phi = \alpha$ describes a cone making angle α with the positive z -axis. The cone can lie above or below the x, y -plane, and $\phi = \pi/2$ describes the x, y -plane.

Example 72 $\theta = \beta$ describes a half plane starting from the z -axis as before.



Example 73 $\rho = 2a \cos \phi$ describes a sphere of radius a centered at $(0, 0, a)$. You can see this by fixing attention on the half plane determined by fixing θ . In that half plane, the locus is the *semi-circle* with the given radius and center. If we then let θ vary, the effect is to rotate the semi-circle about the z -axis and generate the sphere

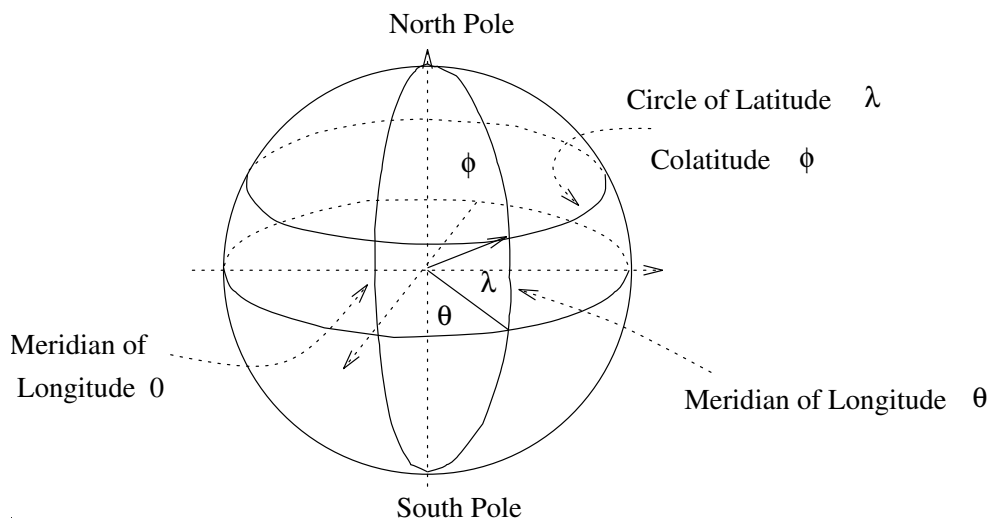


Geometry on the Surface of a Sphere If we fix $\rho = a$, we obtain a sphere of radius a . Then (ϕ, θ) specify the position of a point on that sphere.

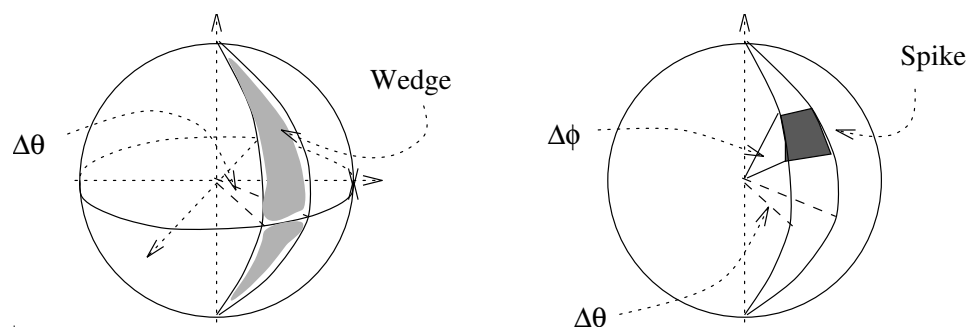
For $\theta = \text{constant}$, we obtain the semi-circle which is the intersection of the half plane for that θ with the sphere. That circle is called a *meridian of longitude*. This is exactly the concept of longitude used to measure position on the surface of the Earth, except that we use radians instead of degrees. Earth longitude is usually measured in degrees east or west of the Greenwich Meridian. That corresponds in our case to the positive and negative directions from the 0-meridian.

For $\phi = \text{constant}$, we obtain the circle which is the intersection of the cone for that ϕ with the sphere. Such circles are called *circles of latitude*. ϕ is related to the notion of latitude on the surface of the Earth, except that the latter is an

angle λ measured in degrees north or south of the *equatorial plane*. The spherical coordinate ϕ is sometimes called *co-latitude*, and we have $\phi = \pi/2 - \lambda$, if both are measured in radians. The unique point with $\phi = 0$ is called the *north pole*, that with $\phi = \pi$ is called the *south pole*, and at the poles θ is not well defined.

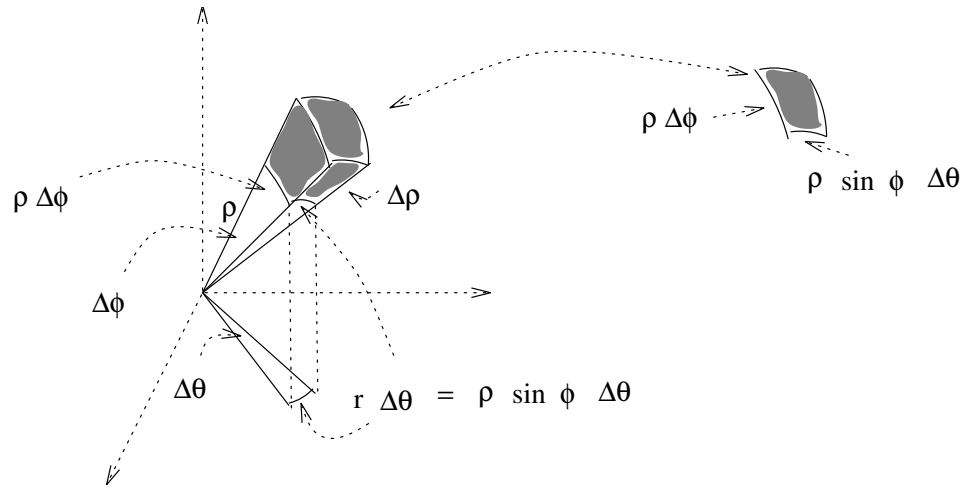


Integrals in Spherical Coordinates We want to evaluate triple integrals $\iiint_E f(x, y, z) dV$ using spherical coordinates. The most common order of integration for spherical coordinates is—from the inside out— $d\rho, d\phi, d\theta$. As before, this is associated with a certain dissection of the solid region E into *spherical cells*. To see what these cells look like, we describe the dissection of the region in the reverse order. First, assume $\alpha \leq \theta \leq \beta$. In this range, decompose the solid into *wedges* formed by a family of half planes emanating from the z -axis. Let $\Delta\theta$ be the angle subtended at the z -axis for the wedge at longitudinal angle θ .



In that wedge, assume $\phi_1(\theta) \leq \phi \leq \phi_2(\theta)$, where the extreme values of ϕ depend in general on θ . Decompose the wedge into *spikes* formed by a family of conical surfaces for different (constant) values of ϕ . Let $\Delta\phi$ be the angle subtended at the origin by the spike at azimuthal angle ϕ .

Finally, assume for that spike that $\rho_1(\phi, \theta) \leq \rho \leq \rho_2(\phi, \theta)$ where the extreme values of ρ depend generally on ϕ and θ . Decompose the spike into *spherical cells* by a family of concentric spherical surfaces. Let $\Delta\rho$ be the radial extension of the cell at radius ρ .



Note that the ‘base’ of this spherical cell is a spherical ‘rectangle’ on the sphere of radius ρ . Two of its sides lie along meridians of longitude, and the length of each of these sides is $\rho\Delta\theta$. The other two sides are circles of latitude. The top circle of latitude has radius $r = \rho \sin \theta$, and if everything is small enough the bottom circle has only a slightly larger radius. The arc which is the top side of the spherical rectangle subtends angle $\Delta\theta$ at the center of the circle of latitude, so its length is $r\Delta\theta = \rho \sin \phi \Delta\theta$. It is not hard to see from this that the area of the spherical rectangle is approximately $\rho\Delta\phi \cdot \rho \sin \phi \Delta\theta = \rho^2 \sin \phi \Delta\phi \Delta\theta$. Multiplying by $\Delta\rho$, we have the following approximate formula for the volume of a spherical cell

$$\Delta V = \rho^2 \sin \phi \Delta\rho \Delta\phi \Delta\theta.$$

(This can be made an exact formula if we use appropriate values of ρ and ϕ inside

the cell instead of the values at one corner.) The iterated integral is

$$\begin{aligned} \iiint_E f(x, y, z) dV \\ = \int_{\alpha}^{\beta} \int_{\phi_1(\theta)}^{\phi_2(\theta)} \underbrace{\int_{\rho_1(\phi, \theta)}^{\rho_2(\phi, \theta)} f(\rho \sin \phi \cos \theta, \rho \sin \phi \sin \theta, \rho \cos \phi) \rho^2 \sin \phi d\rho}_{\text{spike}} d\phi}_{\text{wedge}} d\theta. \end{aligned}$$

solid

Symbolically, we may write

$$dV = \underbrace{\rho^2 \sin \phi}_{\text{correction factor}} d\rho d\phi d\theta.$$

Example 74 We shall evaluate $\iiint_E (x^2 + y^2) dV$ for E a solid sphere of radius a centered at the origin. (This was done in the previous section in cylindrical coordinates.) Here $f(x, y, z) = x^2 + y^2 = r^2 = \rho^2 \sin^2 \phi$. To generate the entire sphere, we let $0 \leq \theta \leq 2\pi$. For each θ , to generate a wedge, we let $0 \leq \phi \leq \pi$. Finally, for each ϕ, θ , to generate a spike, we let $0 \leq \rho \leq a$. The integral is

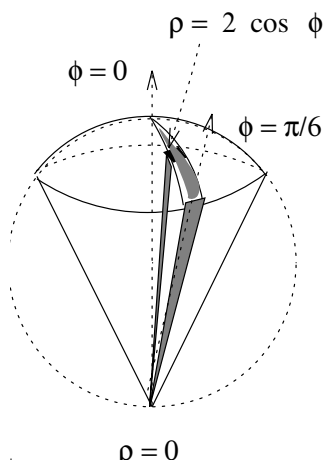
$$\begin{aligned} \iiint_E r^2 dV &= \int_0^{2\pi} \int_{\phi=0}^{\phi=\pi} \int_{\rho=0}^{\rho=a} \rho^2 \sin^2 \phi \rho^2 \sin \phi d\rho d\phi d\theta \\ &= \int_0^{2\pi} \int_{\phi=0}^{\phi=\pi} \int_{\rho=0}^{\rho=a} \rho^4 \sin^3 \phi d\rho d\phi d\theta \\ &= \frac{a^5}{5} \int_0^{2\pi} \int_{\phi=0}^{\phi=\pi} \sin^3 \phi d\phi d\theta \\ &= \frac{a^5}{5} \frac{4}{3} \int_0^{2\pi} d\theta \\ &= \frac{a^5}{5} \frac{4}{3} 2\pi = \frac{8\pi a^5}{15}. \end{aligned}$$

Note that it was not at all apparent whether this problem would be easier to solve in cylindrical or in spherical coordinates. The fact that we were integrating r^2 suggested the former but the fact that the region is a sphere suggested the latter. It turned out that the integral was a trifle easier in spherical coordinates, but there wasn't much difference.

Example 75 We shall find the volume bounded by the cone $z = \sqrt{3}r$ and the sphere $r^2 + (z - 1)^2 = 1$. Recall from Example 73 that the sphere may be described in spherical coordinates by $\rho = 2 \cos \phi$. The cone makes angle α with the positive z -axis where $\tan \alpha = r/z = 1/\sqrt{3}$. Hence, the cone is described in spherical coordinates by $\phi = \alpha = \pi/6$. To generate the solid, let $0 \leq \theta \leq 2\pi$, for each θ , let $0 \leq \phi \leq \pi/6$,

and for each ϕ, θ , let $0 \leq \rho \leq 2 \cos \phi$. Thus, the volume is given by

$$\begin{aligned} \iiint_E 1 \, dV &= \int_0^{2\pi} \int_0^{\pi/6} \int_0^{2 \cos \phi} \rho^2 \sin \phi \, d\rho \, d\phi \, d\theta \\ &= \frac{1}{3} \int_0^{2\pi} \int_0^{\pi/6} 8 \cos^3 \phi \sin \phi \, d\phi \, d\theta \\ &= \frac{8}{3} \int_0^{2\pi} (-\cos^4 \phi) \Big|_0^{\pi/6} d\theta \\ &= \frac{2}{3} \int_0^{2\pi} (1 - (\sqrt{3}/2)^4) d\theta \\ &= \frac{2}{3} \frac{7}{16} 2\pi = \frac{7\pi}{12}. \end{aligned}$$



There are other possible orders of integration in spherical coordinates, and you should try to visualize some of them. For example, suppose the region E is a solid sphere centered at the origin. The order $d\theta, d\phi, d\rho$ is associated with the following dissection. The sphere is first dissected into spherical shells of thickness $d\rho$. Then each shell is dissected into ‘rings’ at different latitudes subtending angle $d\phi$ at the center of the sphere. Finally, each ring is dissected into spherical cells as before each subtending angle $d\theta$ on the z -axis.

Other Notation Unfortunately there is no universally accepted notation for spherical coordinates. First of all, $\rho = |\mathbf{r}|$ is just the magnitude of the position vector $\mathbf{r} = \overrightarrow{OP}$, and another common notation for $|\mathbf{r}|$ is r , which we have reserved for the cylindrical coordinate. Secondly, many texts reverse the meanings of ϕ and θ . Indeed, almost all physics books and most mathematics books—except for calculus books—use θ to denote the azimuthal angle and ϕ for the longitudinal angle. Because of this inconsistency, you should be sure you check the meanings of the symbols whenever you encounter these coordinate systems. In any case, you should concentrate on the geometric and physical meaning of the concepts rather than the symbols used to represent them.

Exercises for 4.6.

- Find spherical coordinates for the points with the given rectangular coordinates:
 - $P(0, 0, -2)$.
 - $P(-1, 0, -1)$.
 - $P(2, -3, 5)$.
 - $P(-2, 0, 1)$.
- Identify the following graphs given in spherical coordinates:

- (a) $\phi = \pi/2$.
 - (b) $\rho = 2$.
 - (c) $\rho = 2 \sin \phi$.
 - (d) $\theta = \pi$.
 - (e) $\rho \sin \phi = 1$.
3. Write equations in spherical coordinates for each of the following:
- (a) $x^2 + y^2 + z^2 = 25$.
 - (b) $x + y + z = 1$.
 - (c) $z^2 = x^2 + y^2$.
 - (d) $z = 4 - x^2 - y^2$.
4. A sphere of radius a centered at the origin has a cylindrical hole of radius $a/2$, centered on the z -axis drilled in it. Describe the solid region that remains by inequalities in spherical coordinates.
5. Two points on the surface of a sphere of radius R have co-latitude and longitude (ϕ_1, θ_1) and (ϕ_2, θ_2) respectively. Show that the great circle distance between them is $R\alpha$ where

$$\cos \alpha = \sin \phi_1 \sin \phi_2 (\cos(\theta_1 - \theta_2)) + \cos \phi_1 \cos \phi_2.$$

- (The great circle determined by two points on a sphere is the circle of intersection of the sphere and the plane determined by the two points and the center of the sphere.) Use this fact to determine the great circle distance from your home town to London, given that London is at 51.5 degrees north latitude (co-latitude 38.5 degrees) and 0 degrees longitude. (If you don't know the latitude and longitude of your home town, look it up.)
6. Find the mass and centroid of a solid hemisphere of radius a if the density varies proportionally to the distance from the center, i.e. $\delta = k\rho$.
7. Find the volume of the intersection of the cylinder $r = 1$ with the sphere $\rho = 2$. Find the mass and center of mass if $\delta = \rho^2$.
8. Find the mass and center of mass of the 'ice cream cone' between the cone $\phi = \pi/6$ and the sphere $\rho = 2$, if $\delta = 2\rho$.
9. Find the moment of inertia of the right circular cylinder given by $0 \leq r \leq 1$ and $0 \leq z \leq \sqrt{3}$ about the z -axis assuming constant density $\delta = 1$. Use *spherical coordinates* even though it would be more natural to use cylindrical coordinates. Hint: You will have to divide the region into two parts.
10. Find the volume left in a sphere of radius a centered at the origin if a cylindrical hole of radius $a/2$ centered on the z -axis is drilled out. Do the problem in both cylindrical and spherical coordinates.

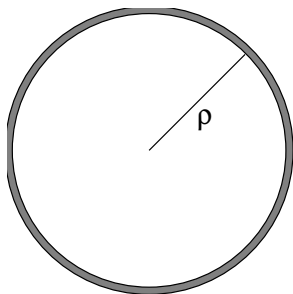
4.7 Two Applications

We illustrate the use of integration in spherical coordinates by giving two historically important applications.

Olbers' Paradox Olbers' Paradox is the 19th century observation that, in an infinite Newtonian universe in which stars are uniformly distributed and which has always existed, the sky would not be dark at night.

The argument for the paradox goes as follows. Assume stars are uniformly distributed through space. (Although stars are discrete objects, the model assumes that on a large scale, we may assume a uniform continuous mass distribution as an approximation. Today, we would replace 'star' by 'galaxy' as the basic unit.) Choose a coordinate system centered on our solar system. Since light intensity follows the *inverse square law*, the stars in a cell dV at distance ρ , would produce intensity proportional to dV/ρ^2 . Choosing our units properly, we would obtain for all the stars in a large sphere of radius R the light intensity

$$\begin{aligned} I &= \iiint_E \frac{1}{\rho^2} dV = \int_0^{2\pi} \int_0^\pi \int_0^R \frac{1}{\rho^2} \rho^2 \sin \phi \, d\rho \, d\phi \, d\theta \\ &= \int_0^{2\pi} \int_0^\pi \int_0^R \sin \phi \, d\rho \, d\phi \, d\theta \\ &= R \int_0^{2\pi} \int_0^\pi \sin \phi \, d\phi \, d\theta \\ &= R \int_0^{2\pi} -\cos \phi \Big|_0^\pi \, d\theta \\ &= R 2 \int_0^{2\pi} d\theta = 4\pi R. \end{aligned}$$



This is unbounded as $R \rightarrow \infty$, whence the conclusion that the sky would not be dark at night. Of course, there are lots of objections to this simple model, but the paradox persists even if one attempts to be more realistic. The resolution of the paradox had to await modern cosmology with its model of a universe expanding from an initial 'big bang'. We won't go into this here, referring you instead to any good book on cosmology.

The usual derivation of the paradox does not explicitly mention spherical coordinates. The argument is that the intensity due to all the stars in a thin shell at distance ρ will be proportional to the product of the area of the shell, $4\pi\rho^2$, with $1/\rho^2$; hence it will be proportional to 4π . In other words, the contribution from each spherical shell is the same and independent of the radius of the shell. If the contributions from all shells in the universe are added up, the result is infinite. You should convince yourself that this is just the same argument in other language.

The Gravitational Attraction of a Solid Sphere Newton discovered his laws of motion and the inverse square law for gravitational attraction about 1665, when

he was quite young, but he waited until 1686 to start his famous *Principia* in which these laws are expounded. Some scholars think the reason is that he was stumped by the problem of showing that the gravitational attraction of a solid sphere on a particle outside the sphere is the same as if the entire mass of the sphere were concentrated at the center. (However, according to the Encyclopedia Britannica, most authorities reject this explanation, thinking instead that he did not have an accurate enough value for the radius of the Earth.) We shall show how to solve that problem using spherical coordinates.

Let a mass M be distributed over a solid sphere of radius a in such a way that the density $\delta = \delta(\rho)$ depends only on the distance ρ to the center of the sphere. Let a test particle of unit mass be located at a point P at distance R from its center, and suppose $R > a$, i.e., P is outside the sphere. Choose the coordinate system so that the origin is at the center of the sphere and so that the z -axis passes through P . We can resolve the force \mathbf{F} exerted on the test particle into components $\langle F_x, F_y, F_z \rangle$, but it is clear by symmetry considerations that $F_x = F_y = 0$, so $\mathbf{F} = F_z \mathbf{k}$ is directed toward the origin. Thus we need only find F_z . Let dV be a small element of volume located at a point inside the sphere with spherical coordinates (ρ, ϕ, θ) . The mass inside dV will be $dm = \delta dV$, and according to the law of gravitation, the force on the test particle will have magnitude $G dm/s^2$, where s is the distance from P to dV . This force will be directed toward dV , but its z -component will be given by

$$dF_z = -\frac{G\delta dV}{s^2} \cos \eta \quad (56)$$

where η is the angle between the vector from P to dV and the z -axis. (See the diagram.)

We calculate the total z -component by *integrating* this over the solid sphere E .

$$F_z = -G \iiint_E \frac{\delta}{s^2} \cos \eta dV. \quad (57)$$

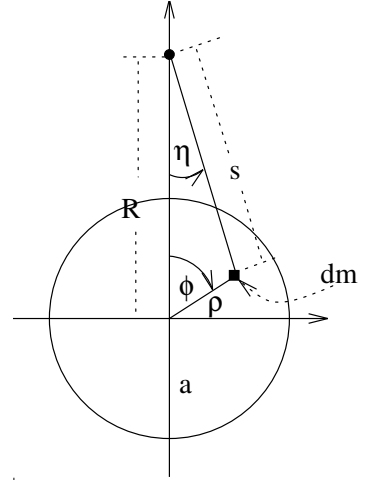
We shall compute the integral by integrating in spherical coordinates in the order $d\theta, d\phi, d\rho$.

$$F_z = -G \int_0^a \int_0^\pi \int_0^{2\pi} \frac{\delta}{s^2} \cos \eta \rho^2 \sin \phi d\theta d\phi d\rho.$$

The first integration with respect to θ is easy since nothing in the integrand depends on θ . It just yields a factor 2π which may be moved in front of the integral signs.

$$\begin{aligned} F_z &= -G(2\pi) \int_0^a \int_0^\pi \frac{\delta}{s^2} \cos \eta \rho^2 \sin \phi d\phi d\rho \\ &= -2\pi G \int_0^a \rho^2 \delta(\rho) \int_0^\pi \frac{1}{s^2} \cos \eta \sin \phi d\phi d\rho. \end{aligned}$$

(Since ρ and $\delta(\rho)$ do not depend on ϕ we have moved them out of the way.) The first integration gives us the contribution from the mass in a ring situated at co-latitude ϕ and distance ρ from the origin. The next integration with respect to ϕ is the



hardest part of the computation. It will give us the contribution from the mass in a spherical shell of radius ρ . It is easier if we *change the variable of integration* from ϕ to s . By the law of cosines, we have

$$s^2 = \rho^2 + R^2 - 2\rho R \cos \phi$$

whence

$$2s \, ds = -2\rho R (-\sin \phi \, d\phi)$$

or

$$\sin \phi \, d\phi = \frac{s \, ds}{\rho R}.$$

Also, at $\phi = 0$ (the north pole), we have $s = R - \rho$, and at $\phi = \pi$ (the south pole), we have $s = R + \rho$. (Look at the diagram.) Hence,

$$\begin{aligned} F_z &= -2\pi G \int_0^a \rho^2 \delta(\rho) \int_{R-\rho}^{R+\rho} \frac{1}{s^2} \cos \eta \frac{s}{\rho R} \, ds \, d\rho \\ &= -2\pi G \int_0^a \frac{\rho \delta(\rho)}{R} \int_{R-\rho}^{R+\rho} \frac{1}{s} \cos \eta \, ds \, d\rho. \end{aligned}$$

To proceed, we need to express $\cos \eta$ in terms of s . Refer to the diagram. By the law of cosines, we have

$$\rho^2 = s^2 + R^2 - 2Rs \cos \eta$$

so

$$\cos \eta = \frac{s^2 + R^2 - \rho^2}{2Rs}$$

and

$$\frac{1}{s} \cos \eta = \frac{1}{s} \frac{s^2 + R^2 - \rho^2}{2Rs} = \frac{1}{2R} \left(1 + \frac{R^2 - \rho^2}{s^2} \right).$$

Hence,

$$\begin{aligned} F_z &= -2\pi G \frac{1}{2R^2} \int_0^a \rho \delta(\rho) \int_{R-\rho}^{R+\rho} \left(1 + \frac{R^2 - \rho^2}{s^2} \right) \, ds \, d\rho \\ &= -\pi G \frac{1}{R^2} \int_0^a \rho \delta(\rho) \left(s - \frac{R^2 - \rho^2}{s} \right) \Big|_{s=R-\rho}^{s=R+\rho} d\rho \\ &= -\pi G \frac{1}{R^2} \int_0^a \rho \delta(\rho) \left(R + \rho - \frac{R^2 - \rho^2}{R + \rho} - (R - \rho) + \frac{R^2 - \rho^2}{R - \rho} \right) d\rho \\ &= -\pi G \frac{1}{R^2} \int_0^a \rho \delta(\rho) (4\rho) \, d\rho = -\frac{G}{R^2} \int_0^a 4\pi \rho^2 \delta(\rho) \, d\rho. \end{aligned}$$

The integral on the right is just the total mass M in the sphere. You can see this by setting up the integral for the mass in spherical coordinates and carrying out

the integrations with respect to θ and ϕ as above. However, since a sphere of radius ρ has surface area $4\pi\rho^2$, it is clear that the mass in a thin shell of radius ρ and thickness $d\rho$ is $4\pi\rho^2\delta(\rho) d\rho$. We get finally the desired result

$$F_z = -\frac{GM}{R^2}$$

as claimed.

The calculation of the force due to a spherical shell depends strongly on the test particle being outside the shell, i.e., $R > \rho$. The expression for $\cos\eta$ is different if the test particle is inside the shell, i.e., $R < \rho$. In that case, it turns out that the force on the test particle is zero. (See the Exercises.)

Exercises for 4.7.

1. How would the argument for Olbers' Paradox change if we assumed the density of stars was constant for $\rho < \rho_0$ and dropped off as $1/\rho$ for $\rho > \rho_0$?
2. Find the force exerted on a test particle of unit mass at the origin by a solid hemisphere of radius a centered at the origin if the density is given by $\delta = k\rho$. Express the answer in terms of the total mass. Note that if the test particle is at the origin, then $s = \rho$ in equations (56) and (57). Note also that by symmetry the x and y components of the force are zero.
3. Find the force exerted on a test particle of unit mass at the origin by a solid cone with vertex at the origin, centered on the z -axis, of height $\sqrt{3}a$ and radius a . Use the same density function $\delta = k\rho$ as in the previous problem.
4. Consider a spherical shell described by $c \leq \rho \leq a$ of constant density δ . It is clear by symmetry that the gravitational force exerted on a unit test mass at the origin is zero. Show that the force exerted on a test mass at any point *inside the shell* is zero. Hint: Change variables from ϕ to s , as was done in the text. Convince yourself that if $\rho \geq R$ (the test mass is inside the shell of radius ρ), then the calculation of $\cos\eta$ in terms of s is still correct, but the limits for s are $\rho - R \leq s \leq \rho + R$. Show that in this case the answer is zero. (Note that if $R < \rho$, it is possible for η to be obtuse, so $w = R - z$ would be negative.)
5. Let a mass M be uniformly distributed over a solid sphere of radius a . Imagine a very narrow tunnel drilled along a diameter and a test particle placed at distance R from the center. Show that the gravitational attraction is proportional to R . Ignore the effect of removing the mass in the tunnel. Hint: According to the previous problem, the mass in the shell $R \leq \rho \leq a$ will exert zero net force. What force will be exerted by the mass of the sphere $0 \leq \rho \leq R$? How does this argument change if the mass density is $\delta = k/\rho$?

4.8 Improper Integrals

One often encounters integrals involving *infinities* of one sort or another. This may occur if either the domain of integration is not bounded or if the function being integrated is not bounded on its domain. The basic method of dissecting the domain, forming a *finite* sum, and taking a limit does not work in such cases, but one can usually do something sensible. The resulting integrals are called *improper integrals*.

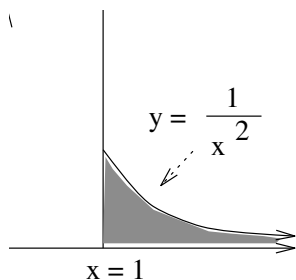
Example 76 We shall find the area bounded by the graphs of $x = 1$, $y = 0$, and $y = 1/x^2$. The region is bounded below by the x -axis and above by a graph which approaches the x -axis asymptotically. The region is cut off by the line $x = 1$ on the left, but it extends without limit to the right.

The area is calculated as follows. Consider a finite portion of the region, bounded on the right by the line $x = U$ (for ‘upper limit’). Its area is the integral

$$A(U) = \int_1^U \frac{dx}{x^2} = -\frac{1}{x} \Big|_1^U = 1 - \frac{1}{U}.$$

Now let the upper limit $U \rightarrow \infty$. The term $1/U \rightarrow 0$, so the area is

$$A = \lim_{U \rightarrow \infty} A(U) = 1.$$

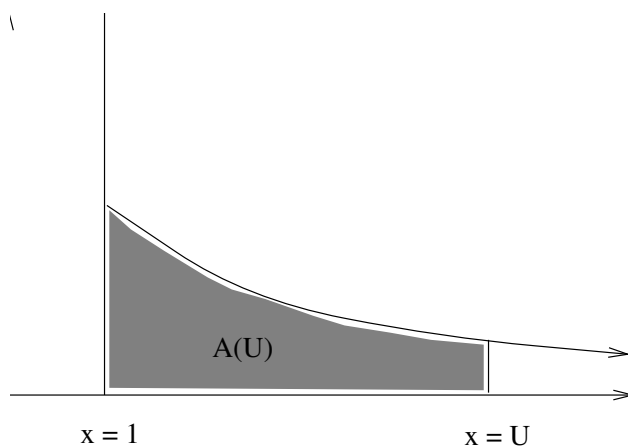


Note that the result seems a trifle paradoxical. *Although the region is unbounded, it does have a finite area* according to this plausible method for finding area.

The above example is a special case of a more general concept. Suppose $y = f(x)$ defines a function for $a \leq x < \infty$. Suppose moreover that the integral $\int_a^U f(x) dx$ exists for each $a < U$. We define the improper integral

$$\int_a^\infty f(x) dx = \lim_{U \rightarrow \infty} \int_a^U f(x) dx$$

provided this limit exists. Similar definitions can be made for $\int_{-\infty}^b f(x) dx$ or for various unbounded regions in \mathbf{R}^2 and \mathbf{R}^3 . (See below for some examples.)



The answer is not always finite.

Example 77 To determine $\int_1^\infty \frac{dx}{\sqrt{x}}$, we consider

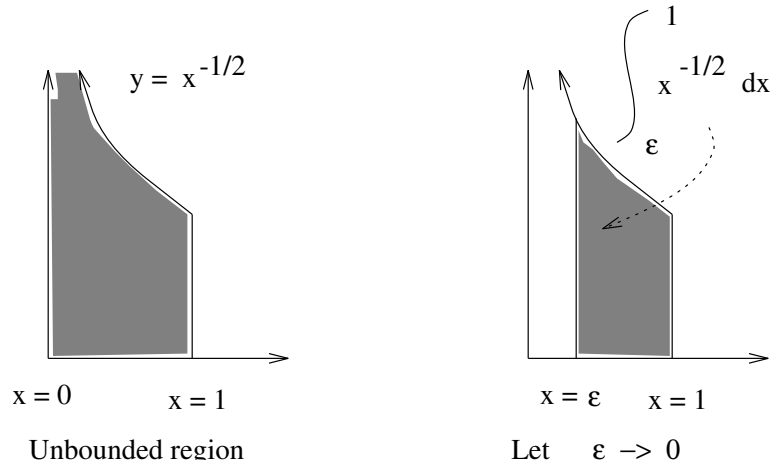
$$\int_1^U \frac{dx}{\sqrt{x}} = \int_1^U x^{-1/2} dx = \left. \frac{x^{1/2}}{1/2} \right|_1^U = 2\sqrt{U} - 2.$$

This does not have a finite limit as $U \rightarrow \infty$, so we say the improper integral *diverges* or that the answer is $+\infty$.

Example 78 We shall evaluate $\int_0^1 \frac{dx}{\sqrt{x}}$. At first glance, this looks like an ordinary integral. Indeed, we have

$$\int_0^1 \frac{dx}{\sqrt{x}} = \left. \frac{x^{1/2}}{1/2} \right|_0^1 = 2.$$

However, if you look carefully, you will notice that there is something not quite right since the integrand is not bounded near the lower limit 0. (The graph approaches the y -axis asymptotically.)



The correct way to do this problem is to treat the integral as an improper integral and to evaluate it as a limit of proper integrals. Let $0 < \epsilon < 1$. Then

$$\int_{\epsilon}^1 \frac{dx}{\sqrt{x}} = \left. \frac{x^{1/2}}{1/2} \right|_{\epsilon}^1 = 2 - 2\sqrt{\epsilon}.$$

If we now let $\epsilon \rightarrow 0$, we have

$$\int_0^1 \frac{dx}{\sqrt{x}} = \lim_{\epsilon \rightarrow 0} \int_{\epsilon}^1 \frac{dx}{\sqrt{x}} = \lim_{\epsilon \rightarrow 0} (2 - 2\sqrt{\epsilon}) = 2.$$

Evaluating the above integral as a limit is a bit silly, since the first method gives the same answer. This is a common state of affairs. What saves us from error in such cases is that the anti-derivative is continuous, so taking a limit or evaluating it yield the same answer. However, as the next example shows, it is possible to go wrong, so one should always be aware that one is really evaluating an improper integral. (Check through the previous sections and you will find several improper integrals in hiding.)

Example 79 We shall try to evaluate $\int_{-1}^1 \frac{dx}{x^2}$. The graph of the function $f(x) = 1/x^2$ is asymptotic to the positive y -axis, so it is unbounded as $x \rightarrow 0$. Suppose we ignore this and just try to do the integral by the usual method.

$$\int_{-1}^1 \frac{dx}{x^2} = \left. -\frac{1}{x} \right|_{-1}^1 = -2.$$

However, this is clearly not a correct answer since it is negative and the function is always positive. Moreover, suppose we divide the integral into two parts: one from -1 to 0 and the other from 0 to 1 . Each of these is an improper integral,

so they should be computed as limits. Looking at the second integral, we have for $0 < \epsilon < 1$,

$$\int_{\epsilon}^1 \frac{dx}{x^2} = -\frac{1}{x} \Big|_{\epsilon}^1 = \frac{2}{\epsilon} - 2,$$

and this does not approach a finite limit as $\epsilon \rightarrow 0$. By symmetry, the same argument works for the other integral, so the sum of the two is not a finite number.

In each of the above examples, the functions were always positive. In cases where we have to combine ‘positive infinities’ with ‘negative infinities’, the situation is a bit more complicated because the answer may depend on how you take limits.

Example 80 Consider $\int_{-1}^1 \frac{dx}{x}$. If we divide this into two improper integrals, we could *try*

$$\int_{-1}^1 \frac{dx}{x} = \int_{-1}^0 \frac{dx}{x} + \int_0^1 \frac{dx}{x}.$$

However

$$\begin{aligned} \int_{-1}^0 \frac{dx}{x} &= \lim_{\epsilon \rightarrow 0} \int_{-1}^{-\epsilon} \frac{dx}{x} = \lim_{\epsilon \rightarrow 0} \ln |x| \Big|_{-1}^{-\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} (\ln |-\epsilon| - \ln |-1|) = \lim_{\epsilon \rightarrow 0} \ln \epsilon = -\infty, \end{aligned}$$

and

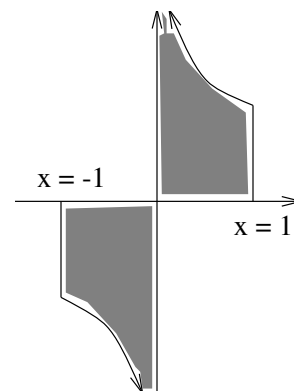
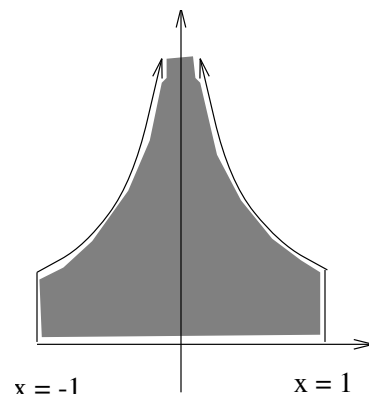
$$\begin{aligned} \int_0^1 \frac{dx}{x} &= \lim_{\eta \rightarrow 0} \int_{\eta}^1 \frac{dx}{x} = \lim_{\eta \rightarrow 0} \ln |x| \Big|_{\eta}^1 \\ &= \lim_{\eta \rightarrow 0} (-\ln \eta) = +\infty. \end{aligned}$$

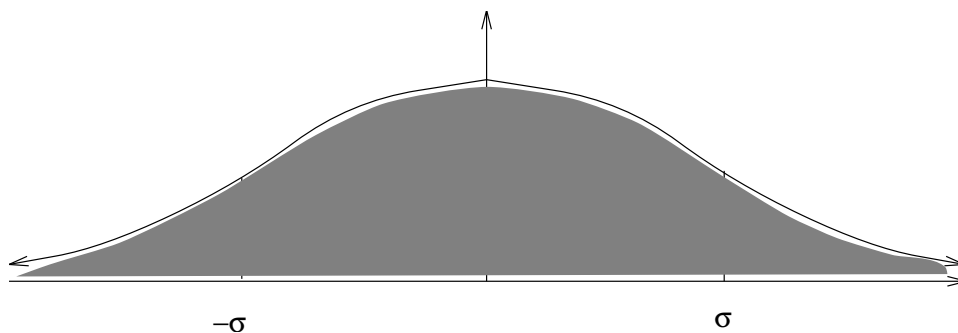
There is no sensible way to combine these infinities to get a unique value. However, we could combine the two integrals as follows

$$\begin{aligned} \int_{-1}^1 \frac{dx}{x} &= \lim_{\epsilon \rightarrow 0} \left[\int_{-1}^{-\epsilon} \frac{dx}{x} + \int_{\epsilon}^1 \frac{dx}{x} \right] \\ &= \lim_{\epsilon \rightarrow 0} [(\ln |-\epsilon| - \ln |-1|) - (\ln 1 - \ln \epsilon)] = 0. \end{aligned}$$

Here we have carefully arranged to approach zero from both directions at *exactly* the same rate, so at each stage the integrals cancel. The result 0, in this case, is called the *Cauchy principal value* of the improper integral.

The Normal Distribution In probability and statistics one encounters the so-called ‘bell shaped curve’. This is the graph of the function $f(x) = Ce^{-x^2/2\sigma^2}$ where C and σ are appropriate constants.





For any interval $[a, b]$ on the real line, the integral $\int_a^b f(x)dx$ is supposed to be the probability of a measurement of the quantity x giving a value in that interval. Here, the mean value of the measured variable is assumed to be 0, and σ , which is called the *standard deviation*, tells us how concentrated the measurements will be about that mean value. Moreover, the constant C should be chosen so that $\int_{-\infty}^{\infty} f(x) = 1$ since it is certain that a measurement will produce *some value*. Hence, C should be the reciprocal of

$$\int_{-\infty}^{\infty} e^{-x^2/2\sigma^2} dx.$$

This is of course an improper integral. The fact that both limits are infinite adds a complication, but since the function is always positive, no significant problem arises. Indeed, by symmetry, we may assume

$$\int_{-\infty}^{\infty} e^{-x^2/2\sigma^2} dx = 2 \int_0^{\infty} e^{-x^2/2\sigma^2} dx,$$

and we shall calculate the latter integral. The first step is to eliminate the parameter σ by making the substitution $u = x/\sigma, du = dx/\sigma$. This gives

$$\int_0^{\infty} e^{-x^2/2\sigma^2} dx = \sigma \int_0^{\infty} e^{-u^2/2} du.$$

The integral $I = \int_0^{\infty} e^{-u^2/2} du$ cannot be done by explicit integration, so we make use of a clever trick. Consider

$$\begin{aligned} I^2 &= \left(\int_0^{\infty} e^{-u^2/2} du \right)^2 = \left(\int_0^{\infty} e^{-u^2/2} du \right) \left(\int_0^{\infty} e^{-u^2/2} du \right) \\ &= \left(\int_0^{\infty} e^{-x^2/2} dx \right) \left(\int_0^{\infty} e^{-y^2/2} dy \right). \end{aligned}$$

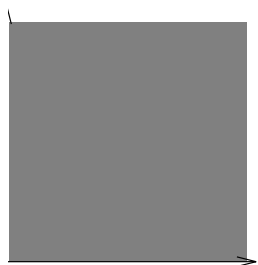
(Here we used the fact that the ‘dummy variable’ in a definite integral can be called anything at all, so we called it first ‘ x ’ and then ‘ y ’.) This product can also be written as an iterated integral

$$\int_0^{\infty} \int_0^{\infty} e^{-x^2/2} e^{-y^2/2} dy dx = \int_0^{\infty} \int_0^{\infty} e^{-(x^2+y^2)/2} dy dx.$$

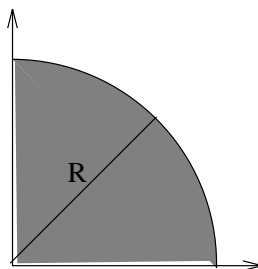
This last integral can be viewed as an *improper double integral*, i.e., as

$$\iint_D e^{-(x^2+y^2)/2} dA$$

where D is the first quadrant in the x, y -plane.



Unbounded region



Let $R \rightarrow \infty$

To calculate this improper integral, we switch to polar coordinates and treat the region D as a limit of quarter discs $D(R)$ of radius R as $R \rightarrow \infty$. Thus,

$$\begin{aligned} \iint_D e^{-(x^2+y^2)/2} dA &= \lim_{R \rightarrow \infty} \iint_{D(R)} e^{-(x^2+y^2)/2} dA \\ &= \lim_{R \rightarrow \infty} \int_0^{\pi/2} \int_0^R e^{-r^2/2} r dr d\theta \\ &= \lim_{R \rightarrow \infty} \int_0^{\pi/2} (-e^{-r^2/2}) \Big|_0^R d\theta \\ &= \frac{\pi}{2} \lim_{R \rightarrow \infty} (1 - e^{-R^2/2}) \\ &= \frac{\pi}{2}, \end{aligned}$$

since $\lim_{R \rightarrow \infty} e^{-R^2} = 0$. It follows that $I^2 = \pi/2$ whence $I = \sqrt{\pi/2}$. Hence,

$$\int_{-\infty}^{\infty} e^{-x^2/2\sigma^2} dx = 2 \int_0^{\infty} e^{-x^2/2\sigma^2} dx = 2\sigma I = 2\sqrt{\frac{\pi}{2}} \sigma = \sqrt{2\pi} \sigma.$$

Thus, we should take $C = 1/(\sqrt{2\pi} \sigma)$, so

$$\int_{-\infty}^{\infty} \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi} \sigma} dx = 1.$$

The adjusted integrand is called the *normal* or *Gaussian* density function.

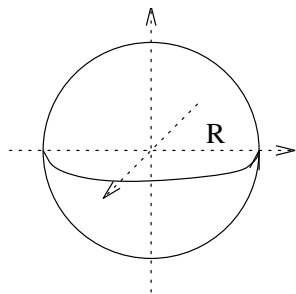
The calculation of the improper double integral involves some hidden assumptions. (See the Exercises.)

Similar calculations for unbounded regions may be done in \mathbf{R}^3

Example 81 We shall determine the improper integral

$$\iiint_{\mathbf{R}^3} e^{-(x^2+y^2+z^2)/2} dV.$$

The method is to calculate the integral for a solid sphere $E(R)$ of radius R , centered at the origin, and then let $R \rightarrow \infty$. Using spherical coordinates, we have



$$\begin{aligned} \iiint_{E(R)} e^{-(x^2+y^2+z^2)/2} dV &= \int_0^{2\pi} \int_0^\pi \int_0^R e^{-\rho^2/2} \rho^2 \sin \phi \, d\rho \, d\phi \, d\theta \\ &= \underbrace{2\pi}_{\text{from } \theta} \underbrace{\int_0^\pi \sin \phi \, d\phi}_2 \int_0^R e^{-\rho^2/2} \rho^2 \, d\rho. \end{aligned}$$

The ρ integral can be done by integrating by parts and the answer is

$$-Re^{-R^2/2} + \int_0^R e^{-\rho^2/2} d\rho.$$

Let $R \rightarrow \infty$. The first limit may be calculated by L'Hôpital's rule.

$$\lim_{R \rightarrow \infty} Re^{-R^2/2} = \lim_{R \rightarrow \infty} \frac{R}{e^{R^2/2}} = \lim_{R \rightarrow \infty} \frac{1}{Re^{R^2/2}} = 0.$$

The second term approaches $\int_0^\infty e^{-\rho^2/2} d\rho = \sqrt{\pi/2}$ by the previous calculation. Hence,

$$\iiint_{\mathbf{R}^3} e^{-(x^2+y^2+z^2)/2} dV = (2\pi)(2) \left(\sqrt{\frac{\pi}{2}} \right) = (2\pi)^{3/2}.$$

Note that the argument for Olbers Paradox in the previous section really involves an improper integral. So do many gravitational force calculations which involve integrating functions with denominators which may vanish.

Exercises for 4.8.

1. Evaluate the following improper integrals, if they converge: (a) $\int_0^1 \frac{1}{x^{5/2}} dx$.

(b) $\int_1^\infty \frac{1}{x^{5/2}} dx$. (c) $\int_{-1}^1 \frac{1}{1-x^2} dx$. (d) $\int_0^1 \frac{1}{x-1} dx$. (e) $\int_{-\infty}^0 \frac{1}{x-1} dx$. (f)

$$\int_{-\infty}^\infty \frac{x}{x+1} dx.$$

2. Consider the curve $y = f(x) = 1/x$, $x \geq 1$.
 - (a) Show that the area under the curve is infinite.
 - (b) Show that the volume formed by rotating the curve around the x -axis is finite.
3. An infinite rod of uniform density δ lies along the x -axis. Find the net gravitational force on a unit test mass located at $(0, a)$.
4. Calculate

$$\int_0^\infty \int_0^\infty \frac{dx dy}{(1 + x^2 + y^2)^2}$$

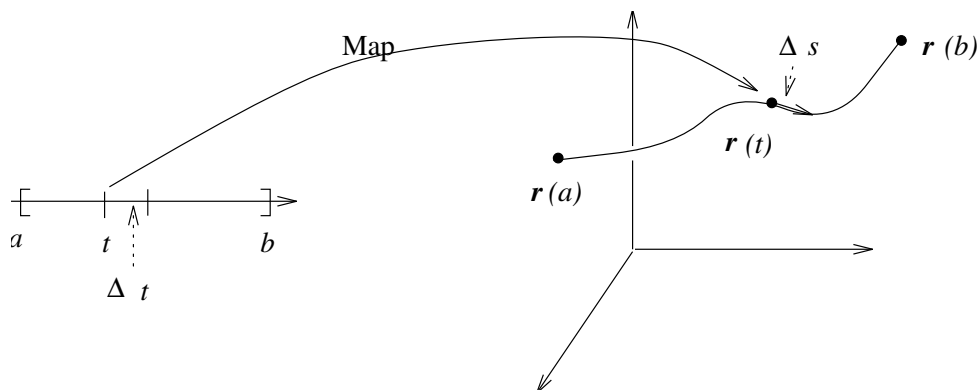
by switching to polar coordinates.

5. Find the total mass in a solid sphere of radius a supposing the mass density is given by $\delta(\rho) = k/\rho$. First find this by doing the integral in spherical coordinates. Next notice that because the integrand is unbounded as $\rho \rightarrow 0$, the integral is in fact an improper integral. With this in mind, recalculate the integral as follows. Find the mass $M(\epsilon)$ in the solid region $E(\epsilon)$ *between* an inner sphere of radius ϵ and the outer sphere of radius a . Then take the limit as $\epsilon \rightarrow 0$. You should get the same answer.
6. Look back over previous integration problems to see if you can find any which involve hidden improper integrals.

4.9 Integrals on Surfaces

The double integrals discussed so far have been for regions in \mathbf{R}^2 . We also want to be able to integrate over *surfaces* in \mathbf{R}^3 . In the former case, we can always dissect the region into true rectangles (except possibly near the boundary), but that won't generally be possible for surfaces which are usually *curved*. We encountered a similar situation in our discussion of arc length and line integrals for paths in \mathbf{R}^2 and \mathbf{R}^3 , so we shall briefly review that here.

Parametric Representations Let $\mathbf{r} = \mathbf{r}(t)$, $a \leq t \leq b$ provide a parametric representation for a path \mathcal{C} in \mathbf{R}^n ($n = 2$ or 3). It is useful to picture this by drawing a diagram which exhibits the domain $a \leq t \leq b$ on the left, \mathbf{R}^n with the image curve on the right, and a curved arrow indicating the action of *mapping* the parameter t to the point $\mathbf{r}(t)$ on the path.



To integrate on the curve, we dissect the parameter domain into small intervals Δt , and that results in a corresponding dissection of the curve into small arcs Δs where

$$\Delta s \approx |\mathbf{r}'(t)|\Delta t.$$

(The quantity on the right is just the length of a small displacement *tangent* to the curve, but it is also a good approximation to the length of the chord connecting the endpoints of the small arc.) Suppose now that $f : \mathbf{R}^n \rightarrow \mathbf{R}$ is a scalar valued function such that the image curve \mathcal{C} is contained in its domain, i.e., $f(\mathbf{r})$ is defined for \mathbf{r} on \mathcal{C} . We can form the sum

$$\sum_{\substack{t\text{-dis-} \\ \text{section}}} f(\mathbf{r})\Delta s \approx \sum_{\substack{t\text{-dis-} \\ \text{section}}} f(\mathbf{r}(t))|\mathbf{r}'(t)|\Delta t$$

which in the limit becomes

$$\int_{\mathcal{C}} f(\mathbf{r}) ds = \int_a^b f(\mathbf{r}(t))|\mathbf{r}'(t)| dt.$$

This generalizes slightly what we did before when discussing *line integrals*. In that case, we have a vector function \mathbf{F} defined on \mathcal{C} , and the scalar function f to be integrated is given by $f(\mathbf{r}) = \mathbf{F}(\mathbf{r}) \cdot \mathbf{T}$ where \mathbf{T} is the unit tangent vector at \mathbf{r} .

Example 82 Suppose a mass is distributed on a thin wire shaped in a circle of radius a in such a way that the density is proportional to the distance r to a fixed point O on the circle. We shall find the total mass. To this end, introduce a coordinate system with the origin at O and the x -axis pointing along the diameter through O . (See the diagram.) Then the mass density will have the form $\delta(\mathbf{r}) = kr = k\sqrt{x^2 + y^2}$, and we want to find $\int_{\mathcal{C}} \delta(\mathbf{r}) ds$. We know that the circle may be described in polar coordinates by $r = 2a \cos \theta$, so using $x = r \cos \theta$, $y = r \sin \theta$, we obtain a parametric representation in terms of θ

$$\mathbf{r} = \langle 2a \cos^2 \theta, 2a \cos \theta \sin \theta \rangle, \quad -\pi/2 \leq \theta \leq \pi/2.$$

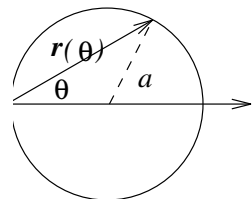
Hence,

$$\begin{aligned}\mathbf{r}'(\theta) &= \langle -4a \cos \theta \sin \theta, -2a \sin \theta \sin \theta + 2a \cos \theta \cos \theta \rangle \\ &= \langle -2a \sin 2\theta, 2a \cos 2\theta \rangle,\end{aligned}$$

so $|\mathbf{r}'(\theta)| = 2a$. In addition, on the curve, $\delta(\mathbf{r}) = kr = k(2a \cos \theta)$, so

$$\int_C \delta(\mathbf{r}) ds = \int_{-\pi/2}^{\pi/2} (2ak \cos \theta)(2a) d\theta = 4a^2k \sin \theta \Big|_{-\pi/2}^{\pi/2} = 8a^2k.$$

(You might also try to do the problem by choosing a coordinate system with origin at the center of the circle. Then the expression for $\delta(\mathbf{r})$ would be a bit more complicated.)

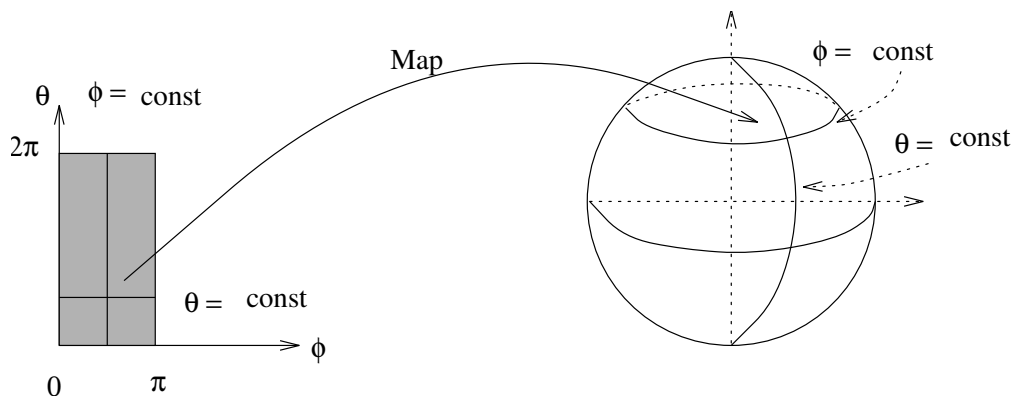


We want to do something similar for surfaces in space. So far, we have met surfaces as graphs of functions $f : \mathbf{R}^2 \rightarrow \mathbf{R}$ or as level sets of functions $g : \mathbf{R}^3 \rightarrow \mathbf{R}$. They may also be represented *parametrically* by vector valued functions $\mathbf{R}^2 \rightarrow \mathbf{R}^3$. Before, discussing the general case, we recall our discussion of the surface of a sphere which is one of the most important applications.

Example 83 In our discussion of ‘geography’, we noted that on the surface of the sphere $\rho = a$, the spherical coordinates ϕ, θ are *intrinsic coordinates* specifying the position of points on that sphere. Moreover, using $x = \rho \sin \phi \cos \theta, y = \rho \sin \phi \sin \theta, z = \rho \cos \phi$, we may specify the relation between (ϕ, θ) and the position vector of the point on the sphere by

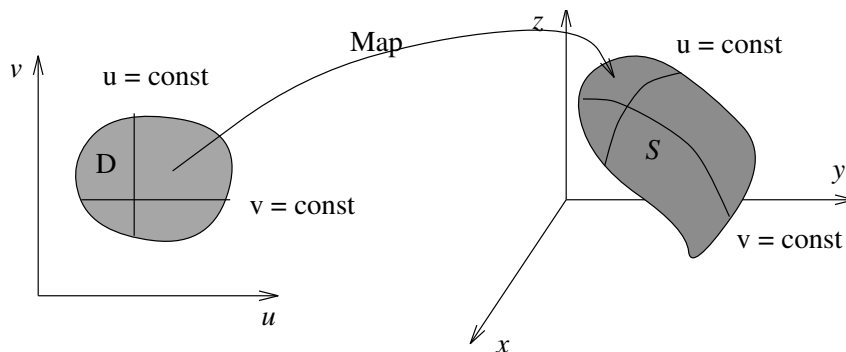
$$\begin{aligned}\mathbf{r} &= \langle a \sin \phi \cos \theta, a \sin \phi \sin \theta, a \cos \phi \rangle, \\ 0 \leq \phi \leq \pi, 0 \leq \theta < 2\pi.\end{aligned}$$

As above, consider a diagram with a ϕ, θ -plane on the left, the sphere imbedded in \mathbf{R}^3 on the right, and a curved arrow indicating the action of *mapping* (ϕ, θ) to the image point on the sphere.



ϕ, θ on the left should be thought of as *rectangular* coordinates in a *map* of the sphere, while the picture on the right represents ‘reality’. In the map, circles of latitude are represented by vertical lines and meridians of longitude by horizontal lines. The entire sphere is covered by mapping the rectangle $0 \leq \phi \leq \pi, 0 \leq \theta \leq 2\pi$. The bottom edge of this rectangle ($\phi = 0$) is mapped to the North Pole on the sphere, and similarly, the upper edge ($\phi = \pi$) is mapped to the South Pole. For points interior to the rectangle, there is a one-to-one correspondence between parameter points (ϕ, θ) and points on the sphere.

The general situation is quite similar. We suppose we are given a smooth vector valued function $\mathbf{r} = \mathbf{r}(u, v)$ defined on some domain D in the u, v -plane and taking values in \mathbf{R}^3 . The subset of \mathbf{R}^3 consisting of image points $\mathbf{r}(u, v)$ for (u, v) in D will generally be a surface, and we say the function $\mathbf{r} = \mathbf{r}(u, v)$ is a parametric representation of this surface. As above, we picture this by a diagram with the u, v -parameter plane on the left, \mathbf{R}^3 with the image surface imbedded on the right, and a curved arrow indicating the action of mapping (u, v) to $\mathbf{r}(u, v)$.



We assume that at least for the interior of the domain, the function is one-to-one, i.e., distinct points in the parameter plane map to distinct points on the surface. However, for the boundary of the domain, the one-to-one condition may fail.

Horizontal lines ($v = \text{constant}$) in the parameter domain, and vertical lines ($u = \text{constant}$) map to curves on the surface, and it is usually worthwhile seeing what those lines are.

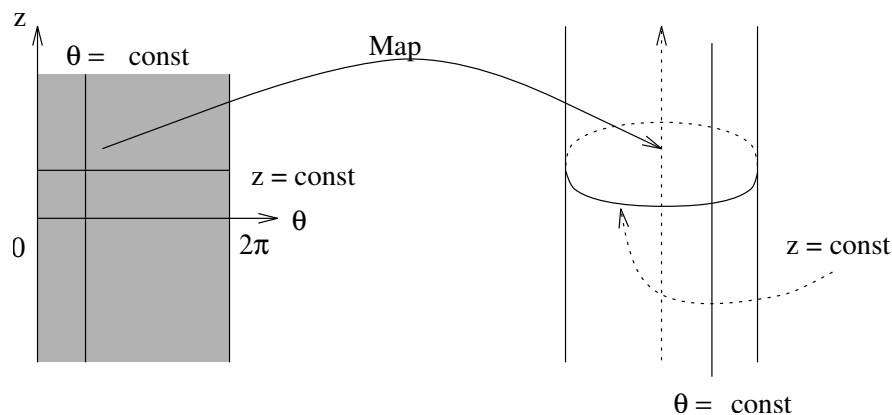
There are of course as many surfaces which can be defined this way as there are functions $\mathbf{R}^2 \rightarrow \mathbf{R}^3$. However, it is not necessary at this point to be familiar with all of them; knowing how to represent certain simple surfaces parametrically will suffice. We started with the surface of a sphere. The case of a cylinder is even easier.

Example 84 Consider a cylinder of radius a centered on the z -axis. In cylindrical coordinates, this is described simply by $r = a$. Putting this in $x = r \cos \theta, y = r \sin \theta$,

we obtain the following parametric representation of the cylinder

$$\mathbf{r} = \mathbf{r}(\theta, z) = \langle a \cos \theta, a \sin \theta, z \rangle$$

$$0 \leq \theta < 2\pi, -\infty < z < \infty.$$



The parameter domain in this case is the infinite strip between the lines $\theta = 0$ and $\theta = 2\pi$ in the θ, z -plane. If we wanted only a finite portion of the cylinder, say between $z = 0$ and $z = h$, we would appropriately limit the domain.

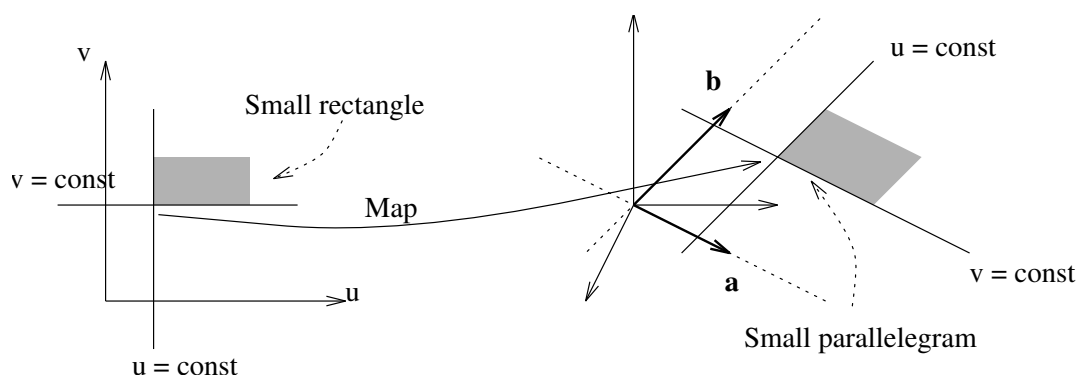
The lines $\theta = \text{constant}$ in the θ, z -plane correspond to vertical lines on the cylinder. The lines $z = \text{constant}$ in the parameter plane correspond to circles on the cylinder, parallel to the x, y -plane.

Example 85 Let $\mathbf{a} = \langle a_1, a_2, a_3 \rangle$ and $\mathbf{b} = \langle b_1, b_2, b_3 \rangle$ be fixed vectors in \mathbf{R}^3 , and consider the function defined by

$$\mathbf{r} = \mathbf{r}(u, v) = u\mathbf{a} + v\mathbf{b} = \langle a_1u + b_1v, a_2u + b_2v, a_3u + b_3v \rangle,$$

$$-\infty < u < \infty, -\infty < v < \infty.$$

Here the domain is the entire u, v -plane and the image surface is just the plane through the origin containing the vectors \mathbf{a} and \mathbf{b} .



The lines $u = \text{constant}$ and $v = \text{constant}$ in the parameter domain correspond to lines in the image plane. Note that these lines won't generally be perpendicular to one another. In fact the line $u = \text{constant}$ (v varying) will meet the line with $v = \text{constant}$ (u varying) in the same angle as that between \mathbf{a} and \mathbf{b} .

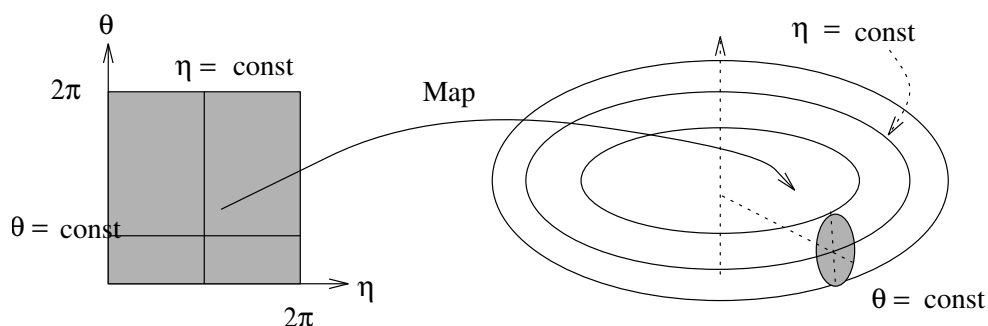
How would you modify this function to represent the plane which passes through the endpoint of the position vector \mathbf{r}_0 and which is parallel to the above plane?

Example 86 Consider the circle of radius a in the x, z -plane centered at $(b, 0, 0)$. The surface obtained by rotating this circle about the z -axis is called a *torus*. Using cylindrical coordinates, you see from the diagram that

$$\begin{aligned} r &= b + a \cos \eta \\ z &= a \sin \eta \end{aligned}$$

where η is the indicated angle. Hence, we obtain the parametric representation

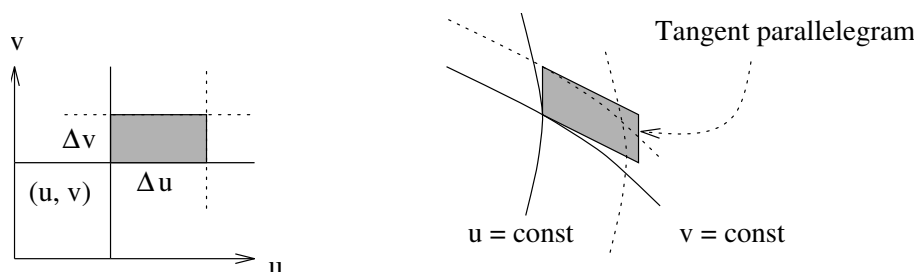
$$\begin{aligned} \mathbf{r} &= \mathbf{r}(\eta, \theta) = \langle (b + a \cos \eta) \cos \theta, (b + a \cos \eta) \sin \theta, a \sin \eta \rangle, \\ 0 &\leq \eta < 2\pi, 0 \leq \theta < 2\pi. \end{aligned}$$



The line $\eta = \text{constant}$ in the η, θ -domain corresponds to a circle on the torus centered on the z -axis and parallel to the x, y -plane. The line $\theta = \text{constant}$ in the η, θ -domain corresponds to a circle on the torus obtained by cutting it crosswise with the half-plane from the z -axis determined by that value of θ .

Integrating on Parametrically Defined Surfaces Let \mathcal{S} denote a surface in \mathbf{R}^3 represented parametrically by $\mathbf{r} = \mathbf{r}(u, v)$. In what follows, we shall make use of the curves on the surface obtained by keeping one of the parameters constant and letting the other vary. For example, if v is constant, $\mathbf{r}(u, v)$ provides a parametric representation of a curve with parameter u . As usual, you can find a tangent vector to the curve by taking the derivative, but since v is constant, it is the partial derivative, $\partial \mathbf{r} / \partial u$. Similarly, $\partial \mathbf{r} / \partial v$ is a vector tangent to the curve obtained by keeping u constant and letting v vary.

Let $f(x, y, z) = f(\mathbf{r})$ be a scalar valued function with domain containing the surface \mathcal{S} . We want to define what we mean by the integral of the function on the surface. Our method parallels what we did in the case of curves. Imagine that the domain D of the parameterizing function is dissected into small rectangles with the area of a typical rectangle being $\Delta A = \Delta u \Delta v$. Corresponding to this dissection is a dissection of the surface into subsets we shall call *curvilinear rectangles*.

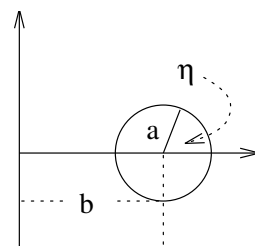


Suppose for the moment that we know how to find the *surface area* ΔS of each such curvilinear rectangle. Also, for each such curvilinear rectangle, choose a point $\mathbf{r} = \mathbf{r}(u, v)$ inside it at which to evaluate $f(\mathbf{r})$. (For reasonable functions, it won't matter much where it is chosen.) Form the sum

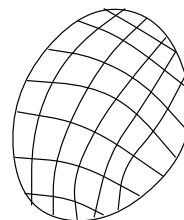
$$\sum_{\text{dissection of } \mathcal{S}} f(\mathbf{r}) \Delta S \quad (58)$$

and consider what happens to such sums as the dissection gets finer and finer and the number of curvilinear rectangles goes to ∞ . If the sums approach a limit, we denote it by

$$\iint_{\mathcal{S}} f(x, y, z) dS \quad \text{or sometimes} \quad \int_{\mathcal{S}} f(x, y, z) dS.$$

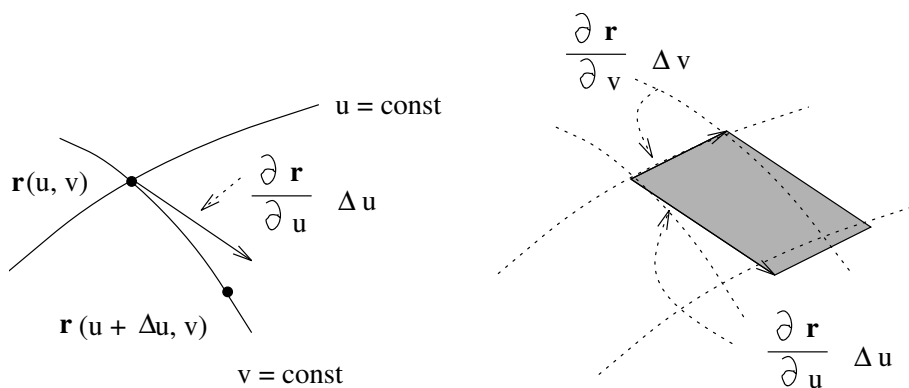


Cross section with constant θ



Dissection of surface \mathcal{S}

The crucial part of this analysis is determining how to identify the element of surface area ΔS for a typical curvilinear rectangle and seeing how that is related to the area $\Delta A = \Delta u \Delta v$ of the corresponding rectangle in the parameter domain. There are several ways to do this, and seeing how they are related is a bit involved. The method we shall use is based on the idea that the *tangent plane* to the surface is a good approximation to the surface. Let (u, v) be the lower left corner of a small rectangle in the parameter domain. Consider the two sides meeting there and their images which form the corresponding sides of the curvilinear rectangle. The image of the side from (u, v) to $(u + \Delta u, v)$ is mapped into the arc from $\mathbf{r}(u, v)$ to $\mathbf{r}(u + \Delta u, v)$. This arc is approximated pretty closely by the tangent vector $(\partial \mathbf{r} / \partial u) \Delta u$. Similarly, the arc from $\mathbf{r}(u, v)$ to $\mathbf{r}(u, v + \Delta v)$ is approximated pretty closely by the tangent vector $(\partial \mathbf{r} / \partial v) \Delta v$.



Thus, the *parallelogram* spanned by these tangent vectors is a good approximation to the curvilinear rectangle, and we may take ΔS to be the area of that parallelogram.

$$\Delta S = \left| \frac{\partial \mathbf{r}}{\partial u} \Delta u \times \frac{\partial \mathbf{r}}{\partial v} \Delta v \right| = \left| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right| \Delta u \Delta v.$$

We can also write this equation as

$$\Delta S = \left| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right| \Delta A,$$

so that $\left| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right|$ is the *correction factor* needed to convert area $\Delta A = \Delta u \Delta v$ in the parameter plane into area ΔS on the surface.

If we put this value for ΔS in formula (58), and take the limit, we obtain

$$\iint_S f(\mathbf{r}) dS = \iint_D f(\mathbf{r}(u, v)) \left| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right| dA$$

where the integral on the right is a normal double integral in the u, v -parameter plane. In particular, if we take $f(\mathbf{r}) = 1$, we obtain a formula for the surface area of \mathcal{S}

$$S = \iint_{\mathcal{S}} 1 dS = \iint_D \left| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right| dA.$$

Example 83, (revisited) We shall find the element of surface area on a sphere. We have

$$\mathbf{r} = \langle a \sin \phi \cos \theta, a \sin \phi \sin \theta, a \cos \phi \rangle,$$

so

$$\begin{aligned} \frac{\partial \mathbf{r}}{\partial \phi} &= \langle a \cos \phi \cos \theta, a \cos \phi \sin \theta, -a \sin \phi \rangle, \\ \frac{\partial \mathbf{r}}{\partial \theta} &= \langle -a \sin \phi \sin \theta, a \sin \phi \cos \theta, 0 \rangle. \end{aligned}$$

These vectors are perpendicular to each other. You can see that directly from the formulas or you can remember that they are tangent respectively to a meridian of longitude and a circle of latitude. *The magnitude of the cross product of two perpendicular vectors is just the product of their magnitudes.* We have

$$\begin{aligned} \left| \frac{\partial \mathbf{r}}{\partial \phi} \right| &= \sqrt{a^2 \sin^2 \phi (\cos^2 \theta + \sin^2 \theta) + a^2 \cos^2 \phi} = a, \\ \left| \frac{\partial \mathbf{r}}{\partial \theta} \right| &= \sqrt{a^2 \sin^2 \phi (\sin^2 \theta + \cos^2 \theta)} = a \sin \phi, \end{aligned}$$

so

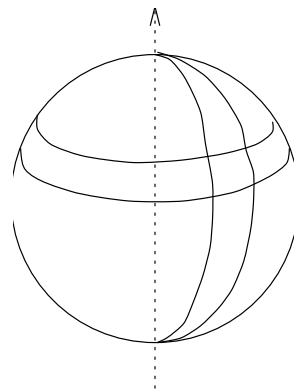
$$\begin{aligned} dS &= \left| \frac{\partial \mathbf{r}}{\partial \phi} \times \frac{\partial \mathbf{r}}{\partial \theta} \right| dA \\ &= a^2 \sin \phi d\phi d\theta. \end{aligned}$$

That of course is the same value that was obtained earlier when we viewed this same curvilinear rectangle on the sphere as the base of a spherical cell in computing the element of volume in spherical coordinates.

Let's use this to calculate the surface area of a sphere of radius a .

$$\begin{aligned} S &= \iint_{\mathcal{S}} 1 dS = \int_0^{2\pi} \int_0^\pi a^2 \sin \phi d\phi d\theta \\ &= a^2 \underbrace{\int_0^{2\pi} d\theta}_{2\pi} \underbrace{\int_0^\pi \sin \phi d\phi}_2 \\ &= 4\pi a^2, \end{aligned}$$

and that is the answer you should be familiar with from high school.



Example 84, (revisited) Assume a mass is distributed over the surface of a right circular cylinder of height h and radius a so that the density δ is proportional to the distance to the base. We shall find the total mass. We represent the cylinder parametrically by

$$\mathbf{r} = \mathbf{r}(\theta, z) = \langle a \cos \theta, a \sin \theta, z \rangle$$

$$0 \leq \theta < 2\pi, 0 \leq z \leq h.$$

Then

$$\frac{\partial \mathbf{r}}{\partial \theta} = \langle -a \sin \theta, a \cos \theta, 0 \rangle$$

$$\frac{\partial \mathbf{r}}{\partial z} = \langle 0, 0, 1 \rangle,$$

and again these are perpendicular. The product of their magnitudes is just a , so the element of area is

$$dS = a \, d\theta \, dz.$$

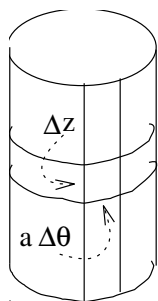
Note that it is quite easy to see why this formula should be correct by looking directly at the curvilinear rectangle on the surface of the cylinder. Its dimensions are $a \, d\theta$ by dz .

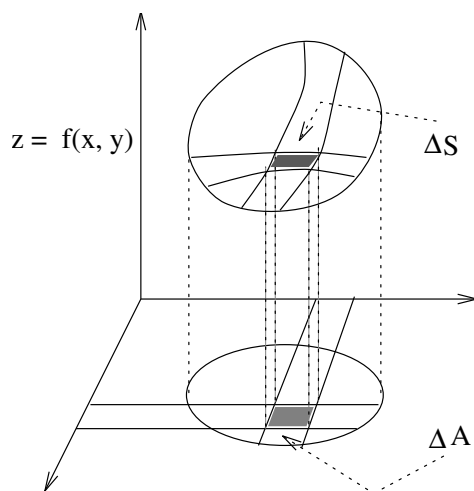
The mass density has been assumed to have the form $\delta(\mathbf{r}) = kz$ for some constant of proportionality k . Hence, the mass is given by

$$\begin{aligned} \iint_S kz \, dS &= \int_0^{2\pi} \int_0^h kz \, a \, d\theta \, dz \\ &= a k \underbrace{\int_0^{2\pi} d\theta}_{2\pi} \int_0^h z \, dz = 2\pi a k \left. \frac{z^2}{2} \right|_0^h \\ &= \pi a h^2 k. \end{aligned}$$

The Graph of a Function One very important case of a surface in \mathbf{R}^3 is that of a graph of a function. This may be treated as a special case of a parametrically defined surface as follows. Suppose $z = f(x, y)$ denotes a function with domain D in \mathbf{R}^2 . Let x, y be the parameters and set

$$\mathbf{r} = \mathbf{r}(x, y) = \langle x, y, f(x, y) \rangle, \quad \text{for } (x, y) \text{ in } D.$$





In this case the element of surface area is calculated as follows.

$$\begin{aligned}\frac{\partial \mathbf{r}}{\partial x} &= \langle 1, 0, f_x \rangle, \\ \frac{\partial \mathbf{r}}{\partial y} &= \langle 0, 1, f_y \rangle.\end{aligned}$$

These vectors are not generally perpendicular, so we need to calculate

$$\frac{\partial \mathbf{r}}{\partial x} \times \frac{\partial \mathbf{r}}{\partial y} = \langle 1, 0, f_x \rangle \times \langle 0, 1, f_y \rangle = \langle -f_x, -f_y, 1 \rangle.$$

(You may remember that we encountered the same vector earlier as a normal vector to the graph.) Hence, the correction factor is $|\partial \mathbf{r} / \partial x \times \partial \mathbf{r} / \partial y| = \sqrt{f_x^2 + f_y^2 + 1}$, and

$$dS = \sqrt{f_x^2 + f_y^2 + 1} dA,$$

where $dA = dy dx$ is the element of area in the domain D of the function f .

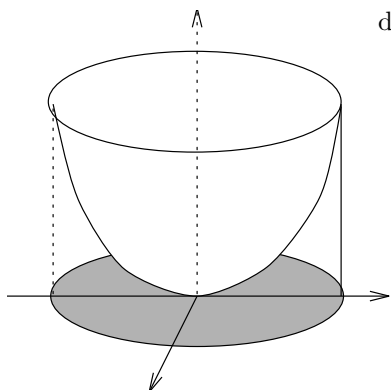
Example 87 We shall find the surface area of that portion of the paraboloid $z = x^2 + y^2$ below the plane $z = 4$. Here $f(x, y) = x^2 + y^2$ and D is the disc in the x, y -plane determined by $x^2 + y^2 \leq 4$. We have $f_x = 2x, f_y = 2y$, so

$$\sqrt{f_x^2 + f_y^2 + 1} = \sqrt{4x^2 + 4y^2 + 1}$$

and

$$S = \iint_S 1 dS = \iint_D \sqrt{4(x^2 + y^2) + 1} dA.$$

The problem has now been reduced to a double integral in the x, y -plane. This can be done by any method you find convenient. For example, since the region D is a



disk, it would seem reasonable to use polar coordinates.

$$\begin{aligned}
 \iint_D \sqrt{4(x^2 + y^2) + 1} \, dA &= \int_0^{2\pi} \int_0^2 \sqrt{4r^2 + 1} \, r \, dr \, d\theta \\
 &= \underbrace{\int_0^{2\pi} d\theta}_{2\pi} \frac{1}{8} \int_0^2 \sqrt{4r^2 + 1} \, 8r \, dr \\
 &= \frac{\pi}{4} \left. \frac{(4r^2 + 1)^{3/2}}{3/2} \right|_0^2 \\
 &= \frac{\pi}{6} (\sqrt{4913} - 1).
 \end{aligned}$$

Note that in effect, we have introduced two correction factors here. The first $\sqrt{4x^2 + 4y^2 + 1}$ converted area in the x, y -plane to surface area on the graph. The second r was needed because we chose to use polar coordinates in the x, y -plane. There is an entirely different approach which introduces only one correction factor. Namely, use the parametric representation

$$\begin{aligned}
 \mathbf{r} &= \langle r \cos \theta, r \sin \theta, r^2 \rangle, \\
 0 &\leq r \leq 2, 0 \leq \theta < 2\pi.
 \end{aligned}$$

In this case, the domain of integration is a *rectangle* in the r, θ -plane, and the correction factor ends up being

$$\left| \frac{\partial \mathbf{r}}{\partial r} \times \frac{\partial \mathbf{r}}{\partial \theta} \right| = \dots = \sqrt{4r^4 + r^2} = \sqrt{4r^2 + 1} \, r.$$

(You should check the ... in the above calculation.)

Example 88 We shall find the area of the circle in the plane $x = 1$ of radius 2 centered on the point $(1, 0, 0)$. This is a bit silly since the answer is clearly $\pi 2^2 = 4\pi$, but let's see how the method gives that answer. The surface in this case may be viewed as the graph of a function $x = g(y, z) = 1$ with domain D a disc of radius 2 centered at the origin in the y, z -plane. In this case, the appropriate element of area would be

$$dS = \sqrt{g_y^2 + g_z^2 + 1} \, dy \, dz$$

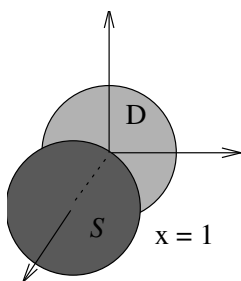
but since $g_y = g_z = 0$, the correction factor is just 1. Hence,

$$S = \iint_D 1 \, dy \, dz.$$

We need not actually do the integral, since we know that the answer will just be the area of the domain D , which is $\pi 2^2 = 4\pi$.

Note that you could also represent the surface parametrically by

$$\begin{aligned}
 \mathbf{r} &= \langle 1, r \cos \theta, r \sin \theta \rangle, \\
 0 &\leq r \leq 2, 0 \leq \theta < 2\pi.
 \end{aligned}$$



Exercises for 4.9.

- Calculate $\int_{\mathcal{C}} f(\mathbf{r}) ds$ for each indicated function over the curve given by the indicated parametrization.
 - $f(x, y) = x - y$, $\mathbf{r} = \langle t^2, t^3 \rangle$, $-1 \leq t \leq 1$.
 - $f(x, y, z) = x + 2y + z$, $\mathbf{r} = \langle \cos \theta, \sin \theta, 1 \rangle$, $0 \leq \theta \leq 2\pi$.
 - $f(x, y) = x$, $\mathbf{r} = \langle x, x^2 \rangle$, $0 \leq x \leq 1$. (\mathcal{C} is the graph of $y = x^2$.)
- A mass is distributed over a thin wire in the shape of the semi-circle $x^2 + y^2 = a^2$, $y \geq 0$. If the density is given by $\delta = y$, find the center of mass.
- Show that the surface area of a right circular cylinder of radius a and height h is $2\pi ah$.
- Find the surface area of the indicated surface.
 - The part of the plane $2x + 3y + z = 2$ contained inside the cylinder $x^2 + y^2 = 4$.
 - The part of the paraboloid $z = 4 - x^2 - y^2$ above the x, y -plane.
 - The part of the sphere $x^2 + y^2 + z^2 = a^2$ above the plane $z = h$ where $0 \leq h \leq a$.
 - The surface parametrized by $\mathbf{r} = \langle u^2, v^2, uv \rangle$, $0 \leq u \leq 1, 0 \leq v \leq 1$.
- Let a sphere of radius a be intersected by two parallel planes. Show that the area of the portion of the sphere between the planes depends only on the distance between the planes. Hint: Use part (c) above.
- Evaluate $\iint_{\mathcal{S}} z dS$ for each of the surfaces in Problem (??).
- Find the centroid of the hemisphere $x^2 + y^2 + z^2 = a^2$, $0 \leq z$.
- Consider the cone described parametrically by $\mathbf{r} = \langle r \cos \theta, r \sin \theta, mr \rangle$, $0 \leq r \leq a, 0 \leq \theta \leq 2\pi$. Show that its surface area is πaL where $L = a\sqrt{1+m^2}$ is its 'slant height'. What is the geometric significance of L ?
- Find the surface area of that part of the cylinder $x^2 + y^2 = 1$
 - under the plane $z = y$ and above the x, y -plane,
 - cut off by the cylinder $x^2 + z^2 = 1$.
- Find the surface area of that part of the sphere $x^2 + y^2 + z^2 = 4a^2$ inside the cylinder $(x - a)^2 + y^2 = a^2$. (Break the surface into two symmetrical parts and double the answer. This is a sphere, but you can also treat each part as the graph of a function.)
- Find the surface area of the torus obtained by rotating the circle $(x - b)^2 + z^2 = a^2$ about the z -axis. Hint: Use the parametric representation in Example 86.

12. A mass of density δ is uniformly distributed over the surface of the cylinder $x^2 + y^2 = a^2$, $0 \leq z \leq h$. Find the gravitation force on a test mass at the origin.

4.10 The Change of Variables Formula

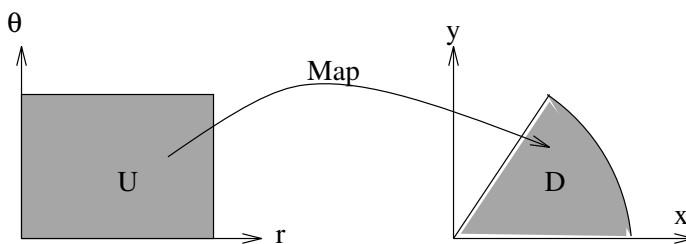
Recall how we use substitution to evaluate an ordinary integral $\int_a^b f(x) dx$. We try to express $x = x(u)$ in terms of some other variable, and then

$$\int_a^b f(x) dx = \int_c^d f(x(u)) \frac{dx}{du} du$$

where the u -limits are chosen so that $a = x(c)$ and $b = x(d)$. A similar method works to evaluate a double integral $\iint_D f(x, y) dA$, but of course it is more complicated. To proceed by analogy, assume we have $x = x(u, v)$ and $y = y(u, v)$. These two functions may be used to define a vector function $\mathbf{R}^2 \rightarrow \mathbf{R}^2$ given by

$$\mathbf{r} = \mathbf{r}(u, v) = \langle x(u, v), y(u, v) \rangle$$

which *transforms* some domain U in the u, v -plane into the domain D in the x, y -plane.



Assume that this function provides a one-to-one correspondence between the interior of U and the interior of D . That will insure that some parts of D are not covered more than once, so we won't have to worry about a part of the domain contributing more than once to the integral. (The restriction can be weakened for the boundaries without creating problems.) We want a formula with relates $\iint_D f(x, y) dy dx$ to an integral of the form $\iint_U f(x(u, v), y(u, v)) C(u, v) du dv$ where $C(u, v)$ is an appropriate 'correction factor'. One way to determine the correction factor is as follows. Think of the function as a mapping into \mathbf{R}^3 with the third coordinate zero

$$\mathbf{r} = \mathbf{r}(u, v) = \langle x(u, v), y(u, v), 0 \rangle.$$

From this point of view, the region D is a surface in \mathbf{R}^3 (represented parametrically) which happens to be contained in the x, y -plane. Then the correction factor for area on this ‘surface’ is $C(u, v) = |(\partial \mathbf{r} / \partial u) \times (\partial \mathbf{r} / \partial v)|$. However,

$$\begin{aligned}\frac{\partial \mathbf{r}}{\partial u} &= \left\langle \frac{\partial x}{\partial u}, \frac{\partial y}{\partial u}, 0 \right\rangle, \\ \frac{\partial \mathbf{r}}{\partial v} &= \left\langle \frac{\partial x}{\partial v}, \frac{\partial y}{\partial v}, 0 \right\rangle.\end{aligned}$$

so

$$\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} = \left\langle 0, 0, \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial y}{\partial u} \frac{\partial x}{\partial v} \right\rangle.$$

Thus,

$$dy \, dx = \left| \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial y}{\partial u} \frac{\partial x}{\partial v} \right| du \, dv.$$

The quantity in absolute values is called the *Jacobian* of the transformation relating x, y to u, v . It is often denoted

$$\frac{\partial(x, y)}{\partial(u, v)}.$$

It may also be characterized as the 2×2 determinant

$$\det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{bmatrix}.$$

Example 89, (Polar Coordinates). Consider the transformation $\mathbf{R}^2 \rightarrow \mathbf{R}^2$ defined by

$$\mathbf{r} = \mathbf{r}(r, \theta) = \langle r \cos \theta, r \sin \theta \rangle.$$

This is the transformation used when expressing points in the x, y -plane in terms of polar coordinates (r, θ) . From this point of view, r and θ are *rectangular* coordinates in a ‘fictitious’ r, θ -plane which through the transformation maps points in the ‘real’ x, y -plane. We have

$$\begin{aligned}\frac{\partial x}{\partial r} &= \cos \theta & \frac{\partial y}{\partial r} &= \sin \theta \\ \frac{\partial x}{\partial \theta} &= -r \sin \theta & \frac{\partial y}{\partial \theta} &= r \cos \theta\end{aligned}$$

so the Jacobian is

$$\frac{\partial(x, y)}{\partial(r, \theta)} = r \cos^2 \theta - (-r \sin^2 \theta) = r.$$

Since $r \geq 0$, we have the change of variables formula

$$\iint_D f(x, y) \, dA = \iint_U f(r \cos \theta, r \sin \theta) r \, dr \, d\theta$$

where U is the domain in the r, θ -plane which describes the region D in polar coordinates. Notice that this is essentially the same as what we derived earlier.

Example 90 We shall find the area enclosed within the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$. (Assume $a, b > 0$.) We make the change of variables $x = au, y = bv$, i.e.,

$$\mathbf{r} = \mathbf{r}(u, v) = \langle au, bv \rangle.$$

Then

$$\begin{aligned} \frac{\partial x}{\partial u} &= a & \frac{\partial y}{\partial u} &= 0 \\ \frac{\partial x}{\partial v} &= 0 & \frac{\partial y}{\partial v} &= b. \end{aligned}$$

It follows that

$$\frac{\partial(x, y)}{\partial(r, \theta)} = ab,$$

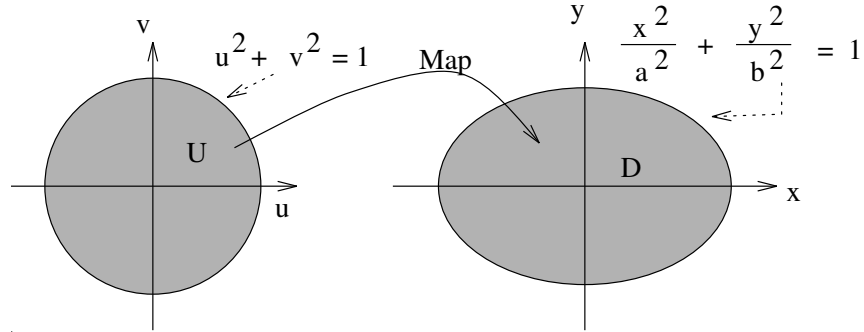
so

$$\iint_D 1 dA = \iint_U 1(ab) du dv,$$

where U in the u, v -plane corresponds to D . However, substituting for x and y in terms of u and v yields

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{a^2 u^2}{a^2} + \frac{b^2 v^2}{b^2} = u^2 + v^2,$$

so the *circle* $u^2 + v^2 = 1$ in the u, v -plane corresponds to the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ in the x, y -plane. It is not hard to see that U is the interior of a circle of radius 1 in the u, v -plane, so its area is easy to find.



Thus,

$$\iint_U ab du dv = (ab) \iint_U du dv = ab \pi 1^2 = \pi ab.$$

Note that in this example, it wasn't actually necessary to work out the u, v integral. This is in keeping with the point. One chooses to use the transformation formula for the multiple integral in the hope that the calculation in the new coordinates (u, v) will be easier than the calculation in the original coordinates (x, y) .

Some Subtle Points in the Theory Our treatment of the change of variables formula can involve elements of circular reasoning if it is not worked out carefully. Our basic argument is that the change of variables formula can be derived from the formula for the integral over a surface. There is an implicit assumption here, namely, that the concept of *area* ($'dA'$) in the domain D , when it is viewed as a subset of \mathbf{R}^2 , is the same as the concept of *surface area* ($'dS'$) when D is viewed as a parametrically defined (albeit flat) surface. Of course, that is a true fact, but one must prove it, and the proof will depend on the precise definitions of the two concepts.

In our previous discussions we were a trifle vague about how these concepts are defined. Let's look a little closer. First, let's consider the definition we introduced for the surface integral

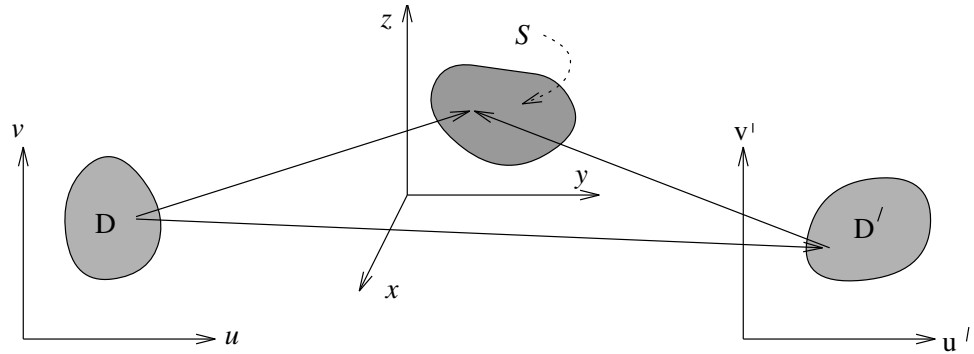
$$\iint_S f(\mathbf{r}) dS = \iint_D f(\mathbf{r}(u, v)) \left| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right| du dv.$$

Suppose the same surface has another parametric representation $\mathbf{r} = \mathbf{r}'(u', v')$ with domain D' . If we compute

$$\iint_{D'} f(\mathbf{r}(u', v')) \left| \frac{\partial \mathbf{r}}{\partial u'} \times \frac{\partial \mathbf{r}}{\partial v'} \right| du' dv',$$

how do we know that we will get the *same answer*? This question merits some thought. For example, note that we won't necessarily get the same answer in general, since if the parameterizing functions are not both one-to-one, parts of the surface may be covered more than once by one or the other of the two functions. Suppose then that both parameterizing functions are one-to-one (except possibly on their boundaries). Consider the relation between u, v -coordinates and u', v' -coordinates of the *same point* on the surface:

$$(u, v) \longrightarrow \mathbf{r}(u, v) = \mathbf{r}'(u', v') \longleftarrow (u', v').$$



Use this to define a transformation $(u', v') = \mathbf{T}(u, v)$. Then some rather involved calculation using the chain rule gives the formula

$$\left| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right| = \left| \frac{\partial \mathbf{r}}{\partial u'} \times \frac{\partial \mathbf{r}}{\partial v'} \right| \left| \frac{\partial(u', v')}{\partial(u, v)} \right|.$$

So to show that the elements of surface area in the two parameterizations are equal, i.e., that

$$\left| \frac{\partial \mathbf{r}}{\partial u'} \times \frac{\partial \mathbf{r}}{\partial v'} \right| du' dv' = \left| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right| du dv = \left| \frac{\partial \mathbf{r}}{\partial u'} \times \frac{\partial \mathbf{r}}{\partial v'} \right| \left| \frac{\partial(u', v')}{\partial(u, v)} \right| du dv,$$

we need to show that

$$du' dv' = \left| \frac{\partial(u', v')}{\partial(u, v)} \right| du dv.$$

But this is just the change of variables formula for the transformation \mathbf{T} relating the element of area in the u', v' -plane to the element of area in the u, v -plane.

The above analysis shows that to define surface integrals in terms of parametric representations and to know that the answer depends only on the surface, we must first prove the change of variables formula for double integrals. That can be done with some effort, but the technical hurdles are difficult to surmount. If you go back to the discussion of integrals in polar coordinates, you can see the source of some of the problems. The double integral is defined originally in terms of *rectilinear* partitions by a network of lines parallel to the coordinate axes. To use the other coordinate system, we need to use *polar rectangles* (curvilinear rectangles in the general case), so we have to prove that the limits for curvilinear partitions and rectilinear partitions are the same. We leave further discussion of such issues for a course in real analysis.

We noted in passing that both for integrals over surfaces and for change of variables, the relevant functions should be *one-to-one* except possibly on boundaries. That will insure that sets with positive area are counted exactly once in any integrals. There is a related property of the ‘correction factor’, namely that it should not

vanish. Consider the case of surface integrals. We have

$$dS = \left| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right| du dv$$

and we generally want it to be true that

$$\left| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right| \neq 0.$$

Otherwise, the cross product will vanish, meaning that the two tangent vectors generating the sides of the parallelogram are *collinear*. That means that the tangent plane might degenerate into a line (or even a point), and our whole analysis might break down. Points at which this happens are called *singular points*, and it is often (but not always) the case that the one-to-one property fails near such a point. Similar remarks apply to the change of variables formula where one wants the Jacobian not to vanish.

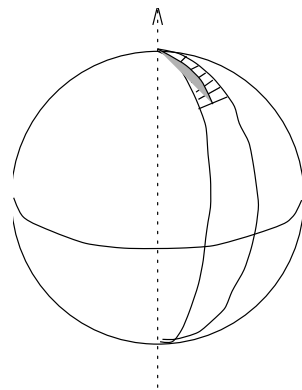
Example 91

$$\mathbf{r} = \langle a \sin \phi \cos \theta, a \sin \phi \sin \theta, a \cos \phi \rangle$$

provides a parametric representation of a sphere of radius a centered at the origin. For $(\phi, \theta) = (0, 0)$, we have

$$\begin{aligned} \frac{\partial \mathbf{r}}{\partial \phi} &= \langle a, 0, 0 \rangle \\ \frac{\partial \mathbf{r}}{\partial \theta} &= \langle 0, 0, 0 \rangle \end{aligned}$$

so the cross product and the correction factor are zero. Of course, the parametric representation fails to be one-to-one at the corresponding point (the North Pole of the sphere) because the entire boundary segment $\phi = 0, 0 \leq \theta \leq 2\pi$ maps into it. Of course, this does not affect the validity of the integration formulas because the point is on the boundary of the parameter domain.



Generalizations to Higher Dimensions The change of variables formula may be generalized to three dimensions. Let E be a region in let $f(x, y, z)$ denote a reasonably smooth function defined on E . Suppose we change variables by smooth functions $x = x(u, v, w), y = y(u, v, w), z = z(u, v, w)$ so that the vector valued function $\mathbf{R}^3 \rightarrow \mathbf{R}^3$ given by

$$\mathbf{r} = \mathbf{r}(u, v, w) = \langle x(u, v, w), y(u, v, w), z(u, v, w) \rangle$$

carries a subset U in the parameter space onto E . Suppose moreover that this function is one-to-one except possibly on the boundary of U . Define the Jacobian of the transformation to be

$$\frac{\partial(x, y, z)}{\partial(u, v, w)} = \det \begin{bmatrix} \partial x / \partial u & \partial y / \partial u & \partial z / \partial u \\ \partial x / \partial v & \partial y / \partial v & \partial z / \partial v \\ \partial x / \partial w & \partial y / \partial w & \partial z / \partial w \end{bmatrix}.$$

Then the change of variables formula says that

$$\iiint_E f(x, y, z) \, dx \, dy \, dz = \iiint_U f(x(u, v, w), y(u, v, w), z(u, v, w)) \left| \frac{\partial(x, y, z)}{\partial(u, v, w)} \right| \, du \, dv \, dw.$$

Example 92 Consider the transformation

$$\mathbf{r} = \mathbf{r}(\rho, \phi, \theta) = \langle \rho \sin \phi \cos \theta, \rho \sin \phi \sin \theta, \rho \cos \phi \rangle$$

which relates the rectangular coordinates (x, y, z) of a point in space to its spherical coordinates. We have

$$\begin{array}{lll} \frac{\partial x}{\partial \rho} = \sin \phi \cos \theta & \frac{\partial y}{\partial \rho} = \sin \phi \sin \theta & \frac{\partial z}{\partial \rho} = \cos \phi \\ \frac{\partial x}{\partial \phi} = \rho \cos \phi \cos \theta & \frac{\partial y}{\partial \phi} = \rho \cos \phi \sin \theta & \frac{\partial z}{\partial \phi} = -\rho \sin \phi \\ \frac{\partial x}{\partial \theta} = -\rho \sin \phi \sin \theta & \frac{\partial y}{\partial \theta} = \rho \sin \phi \cos \theta & \frac{\partial z}{\partial \theta} = 0. \end{array}$$

We leave it to you to calculate the determinant of this 3×3 array. It is not surprising that the answer is

$$\frac{\partial(x, y, z)}{\partial(\rho, \phi, \theta)} = \rho^2 \sin \phi.$$

The change of variables formula in fact holds in \mathbf{R}^n for any positive integer n , but of course to make sense of it, one must first define multiple integrals and determinants in higher dimensions. We shall come back to such matters later in this text.

Exercises for 4.10.

- In each case find the Jacobian of the indicated transformation.
 - $x = uv, y = u^2 + v^2$.
 - $x = au + bv, y = cu + dv$ where a, b, c, d are constants.
 - $x = r \cos \theta, y = 2r \sin \theta$.
- The transformation given by $x = 2u + 3v, y = u + 2v$ carries the unit square $0 \leq u \leq 1, 0 \leq v \leq 1$ into a parallelogram. Show that the parallelogram has area 1.
- Use the transformation $x = ar \cos \theta, y = br \sin \theta$ to evaluate the integral $\iint_D (x^2 + y^2) dA$ where D is the region inside the ellipse $x^2/a^2 + y^2/b^2 = 1$.

4. Use the transformation $x = au, y = bv, z = cw$ to show that the volume inside the ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$ is $(4/3)\pi abc$.
5. Show that the Jacobian of the transformation

$$\begin{aligned}x &= \rho \sin \phi \cos \theta \\y &= \rho \sin \phi \sin \theta \\z &= \rho \cos \phi\end{aligned}$$

is $\rho^2 \sin \phi$.

4.11 Properties of the Integral

Multiple integrals satisfy the rules you learned for single variable integrals. It isn't necessary at this point for you to go through the proofs of the rules, and in the ordinary course of events, most people just use these rules without having to be told about them. Unfortunately, there are some hypotheses which must hold for a rule to apply, so it is possible to go wrong. Such matters, including proofs are usually studied in a course in real analysis. Just for the record, here are some of the relevant rules, without the proofs. We state them for double integrals, but they also hold much more generally, e.g., for triple integrals.

(i) **Existence.** Let D be a closed, bounded subset of \mathbf{R}^2 such that its boundary consists of a finite set of smooth curves. If f is continuous on D , then $\iint_D f \, dA$ exists.

(ii) **Linearity.** If f and g are both integrable on D , then so is $af + bg$ for any constants a and b , and

$$\iint_D (af + bg) \, dA = a \iint_D f \, dA + b \iint_D g \, dA.$$

(iii) **Additivity.** If D_1 and D_2 are disjoint sets such that f is integrable on both, then f is integrable on their union $D_1 \cup D_2$, and

$$\iint_{D_1 \cup D_2} f \, dA = \iint_{D_1} f \, dA + \iint_{D_2} f \, dA.$$

This can be extended slightly in most cases to regions which intersect only on their common boundary, provided that boundary is at worst a finite collection of smooth curves. This rule was referred to earlier when we discussed decomposing a general region into ones bounded by graphs. (iv) **Inequality Property.** If f and g are

integrable on D and $f(\mathbf{r}) \leq g(\mathbf{r})$ for every point \mathbf{r} in D , then

$$\iint_D f \, dA \leq \iint_D g \, dA.$$

(v) **Average Value Property** Suppose f is continuous on the closed bounded set D in \mathbf{R}^2 . Then there is a point \mathbf{r}_0 in D such that

$$f(\mathbf{r}_0) = \frac{1}{A(D)} \iint_D f(\mathbf{r}) \, dA,$$

where $A(D)$ is the area of the region D .

Rule (v) is actually a simple consequence of rule (iv), so let's derive it. Since f is continuous, and since the domain D is *closed and bounded*, f takes on a maximum value M and a minimum value m . Since, $m \leq f(\mathbf{r}) \leq M$, rule (iv) gives

$$\iint_D m \, dA \leq \iint_D f(\mathbf{r}) \, dA \leq \iint_D M \, dA.$$

The constants m and M may be moved out of the integrals (by rule (ii)), so dividing by $A(D) = \iint_D dA$, we obtain

$$m \leq \frac{1}{A(D)} \iint_D f(\mathbf{r}) \, dA \leq M.$$

Again, since f is continuous, it assumes every possible value between its minimum value m and its maximum value M . (That is called the *intermediate value property of continuous functions*.) the quantity in the middle of the inequality is one such value, so there is a point \mathbf{r}_0 in D such that

$$f(\mathbf{r}_0) = \frac{1}{A(D)} \iint_D f(\mathbf{r}) \, dA.$$

For the proofs of the facts about continuous functions that we just used, we must refer you to that same course in real analysis. (Maybe your curiosity will be whetted enough to study the subject some day.)

Non-integrable Functions, Measure Theory, and Some Bizarre Phenomena The significance of rule (i) is not too clear unless one has seen an example of a non-integrable function. 'Non-integrable' does not mean that you can't calculate the integral. It means that when you consider the (Riemann) sums which are supposed to approximate the integral, they don't stabilize around any fixed limit as the dissections get finer and finer. Since the integral is defined as that limit, there can be no well defined integral if there is no limit.

Everyone's favorite non-integrable function (in the plane) is defined as follows. Let the domain D be the unit square $0 \leq x \leq 1, 0 \leq y \leq 1$. Define $f(x, y) = 1$ if both coordinates x and y are rational numbers and $f(x, y) = 0$ if either x or y is not a

rational number. Note that this is a highly discontinuous function since near any point there are both points of the first type and points of the second type. When one forms a Riemann sum

$$\sum f(x, y) \Delta A$$

the point (x, y) in a typical element of area in the dissection could be either of the first type or of the second type. If the choices are made so they are all of the first type, the answer will be $\sum \Delta A$ which is just the area of D , which is 1. If the choices are made so they are all of the second type, then the sum will be 0. No matter how fine the dissection is, we can always make the choices that way, so 1 and 0 are always possible outcomes for the Riemann sum. That shows there is no stable limit for these sums.

Late in the 19th century, mathematicians became dissatisfied with the definition of the integral given by Riemann. The example indicates one of the reasons. Rational numbers are much ‘rarer’ than irrational numbers. One can make a plausible argument that the function f defined above is practically always equal to 0, so its integral should be zero. In response to such concerns, the French mathematician Lebesgue developed a more general concept of an integral which subsumes Riemann’s theory but allows more functions (still not all functions) to be integrable. Lebesgue’s theory was vastly generalized in the 20th century to what is called general *measure theory*. Although the basic ideas of this theory are quite simple, it requires a high level of technical proficiency with abstract concepts and proofs, so it is usually postponed until the first year of graduate study in mathematics. For this reason, although the theory is very general and very powerful, it is not well understood by non-mathematicians. Despite that, many ideas in modern physics (and other areas such as probability and statistics) use measure theory, so you will probably encounter it in some form.

Of course, I can’t really tell you in a brief section what this theory is about, but it is worth giving some hints. To explain it, consider the physical problem of determining the total mass of a mass distribution distributed over space. As usual, we denote the density function $\delta(x, y, z)$. Then for any subset E of \mathbf{R}^3 , the mass inside E is given by

$$m(E) = \iiint_E \delta(x, y, z) dV.$$

Such a function is called a *set function*. It attaches to each set a number, called the *measure* of the set; in this case the mass inside it. So far, this is just a matter of notation, but the general theory allows us to consider measures $m(E)$ which cannot be expressed in terms of any density function. For example, suppose we have a collection of point masses m_1, m_2, \dots, m_n at positions $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$. Then $m(E)$ can be defined to be the sum of the masses which happen to be inside E . There is no *function* $\delta(x, y, z)$ which could give the density of such a mass distribution. $\delta(x, y, z)$ would have to be zero except at the points where the masses are. At such points, the density function would have to be infinite, but there would have to be some way to account for possible differences among the masses. The power of the general theory is that one can apply it in a very similar way to continuous

distributions (for which there is a density function) or to discrete distributions or even to combinations of the two.

Given such a set function $m(E)$, one can define the integral of a function f with respect to that measure. Namely, consider appropriate dissections of the region E into subsets E_i , and form sums

$$\sum_i f(\mathbf{r}_i) m(E_i),$$

where \mathbf{r}_i is in E_i . If these stabilize around a limiting value, that limit is called the integral and is denoted

$$\iiint_E f dm.$$

(The details of what dissections to allow and how to take the limit are quite involved.) To see how general this is, note that for the measure associated as above with a finite set of point masses m_1, m_2, \dots, m_n at points $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$ in E , we have simply

$$\iiint_E f dm = \sum_i f(\mathbf{r}_i) m_i.$$

This also generalizes the ordinary concept of integral since the set function $V(E)$ which attaches to each set its volume is a perfectly good measure, and the resulting integral is just $\iiint_E f dV$.

There are two points raised in the above discussion meriting further discussion. The concept of a density function is so useful, that physicists were reluctant to give it up, even in the case of a point mass. For this reason, the Nobel prize winning physicist Dirac invented a ‘function’ which today is called the ‘Dirac δ function’. Here is how he reasoned. We will explain it in the one-dimensional case, but it generalizes easily to two and three dimensions. Consider a unit mass on the real line \mathbf{R} placed at the origin. Associated with this is a measure on subsets I of \mathbf{R} defined by $m(I) = 1$ if I contains the origin and $m(I) = 0$ otherwise. This measure can also be described by its so-called distribution function: $F(x) = 0$ if $x < 0$, and $F(x) = 1$ if $x > 0$, which gives the value of $m(I)$ for $I = (-\infty, x)$. (It is not hard to see that if you know the distribution function $F(x)$, you can reconstruct the measure.) Dirac’s idea amounts to choosing the density function $\delta(x)$ to be the derivative $F'(x)$. This derivative is clearly 0 except at $x = 0$ where it is undefined. (F is not even continuous at 0, so it certainly isn’t differentiable.) However, imagine that the derivative did make sense at zero, and that the usual relation between derivatives and integrals holds for this derivative. Then we would have

$$\int_a^b \delta(x) dx = \int_a^b F'(x) dx = F(b) - F(a)$$

and this would be 0 unless $a < 0 \leq b$ in which case it would be 1. Thus the integral $\int_a^b \delta(x) dx$ would have exactly the right properties for the set function associated with a point mass at the origin. Physicists liked this so much that they adopted

the useful fiction that there actually is such a function $\delta(x)$, and they use it freely nowadays in formulas and calculations. It all works out correctly, if one is careful, because a statement involving the δ function usually makes sense as a statement about set functions if one puts integral signs around it. We will see the δ function on several occasions in this course.

It turns out that Dirac had a better idea than he might have realized. Subsequently, mathematicians developed a rigorous concept called a *generalized function* or *distribution* which provides a firm foundation for practically everything the physicists want to do. (The main exponent of this theory was the French mathematician Laurent Schwartz.)

The final remark concerns the concept of measure. We said that the set function $m(E)$ is defined for *every* subset, but that was wrong. If the measure is to have reasonable properties, for example, if it gives the expected values for length, area, or volume (depending on whether we are discussing the theory in \mathbf{R} , \mathbf{R}^2 , or \mathbf{R}^3), then it is not possible for every subset to have a measure. There must be what are called *non-measurable sets*. There is an interesting sidelight to the theory called the *Banach–Tarski Paradox*. Banach and Tarski showed that it is possible to take a solid sphere of radius 1 in \mathbf{R}^3 and decompose it into a small number of *non-overlapping* subsets (about 5), move these sets by rigid motions (combinations of translations and rotations), and then reassemble them into two non-overlapping solid spheres, each of radius 1. In this process, no point in any of the three spheres is unaccounted for. This seems to say that $2 = 1$, and that would indeed be a consequence if the subsets of the dissection were measurable sets and so had well defined volumes. (Volume certainly has to satisfy the additivity rule for non-overlapping sets, and it is certainly not changed by translation and rotations.) However, the subsets of the dissection are *not* measurable sets, so there is no contradiction, just an apparent contradiction or *paradox*. Moreover, the argument just shows that the relevant dissection *exists*; it doesn't give a physically realizable method to do it. In fact, no one believes that this is physically possible.

Exercises for 4.11.

1. Use

$$-|f(\mathbf{r})| \leq f(\mathbf{r}) \leq |f(\mathbf{r})|$$

and rule (iv) in this section to prove

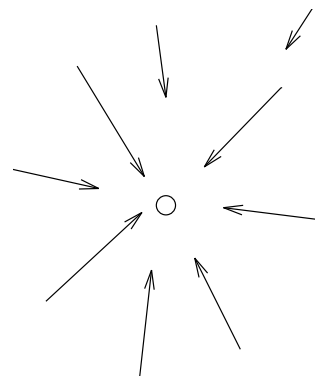
$$\left| \iint_D f \, dA \right| \leq \iint_D |f| \, dA.$$

Chapter 5

Caculus of Vector Fields

5.1 Vector Fields

Multidimensional calculus may be defined as the study of derivatives and integrals for functions $\mathbf{R}^n \rightarrow \mathbf{R}^m$ for various n and m . We have mostly studied scalar valued functions $\mathbf{R}^n \rightarrow \mathbf{R}$ (where $n = 2$ or $n = 3$.) In physics, such functions are called *scalar fields*. We now want to study functions $\mathbf{F} : \mathbf{R}^n \rightarrow \mathbf{R}^n$ (again with $n = 2$ or $n = 3$.) These are called *vector fields*. In general, one wants to picture the set on which the function is defined (the domain) and the set in which it assumes values as being in different places, but in the case $m = n$, they are both in the same place, namely \mathbf{R}^n . Hence, there is another way to view such a function, and it turns out to be quite useful in physics. At each point in the domain of the function \mathbf{F} , imagine the vector $\mathbf{F}(\mathbf{r})$ placed with its tail at the point.



Example 93 Consider the function $\mathbf{F} : \mathbf{R}^3 \rightarrow \mathbf{R}^3$ defined by

$$\mathbf{F}(\mathbf{r}) = -\frac{GM}{|\mathbf{r}|^2} \mathbf{u}_\rho$$

where $\mathbf{u}_\rho = \frac{\mathbf{r}}{|\mathbf{r}|}$ is a unit vector pointing directly away from the origin. This function gives the gravitational force on a test particle at the point with position vector \mathbf{r} due to a mass M at the origin. It certainly makes sense in this case to view the vector $\mathbf{F}(\mathbf{r})$ with its tail at the point, because that is where the force would act on a test particle. Because of the minus sign, the force vector in fact points to the origin. As we get closer to the origin its magnitude increases. It is not defined at the origin, which should be excluded from the domain of the function \mathbf{F} .

This vector field can also be specified in rectangular coordinates by using $\mathbf{r} =$

$\langle x, y, z \rangle$ and $|\mathbf{r}| = (x^2 + y^2 + z^2)^{1/2}$. After some algebra, we get

$$\mathbf{F}(x, y, z) = -GM \left\langle \frac{x}{(x^2 + y^2 + z^2)^{3/2}}, \frac{y}{(x^2 + y^2 + z^2)^{3/2}}, \frac{z}{(x^2 + y^2 + z^2)^{3/2}} \right\rangle.$$

One of the problems in dealing with vector fields is to see through complicated algebra to what is often quite simple geometry. The above expression is a good example.

You can always think of a vector field as specifying the *force* at each point in space, (or for plane vector fields at each point in the plane.) Vector fields are used extensively in this way in the theory of gravitation and also in electromagnetic theory.

Example 94 Consider the vector field in \mathbf{R}^2 defined by

$$\mathbf{v}(\mathbf{r}) = r\mathbf{u}_\theta$$

where \mathbf{u}_θ is the unit vector in the positive θ direction as discussed previously. This can also be expressed in rectangular coordinates

$$\mathbf{v}(x, y) = \langle -y, x \rangle.$$

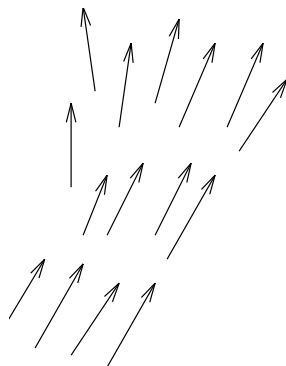
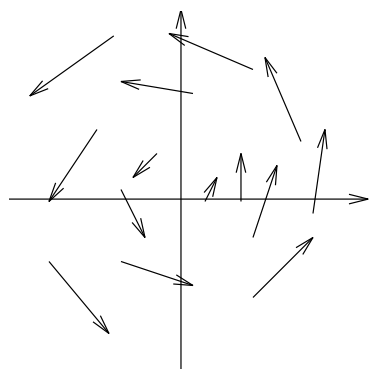
To see this, note that

$$\langle -y, x \rangle = \langle -r \sin \theta, r \cos \theta \rangle = r \langle -\sin \theta, \cos \theta \rangle = r\mathbf{u}_\theta.$$

(See Chapter 1, Section 2.)

At each point in the plane, the vector $\mathbf{v}(\mathbf{r})$ is perpendicular to the position vector \mathbf{r} for that point. It is also tangent to a circle through the point with center at the origin, and it is directed counter-clockwise. We have $|\mathbf{v}(\mathbf{r})| = r$, so the magnitude increases as we move away from the origin (where $\mathbf{v}(\mathbf{0}) = \mathbf{0}$.) One can get a physical picture of this vector field as follows. Imagine a disk (such as a phonograph record) rotating with constant angular velocity ω . Suppose that, by a new photographic process, it is possible to snap a picture which shows the *velocity vector* \mathbf{v} of each element of the disk at the instant the picture is taken. These velocity vectors will look like the vector field described above, except for a factor of proportionality. They will be tangent to circles centered at the origin, and their magnitudes will be proportional to the distance to the center ($|\mathbf{v}| = \omega r$). Note that in this model, the picture will be the same whenever the picture is snapped. At a given point in the plane with position vector \mathbf{r} , the velocity vector $\mathbf{v}(\mathbf{r})$ will just depend on \mathbf{r} , although the particular element of the disk which happens to be passing through that point will change (unless the camera, i.e., the coordinate system, rotates with the disk). Of course, we could envision a somewhat more complicated situation in which the disk speeds up or slows down, and then \mathbf{v} would be a function both of position \mathbf{r} and time t .

The above discussion suggests a second physical model to picture vector fields, that of *fluid flow*. The example was of a 2-dimensional field, but the idea works just

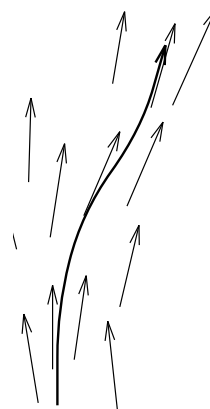


as well in \mathbf{R}^3 . Imagine a fluid flowing in a region of space, for example, Lake Michigan, or the atmosphere of the Earth. In the example, the relative positions of the elements of the medium (a phonograph record) do not change, but in general, that won't be the case. Imagine a super hologram which will show at each instant the velocity vector \mathbf{v} of each element of the fluid. This will generally be a function $\mathbf{v} = \mathbf{v}(\mathbf{r}, t)$ of position and of time. If we ignore the dependence on time, we get a vector function $\mathbf{v} : \mathbf{R}^3 \rightarrow \mathbf{R}^3$, or vector field. (In fluid dynamics, one often distinguishes between *steady flows* where \mathbf{v} does not depend on time and *time dependent* flows where it does.) In fluid mechanics, one is actually more interested in the *momentum* field defined by $\mathbf{F}(\mathbf{r}, t) = \delta(\mathbf{r}, t)\mathbf{v}(\mathbf{r}, t)$, where $\delta(\mathbf{r}, t)$ is a scalar function giving the *density* of the fluid at position \mathbf{r} at time t .

Streamlines Let $\mathbf{F} : \mathbf{R}^n \rightarrow \mathbf{R}^n$ ($n = 2$ or $n = 3$) denote a vector field. (Assume for simplicity that \mathbf{F} does not also depend explicitly on another variable like time.) Consider paths $\mathbf{r} = \mathbf{r}(t)$ with the property that at each point of the path, the vector field $\mathbf{F}(\mathbf{r}(t))$ at that point is tangent to the path. Since the velocity vector is also tangent to the path, we can state this symbolically

$$\frac{d\mathbf{r}}{dt} = \mathbf{F}(\mathbf{r}). \quad (59)$$

(Actually, there should be a factor $c(\mathbf{r})$ of proportionality since the vectors are only parallel. However, we are ordinarily only interested in the geometry of the path, not how it is traced out in time. By a suitable change of parameterization, we can arrange for c to be 1.) Such paths are called either *lines of force* or *streamlines* depending on which physical model we have in mind. In the case of fluid flow, the term “streamline” is self-explanatory. It is supposed to be the path followed by an element of fluid as it moves with the flow. In the case of force fields, the concept, line of force, is a bit mysterious. It was first introduced by Faraday in his work on electricity and magnetism. He thought of the lines of force as having physical reality, behaving almost like elastic bands pulling the objects together (or pushing them apart in the case of repulsive forces.) I will leave further discussion of the meaning to your physics professors.



Example 95 We shall find the streamlines for the vector field in \mathbf{R}^2 defined by

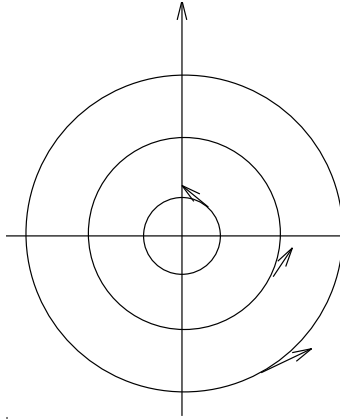
$$\mathbf{v}(x, y) = \langle -y, x \rangle.$$

From the previous discussion of this field, it is clear that they will be circles centered at the origin, but let's see if we can that derive that from equation (59). We get

$$\left\langle \frac{dx}{dt}, \frac{dy}{dt} \right\rangle = \langle -y, x \rangle,$$

which may be rewritten in terms of components

$$\begin{aligned} \frac{dx}{dt} &= -y, \\ \frac{dy}{dt} &= x. \end{aligned}$$



This is a *system* of differential equations, and we won't study general methods for solving systems until later. However, in the 2-dimensional case, there is a simple trick which usually allows one to find the curves. Namely, divide the second equation by the first to obtain

$$\frac{dy}{dx} = -\frac{x}{y}.$$

This may be solved by separation of variables,

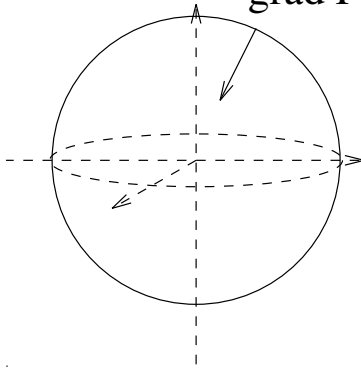
$$y \, dy = -x \, dx$$

$$\text{or} \quad \frac{y^2}{2} = -\frac{x^2}{2} + c$$

$$\text{or} \quad x^2 + y^2 = 2c.$$

These curves are circles centered at the origin as expected. Note that we don't learn how they are traced out as functions of t .

grad F



Gradient Fields Given a scalar field $f : \mathbf{R}^n \rightarrow \mathbf{R}$, we may always form a vector field by taking the gradient.

$$\mathbf{F}(\mathbf{r}) = \nabla f(\mathbf{r}).$$

Example 96 Let $f(\mathbf{r}) = \frac{GM}{\rho} = \frac{GM}{\sqrt{x^2 + y^2 + z^2}}$ for $\mathbf{r} \neq \mathbf{0}$. We in essence calculated the gradient of this function in Example 2, Section 7, Chapter III. The answer is

$$\nabla f = -\frac{GM}{\rho^2} \mathbf{u}_\rho$$

which is the gravitational field of a point mass at the origin.

Example 97 A metal plate is heated so the temperature at (x, y) is given by $T(x, y) = x^2 + 2y^2$. A heat seeking robot starts at the point $(1, 2)$. We shall find the path it follows. "Heat seeking" means that at any point, it will move in the direction of maximum increase of temperature, i.e., in the direction of $\nabla T = \langle 2x, 4y \rangle$. Since we aren't told how fast it responds to the temperature gradient, its equations of motion will be

$$\frac{dx}{dt} = c(2x),$$

$$\frac{dy}{dt} = c(4y),$$

where $c = c(x, y)$ is an unknown function. We can't actually determine the motion along the path, but using the same trick as above, we can determine its shape.

$$\frac{dy}{dx} = 2 \frac{y}{x}$$

$$\text{so} \quad \frac{dy}{y} = 2 \frac{dx}{x}$$

$$\text{and} \quad \ln |y| = \ln |x|^2 + c.$$

If we exponentiate, we obtain

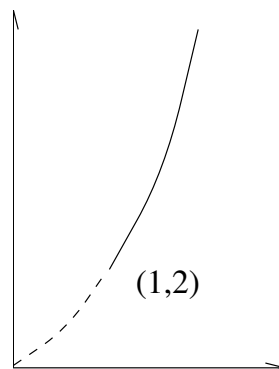
$$|y| = |x|^2 e^c = C|x|^2$$

where $C = e^c$ is just another constant. Putting $x = 1, y = 2$ (where the robot starts) yields $C = 2$. One can make a convincing argument that, considering the starting point and the nature of the possible path, we may take x and y positive, so the solution is

$$y = 2x^2,$$

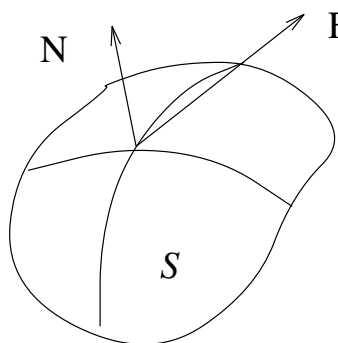
which describes a parabola. Note that, as before, the method depends on this being a problem in the plane. If we had a z to deal with, it wouldn't have worked.

It should be noted that while gradient fields are quite important, not every field is a gradient. As we shall see later, gradient fields must have the property that line integrals depend only on end points, and as we saw earlier, not every force field has that property.



Exercises for 5.1.

- Sketch the following vector fields by drawing some vectors at representative points.
 - $\mathbf{F}(x, y) = x^2\mathbf{i} + y^2\mathbf{j}$.
 - $\mathbf{F}(x, y) = y^2\mathbf{i} + x^2\mathbf{j}$.
 - $\mathbf{F}(x, y) = 3\mathbf{i} - 4\mathbf{j}$.
 - $\mathbf{F}(x, y, z) = \langle x, y, z \rangle$.
 - $\mathbf{F}(x, y) = (x\mathbf{i} - y\mathbf{j})(x^2 + y^2)^{-1/2}$.
- For each of the scalar fields, sketch its gradient field by drawing vectors at representative points
 - $f(x, y) = x + 3y$.
 - $f(x, y) = 3x^2 + 4y^2$.
 - $f(x, y, z) = x^2 + y^2$.
- Find the streamlines of the plane vector field defined by $\mathbf{F}(\mathbf{r}) = y\mathbf{i} + x\mathbf{j}$.
- The shape of a mountain is described by the equation $z = 4000 - x^2 - 4y^2$. There is a shelter at $(10, 14)$. Find a path from the shelter to the bottom of the mountain (at $z = 0$) which descends everywhere as rapidly as possible. Describe the projection of the path in the x, y -plane, thought of as a map of the mountain, rather than the actual path on the mountain.



5.2 Surface Integrals for Vector Fields

In this section, we shall use the notation \mathbf{F} to denote a general vector field. The mathematics won't care, but you may find it useful on occasion to think of it as either a force field or a momentum field for a fluid flow. Also, any dependence of \mathbf{F} on time won't matter in what we do, so we shall assume $\mathbf{F} = \mathbf{F}(\mathbf{r})$ is a function of position alone.

Let $\mathbf{F} : \mathbf{R}^3 \rightarrow \mathbf{R}^3$ denote a vector field, and suppose \mathcal{S} is a surface in \mathbf{R}^3 . Suppose everything is reasonably smooth, so we don't have to worry about singularities or other bizarre behavior. At each point on \mathcal{S} , choose a unit vector \mathbf{N} perpendicular to the surface. (Hence, \mathbf{N} will usually vary from point to point on \mathcal{S} .) Consider the scalar valued function of position $\mathbf{F} \cdot \mathbf{N}$. The integral of this function over the surface

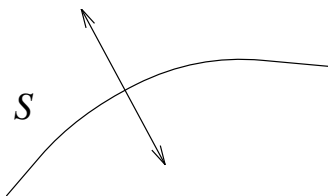
$$\iint_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} \, dS$$

is called the surface integral of the vector field \mathbf{F} over the surface \mathcal{S} . One must exercise care in choosing the normal vectors since at any given point on the surface, there will be *two* unit vectors perpendicular to the surface. Clearly, the direction of the normal should not change precipitously between nearby points on the surface. (The choice of the normals, called the *orientation* of the surface, can be a bit subtle, and we shall come back to this important point later.) One often uses the notation

$$d\mathbf{S} = \mathbf{N} \, dS$$

so that the element of surface area becomes a *vector quantity*. Then the notation for the surface integral becomes

$$\iint_{\mathcal{S}} \mathbf{F} \cdot d\mathbf{S}.$$



2 possible \mathbf{N} 's

Other notational variations include using a single integral sign, \int , and using some other symbol, such as $d\sigma$ or $d\mathbf{A}$, for the vector element of surface area.

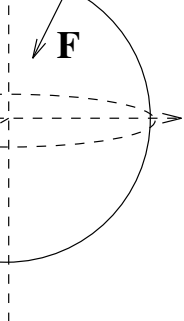
Example 98 Let $\mathbf{F}(\mathbf{r}) = -\frac{GM}{\rho^2} \mathbf{u}_\rho$ and let \mathcal{S} be the surface of a sphere of radius a . Assume all the normals point outward from the origin. Thus in this case, at each point on the surface, $\mathbf{N} = \mathbf{u}_\rho$. Hence, since $\rho = a$ on the sphere \mathcal{S} ,

$$\mathbf{F} \cdot \mathbf{N} = -\frac{GM}{a^2} \mathbf{u}_\rho \cdot \mathbf{u}_\rho = -\frac{GM}{a^2}.$$

Thus,

$$\iint_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} \, dS = -\frac{GM}{a^2} \iint_{\mathcal{S}} dS = -\frac{GM}{a^2} (4\pi a^2) = -4\pi GM.$$

Notice that most of the algebra is really superfluous here. The vector field is parallel to the unit normal, so the dot product is just the magnitude of the vector field on the sphere, $-\frac{GM}{a^2}$. However, since this is constant, integrating it over the sphere



just results in that constant times the surface area of the sphere. With a bit of practice, you should be able to do such a simple example in your head.

Interpretation for Fluid Flow Let \mathbf{F} denote the momentum field of a steady fluid flow (i.e., $\mathbf{F}(\mathbf{r}) = \delta(\mathbf{r})\mathbf{v}(\mathbf{r})$, where $\delta(\mathbf{r})$ denotes density and $\mathbf{v}(\mathbf{r})$ velocity at the point with position vector \mathbf{r} .) Then it turns out that the surface integral $\iint_S \mathbf{F} \cdot \mathbf{N} dS$ gives the *rate of flow* or the *flux* of the fluid through the surface. To see this, consider first the special case where \mathbf{F} , δ , and \mathbf{v} are all constant, and the surface is a parallelogram spanned by vectors \mathbf{a} and \mathbf{b} . In one unit of time, each element of fluid is displaced by the vector \mathbf{v} , so all the fluid which flows through the parallelogram will lie within the parallelepiped spanned by \mathbf{a} , \mathbf{b} , and \mathbf{v} . The volume contained therein is $\mathbf{v} \cdot (\mathbf{a} \times \mathbf{b})$. (See Chapter I, Section 4.)

Hence, the mass passing through the parallelogram per unit time is

$$\delta \mathbf{v} \cdot (\mathbf{a} \times \mathbf{b}) = (\delta \mathbf{v}) \cdot (\mathbf{a} \times \mathbf{b}) = \mathbf{F} \cdot (\mathbf{a} \times \mathbf{b}).$$

Since $A = |\mathbf{a} \times \mathbf{b}|$ is the area of the parallelogram, we may rewrite $\mathbf{a} \times \mathbf{b} = AN$ where \mathbf{N} is a unit vector perpendicular to the parallelogram. Hence,

$$\text{flux through parallelogram} = \mathbf{F} \cdot \mathbf{N} A.$$

Note the significance of the direction of the normal vector in this context. If \mathbf{F} and \mathbf{N} point to the same side of the parallelogram, the flux is positive, and if they point to opposite sides, the flux is negative.

The above analysis may be extended to curved surfaces. The quantity $\mathbf{F} \cdot \mathbf{N} dS$ represents the flux through a small element of surface area dS , and integrating gives the net flux through the entire surface.

It is common to use the term *flux* quite generally, even where the fluid flow model is not appropriate. For example, in gravitational theory (and also the theory of electromagnetic fields), there is nothing “flowing” in the ordinary sense. In this case, the flux is interpreted as a measure of the “number” of lines of force passing through the surface. Since this is a mathematics course, I will leave such matters for your physics professor to explain.

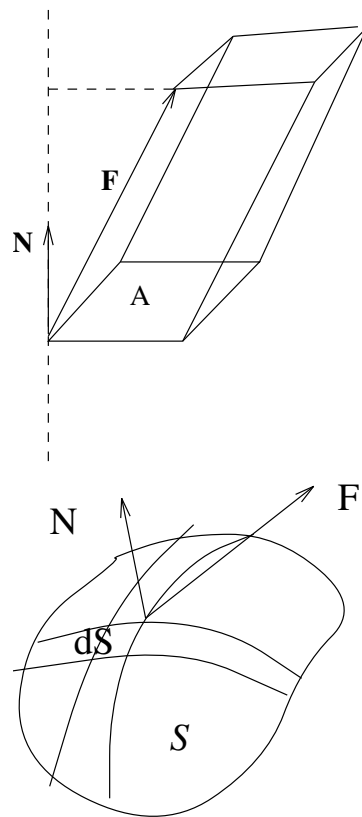
We continue with further examples.

Example 99 Let $\mathbf{F}(\mathbf{r}) = z\mathbf{k}$ and let S be the top hemisphere of the sphere of radius a centered at the origin. Use outward normals. As above $\mathbf{N} = \mathbf{u}_\rho$. On the surface of the sphere, we use ϕ, θ as intrinsic coordinates as usual. Then $\rho = a$, $z = a \cos \phi$, and $dS = a^2 \sin \phi d\phi d\theta$. Moreover, $\mathbf{N} = \mathbf{u}_\rho$, and since the angle between \mathbf{N} and \mathbf{k} is ϕ , we have $\mathbf{k} \cdot \mathbf{N} = \cos \phi$ and

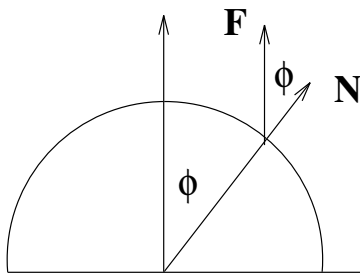
$$\mathbf{F} \cdot \mathbf{N} = z\mathbf{k} \cdot \mathbf{N} = a \cos \phi \cos \phi = a \cos^2 \phi.$$

Hence,

$$\iint_S \mathbf{F} \cdot \mathbf{N} dS = \int_0^{2\pi} \int_0^{\pi/2} a \cos^2 \phi a^2 \sin \phi d\phi d\theta$$



where the limits were chosen to cover the hemisphere. Thus,



$$\begin{aligned}\iint_S \mathbf{F} \cdot \mathbf{N} \, dS &= a^3 \int_0^{2\pi} \int_0^{\pi/2} \cos^2 \phi \sin \phi \, d\phi \, d\theta \\ &= a^3 (2\pi) \left. -\frac{\cos^3 \phi}{3} \right|_0^{\pi/2} \\ &= \frac{2\pi a^3}{3} (1) = \frac{2\pi a^3}{3}.\end{aligned}$$

If the integration had been over the entire sphere, the answer would have been $4\pi a^3/3$. Can you see why without actually doing the calculation? (Try drawing a picture showing \mathbf{F} and \mathbf{N} on the bottom hemisphere and comparing with the picture on the top hemisphere.)

It is worth remembering that on the surface of a sphere of radius a ,

$$d\mathbf{S} = \mathbf{N} \, dS = \mathbf{u}_\rho \, a^2 \sin \phi \, d\phi \, d\theta.$$

Similarly, on the *lateral* surface of a right circular cylinder of radius a centered on the z -axis, the outward unit normal vector $\mathbf{N} = \mathbf{u}_r$ points directly away from the z -axis, and $dS = a \, d\theta \, dz$, so

$$d\mathbf{S} = \mathbf{N} \, dS = \mathbf{u}_r \, a \, d\theta \, dz.$$

Surface Integrals for Parametrically Defined Surfaces In the previous chapter, we calculated the *surface area* of a parametrically defined surface as follows. Each small curvilinear rectangle on the surface was approximated by a tangent parallelogram, the sides of which were the vectors $\mathbf{a} = \frac{\partial \mathbf{r}}{\partial u} \Delta u$ and $\mathbf{b} = \frac{\partial \mathbf{r}}{\partial v} \Delta v$. The area of the latter was then used as an approximation for the area of the former. The same reasoning may be used to calculate flux. The flux through a curvilinear rectangle on the surface may be approximated by the flux through the (tangent) parallelogram, i.e., as above, by $\mathbf{F} \cdot (\mathbf{a} \times \mathbf{b})$ where \mathbf{F} is the value of the vector field at a corner of the parallelogram. Expanding this out yields

$$\mathbf{F}(\mathbf{r}(u, v)) \cdot \left(\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right) \Delta u \, \Delta v.$$

Hence,

$$\iint_S \mathbf{F} \cdot d\mathbf{S} = \iint_D \mathbf{F}(\mathbf{r}(u, v)) \cdot \left(\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right) du \, dv \quad (60)$$

where D is the domain of the parameterizing function $\mathbf{r} = \mathbf{r}(u, v)$. Note that the normal vectors in this case are given by

$$\mathbf{N} = \frac{1}{\left| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right|} \left(\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right)$$

so their directions are determined by the parametric representation. However, if those directions don't square with your expectations, it is easy enough to reverse them. Just change the parametric representation by reversing the roles of the parameters, hence, the order in the cross product.

It is seldom the case that one wants to use formula (60), although it underlies the other calculations. In most cases, the surface is a familiar one such as a sphere or a cylinder, in which case one can visualize the unit normals \mathbf{N} directly, or it is the graph of a function. (See the exercises for some examples where one must resort to the general formula.)

Surface Integrals for the Graph of a Function Suppose the surface \mathcal{S} is the graph of a function $z = f(x, y)$. In that case, we use the parametric representation $\mathbf{r} = \mathbf{r}(x, y) = \langle x, y, f(x, y) \rangle$, and some calculation shows

$$d\mathbf{S} = \frac{\partial \mathbf{r}}{\partial x} \times \frac{\partial \mathbf{r}}{\partial y} dx dy = \langle -f_x, -f_y, 1 \rangle dx dy.$$

(See the corresponding discussion for the surface area of a graph in Section 9, Chapter IV.)

Hence, if the vector field is resolved in components, $\mathbf{F} = \langle F_x, F_y, F_z \rangle$, the surface integral takes the form

$$\iint_{\mathcal{S}} \mathbf{F} \cdot d\mathbf{S} = \iint_D (-F_x f_x - F_y f_y + F_z) dx dy$$

where D is the domain of the function f in the x, y -plane. Note that in this case the unit normal vector

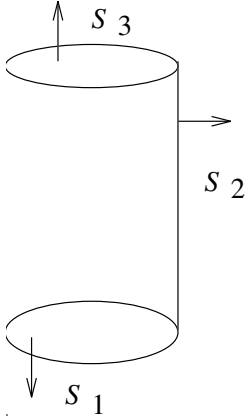
$$\mathbf{N} = \frac{1}{\sqrt{f_x^2 + f_y^2 + 1}} \langle -f_x, -f_y, 1 \rangle$$

points generally upward.

Example 99, (revisited) We use the same hemisphere but view it as the graph of the function $f(x, y) = \sqrt{a^2 - x^2 - y^2}$ with domain D a disk in the x, y -plane of radius a , centered at the origin. Since $\mathbf{F}(x, y, z) = \langle 0, 0, z \rangle$, we need not calculate f_x and f_y . We have

$$\iint_{\mathcal{S}} \mathbf{F} \cdot d\mathbf{S} = \iint_D \underbrace{\sqrt{a^2 - x^2 - y^2}}_z dy dx.$$

However, looking at this carefully reveals it to be the same integral as that for the volume under the hemisphere, which is $(1/2)(4\pi a^3)/3 = 2\pi a^3/3$. Since the calculation of the double integral is so easy, this approach is simpler than the previous approach using the intrinsic coordinates ϕ and θ on the sphere. However, in most cases the intrinsic approach is superior.



Sometimes, the surface needs to be decomposed into pieces in order to calculate the flux.

Example 100 Let $\mathbf{F}(\mathbf{r}) = \mathbf{r} = \langle x, y, z \rangle$. Let \mathcal{S} be the surface enclosing a right circular cylinder of radius a and height h , centered on the z -axis, where this time we include both the top and bottom as well as the lateral surface. We shall find the flux out of the cylinder, that is the normals will be chosen so as to point out of the cylinder.

We decompose the surface into three parts: the bottom \mathcal{S}_1 , the lateral surface \mathcal{S}_2 , and the top \mathcal{S}_3 .

The bottom surface is easiest. The outward unit normal is $\mathbf{N} = -\mathbf{k}$, and $\mathbf{F}(\mathbf{r}) = \mathbf{r} = \langle x, y, 0 \rangle$ is perpendicular to \mathbf{N} for points in the x, y -plane. Hence, $\mathbf{F} \cdot \mathbf{N} = 0$ for \mathcal{S}_1 , so the flux is zero.

The top surface is also not too difficult. The unit normal is $\mathbf{N} = \mathbf{k}$, but in the plane $z = h$, we have $\mathbf{F}(\mathbf{r}) = \langle x, y, h \rangle$. Hence, $\mathbf{F} \cdot \mathbf{N} = h$, and

$$\iint_{\mathcal{S}_3} \mathbf{F} \cdot \mathbf{N} dS = h \iint_{\mathcal{S}_3} dS = h(\pi a^2) = \pi a^2 h$$

since the top surface is a disk of radius a .

Finally, we consider the lateral surface \mathcal{S}_2 . We may resolve \mathbf{F} into horizontal and vertical components by writing

$$\mathbf{F}(\mathbf{r}) = \langle x, y, z \rangle = \langle x, y, 0 \rangle + \langle 0, 0, z \rangle = r\mathbf{u}_r + z\mathbf{k},$$

but since $r = a$ on the cylinder, we have

$$\mathbf{F}(\mathbf{r}) = a\mathbf{u}_r + z\mathbf{k}.$$

On the other hand, the outward unit normal is $\mathbf{N} = \mathbf{u}_r$, so $\mathbf{F} \cdot \mathbf{N} = a\mathbf{u}_r \cdot \mathbf{u}_r + z\mathbf{k} \cdot \mathbf{u}_r = a$. Hence,

$$\iint_{\mathcal{S}_2} \mathbf{F} \cdot \mathbf{N} dS = a \iint_{\mathcal{S}_2} dS = a(2\pi ah) = 2\pi a^2 h.$$

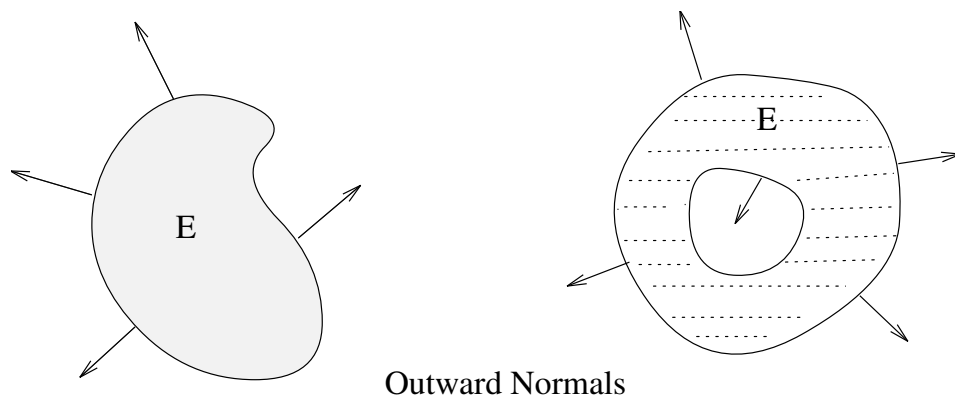
Here we used the fact that the surface area of the lateral surface of cylinder is the height times the circumference of the base. Note that, in the above calculation, the vertical component of the force is irrelevant since the normal is horizontal. Also, the horizontal component is normal to the lateral surface of the cylinder, and that makes calculating the dot product quite easy.

The total flux is the sum for the three surfaces

$$0 + \pi a^2 h + 2\pi a^2 h = 3\pi a^2 h.$$

Orientation As we noted, at each point of a surface, there are generally two possible directions for the unit normal vector. Explaining how one goes about

specifying those directions and seeing how they are related as one moves about the surface is a bit trickier than you might imagine. The simplest case is that of a closed surface, i.e., a surface \mathcal{S} , bounding a solid region E in space. In that case, there are two obvious *orientations* or ways of choosing the normal directions. We may use the *outward* orientation, in which all the normals point away from the region, or we may use the *inward* orientation, in which they all point into the region. If the solid region is something like a solid sphere, cylinder, cube, etc., this is quite easy to visualize. However, the principle works even in regions a bit more complicated than this. Consider for example, the solid region E *between* two concentric spheres. The boundary of that region consists of the two spheres. (It comes in two pieces, or in technical jargon, it is not *connected*.) The outward normals, relative to E point away from the origin on the outer sphere and towards the origin for the inner sphere. The same sort of thing happens for any surface consisting of the (disconnected) boundary of any solid region having one or more holes. (Imagine a hunk of swiss cheese.)

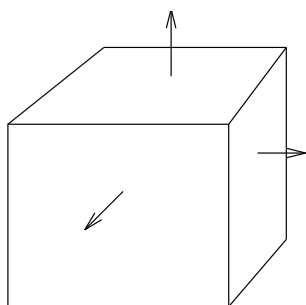


Outward Normals

For surfaces, which are not boundaries of solid regions, it is sometimes not so clear how to choose an orientation. If the surface is given parametrically, $\mathbf{r} = \mathbf{r}(u, v)$, with the parameters given in some specified order, then this implies an orientation, since at each point

$$\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}$$

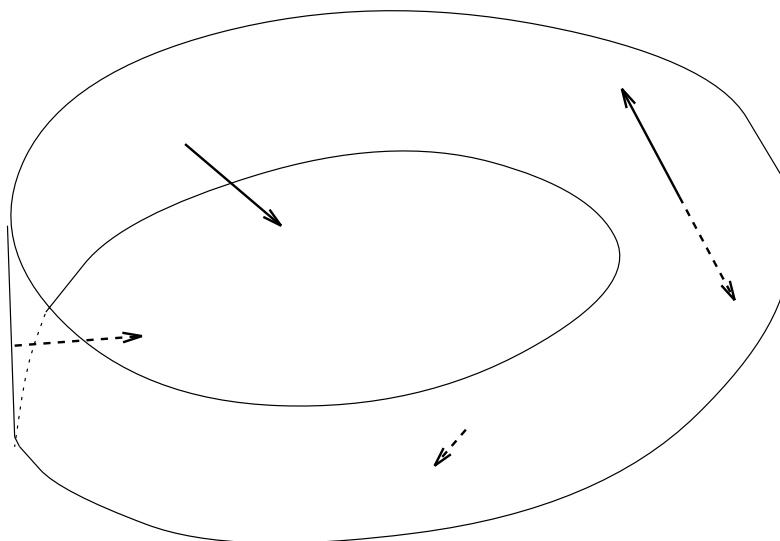
determines the direction of the normal. Suppose however, the surface consists of separate pieces which are connected one to another, but such that no one parameterization works for the entire surface. The closed surface bounding a solid cylinder, including top and bottom, is such an example. Another example, which is not closed, would be the 5 faces of a cubic box open at the top. There is a fairly straightforward way to specify an orientation for such a surface. Suppose that such a surface is decomposed into separate patches, each with a smooth parametric representation. That gives each patch a specific orientation, but it can always be



coherent normals

reversed by reversing the order of the parameters. We hope to be able to choose the orientations for the parametric patches so that they are coherently related. If this is possible, we say the surface has an orientation. The difficulty is that as one crosses an edge which is the common boundary between two adjacent patches, the direction of the normal may change quite radically, so it is a bit difficult to make precise what we mean by saying the normals are coherently related. Fortunately, in most interesting cases, it is fairly clear how the normals on adjacent patches *should be related*. We will return to this point later during our discussion of Stokes's Theorem and related matters.

It is quite surprising that there are fairly simple surfaces in \mathbf{R}^3 without coherent orientations. The so called *Moebius band* is such a surface. One can make a Moebius band by taking a long strip of paper, twisting it once and connecting the ends. (Without the twist, we would get a long narrow “cylinder”.) See the diagram. To see that the Moebius band is not orientable, start somewhere and choose a direction for the normals near the starting point. Then continue around the band, as indicated, choosing normals in roughly the same direction as you progress. When you get all the way around the band, you will find that the normal direction you have “dragged” with you now points *opposite* to the original direction. (To get back the original direction, you will have to go around the band once again.) See the Exercises for a parametric representation of the Moebius band—less one segment—which exhibits this discontinuous reversal of the normal.



It does not make sense to try to calculate the flux through a non-orientable surface such as a Moebius band, since you won't be able to assign positive and negative signs for the contributions from different parts of the surface in a coherent fashion.

Exercises for 5.2.

- Calculate the flux $\iint_{\mathcal{S}} \mathbf{F} \cdot \mathbf{N} \, dS$ through the indicated surface for the indicated vector field.
 - $\mathbf{F} = 2x\mathbf{i} + 3y\mathbf{j} - z\mathbf{k}$, \mathcal{S} the portion of the plane $x + y + z = 1$ in the first octant. Use the normals pointing away from the origin. (You can solve for z and treat the surface as the graph of a function.)
 - $\mathbf{F} = z\mathbf{k}$, \mathcal{S} the lower half of the sphere $x^2 + y^2 + z^2 = 1$. Use normals pointing towards the origin.
 - $\mathbf{F} = \langle 2y, 3x, z \rangle$, \mathcal{S} the portion of the paraboloid $z = 9 - x^2 - y^2$ above the xy -plane. Use normals pointing away from the origin. (The surface is the graph of a function.)
- Let $\mathbf{F} = -y\mathbf{i} + x\mathbf{j}$ and let \mathcal{S} be the portion of the cone $z = \sqrt{x^2 + y^2}$ within the cylinder $x^2 + y^2 = 4$. Find the flux through the surface using normals pointing away from the z -axis. You can treat the surface as the graph of a function, but it might be faster to do the problem by visualizing geometrically the relation of \mathbf{F} to the normals to the surface.
- Let \mathcal{S} be a sphere of radius a centered at the origin, For each of the indicated vector fields, find the flux *out of* the surface \mathcal{S} .
 - $\mathbf{F} = z^3\mathbf{k}$. Hint: $\mathbf{k} \cdot \mathbf{N} = \cos \phi$.
 - $\mathbf{F} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k} = \mathbf{r}$.
 - $\mathbf{F} = x\mathbf{i} + y\mathbf{j} = r \mathbf{u}_r$. Hint: $\mathbf{u}_r \cdot \mathbf{N} = \sin \phi$.
 - $\mathbf{F} = -y\mathbf{i} + x\mathbf{j}$. Hint: At each point of \mathcal{S} , the vector field \mathbf{F} is tangent to the circle of latitude through that point.
- Let \mathcal{S} be the closed cylinder of radius a , centered on the z -axis, with its base in the x, y -plane, and extending to the plane $z = h$. Both the top and bottom of the cylinder are considered part of \mathcal{S} in addition to the cylindrical lateral surface. Find the flux *out of* \mathcal{S} for each of the following vector fields. Note that the computations have to be done for each of the three components of \mathcal{S} .
 - $\mathbf{F} = z\mathbf{k}$.
 - $\mathbf{F} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k} = \mathbf{r}$.
 - $\mathbf{F} = x\mathbf{i} + y\mathbf{j} = r \mathbf{u}_r$. Hint: On the lateral surface, $\mathbf{N} = \mathbf{u}_r$.
 - $\mathbf{F} = -y\mathbf{i} + x\mathbf{j}$.
- Let $\mathbf{F}(\mathbf{r}) = \mathbf{r} = \langle x, y, z \rangle$, and let \mathcal{S} be the surface bounding the unit cube in the first octant. (\mathcal{S} consists of 6 squares, each of side 1.) Find the flux through \mathcal{S} using outward normals. Hint: The surface integral breaks up into 6 parts, three of which are zero. (Why?) The remaining 3 involving integrating a constant over a square.

6. Let $\mathbf{F}(\mathbf{r}) = z\mathbf{k}$ and let \mathcal{S} be the surface bounding the tetrahedron cut off in the first octant by the plane $x + 2y + z = 2$. Find the flux *out of* \mathcal{S} . (Note that \mathcal{S} breaks up into four pieces.)
7. Consider the cylindrical surface given parametrically by

$$\mathbf{r} = \langle a \cos \theta, a \sin \theta, s \rangle$$

where $0 \leq \theta < 2\pi$, $-b \leq s \leq b$. It has radius a and height $2b$. Note that $\mathbf{r}(\theta, s)$ approaches $\mathbf{r}(0, s)$ as $\theta \rightarrow 2\pi$. Hence, the entire surface is represented smoothly by the parametric representation; the “seam” at $\theta = 2\pi$ being illusory.

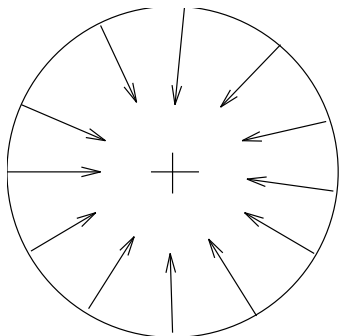
- (a) Find $\mathbf{n}(\theta, s) = \frac{\partial \mathbf{r}}{\partial \theta} \times \frac{\partial \mathbf{r}}{\partial s}$.
- (b) Show that $\mathbf{n}(\theta, 0) \rightarrow \mathbf{n}(0, 0)$ as $\theta \rightarrow 2\pi$.

8. Consider the Moebius surface given parametrically by

$$\mathbf{r} = \langle (a + s \cos(\theta/2)) \cos \theta, (a + s \cos(\theta/2)) \sin \theta, s \sin(\theta/2) \rangle$$

where $0 \leq \theta < 2\pi$, $-b \leq s \leq b$. (Assume $b < a$.) Note that $\mathbf{r}(\theta, s)$ approaches $\mathbf{r}(0, -s)$ as $\theta \rightarrow 2\pi$. Hence, the “seam” at $\theta = 2\pi$ is real, and the parametric representation fails to be continuous across it.

- (a) Find $\mathbf{n}(\theta, 0) = \frac{\partial \mathbf{r}}{\partial \theta}(\theta, 0) \times \frac{\partial \mathbf{r}}{\partial s}(\theta, 0)$.
- (b) Follow $\mathbf{n}(\theta, 0)$ as $0 \leq \theta < 2\pi$. In particular, note that $\mathbf{n}(\theta, 0) \rightarrow -\mathbf{n}(0, 0)$ as $\theta \rightarrow 2\pi$.



5.3 Conservative Vector Fields

Let \mathbf{F} be a vector field defined on some open set in \mathbf{R}^n (where as usual $n = 2$ or $n = 3$). For technical reasons, we also want to assume that the domain of the function is *connected*, i.e., it can't be decomposed into separate disjoint open sets. We say that \mathbf{F} is *conservative* if it is the gradient $\mathbf{F} = \nabla f$ of some scalar field f .

For example, in \mathbf{R}^3 , if $f(\mathbf{r}) = 1/|\mathbf{r}|$, then $\mathbf{F} = -(1/|\mathbf{r}|^2)\mathbf{u}_\rho$ is conservative. This *inverse square law field* arises in gravitation and also in electrostatics.

Note that the function f is not generally unique. For, if f is one such function, then for any constant c , we have

$$\nabla(f + c) = \nabla f + \nabla c = \nabla f = \mathbf{F}.$$

Hence, $f + c$ is another such function. The converse is also true. If f_1 and f_2 both work for the same conservative field \mathbf{F} , then

$$\mathbf{F} = \nabla f_1 = \nabla f_2 \Rightarrow \nabla f_1 = \nabla f_2 \Rightarrow \nabla(f_1 - f_2) = \mathbf{0}.$$

However, it is not hard to see that a (smooth) function with zero gradient must be constant. (Can you prove it? You need to use the fact that the domain is connected!) It follows that $f_1 - f_2 = c$, i.e., *any two functions with the same gradient differ by a constant*.

It is sometimes much easier to find the scalar function f and then take its gradient than it is to find the vector field \mathbf{F} directly.

Example 101, (gravitational dipole) Suppose two point masses of equal mass m are located at the points $(a, 0, 0)$ and $(-a, 0, 0)$. Let \mathbf{F}_1 denote the gravitational force due to the first mass and \mathbf{F}_2 the force due to the second mass. To find the combined force $\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2$ directly requires some complicated vector algebra. However, $\mathbf{F}_1 = \nabla f_1$ where

$$f_1(x, y, z) = \frac{Gm}{\sqrt{(x-a)^2 + y^2 + z^2}}.$$

Here, the expression in the denominator is the distance from (x, y, z) to the first mass. Similarly, the force due to the second mass is $\mathbf{F}_2 = \nabla f_2$ where

$$f_2(x, y, z) = \frac{Gm}{\sqrt{(x+a)^2 + y^2 + z^2}}.$$

Hence, the combined force is

$$\mathbf{F} = \nabla f_1 + \nabla f_2 = \nabla \underbrace{(f_1 + f_2)}_f.$$

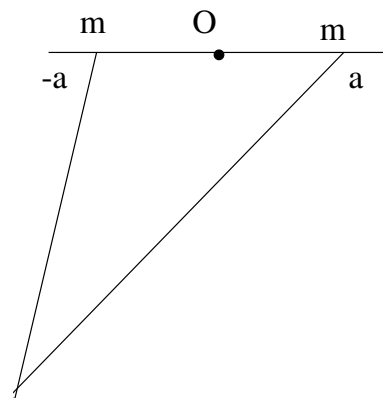
Calculating $f = f_1 + f_2$ is much simpler than calculating $\mathbf{F}_1 + \mathbf{F}_2$.

$$f(x, y, z) = Gm \left(\frac{1}{\sqrt{(x-a)^2 + y^2 + z^2}} + \frac{1}{\sqrt{(x+a)^2 + y^2 + z^2}} \right).$$

Of course, you still have to find the gradient of this function to find $\mathbf{F} = \nabla f$.

Line Integrals for Conservative Fields Conservative fields are particularly nice to deal with when computing line integrals. Suppose $\mathbf{F} = \nabla f$ is conservative, and \mathcal{C} is an oriented path in its domain going from point A to point B . (A and B might be the same point, in which case the path would be a closed loop.) We may calculate the line integral $\int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r}$ as follows. Choose a parametric representation $\mathbf{r} = \mathbf{r}(t)$, $a \leq t \leq b$ for \mathcal{C} . Then

$$\int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r} = \int_a^b \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) dt = \int_a^b \nabla f(\mathbf{r}(t)) \cdot \mathbf{r}'(t) dt.$$



However, by the *chain rule*

$$\nabla f(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = \frac{d}{dt} f(\mathbf{r}(t)).$$

Hence,

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_a^b \frac{d}{dt} f(\mathbf{r}(t)) dt = f(\mathbf{r}(t)) \Big|_a^b$$

so

$$\int_C \mathbf{F} \cdot d\mathbf{r} = f(\mathbf{r}(b)) - f(\mathbf{r}(a)) = f(B) - f(A). \quad (61)$$

In particular, the value of the line integral is *path independent* since it depends only on the endpoints of the path.

Example 102 Consider the inverse square field $\mathbf{F} = -(1/|\mathbf{r}|^2)\mathbf{u}_\rho$ which is the gradient of the function f defined by $f(\mathbf{r}) = 1/|\mathbf{r}|$. We have

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \frac{1}{|\mathbf{r}_B|} - \frac{1}{|\mathbf{r}_A|}.$$

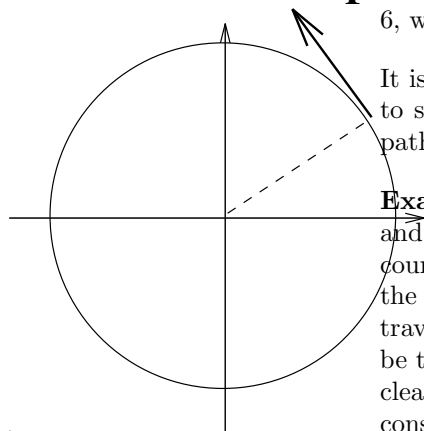
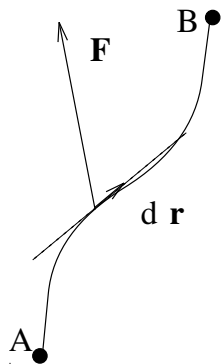
F Compare this with the calculation in the example at the end of Chapter I, Section 6, where the relevant integral in the plane case was calculated directly.

It is important to note that *not every vector field is conservative*. The easiest way to see this is to note that there are plenty of vector fields which do not have the path independence property for line integrals.

Example 103 Let $\mathbf{F}(x, y) = \langle -y, x \rangle = r\mathbf{u}_\theta$. Let C_1 be the closed path which starts and ends at the point $(1, 0)$ (so $A = B$) and traverses a circle of radius 1 in the counter-clockwise direction. This line integral has been calculated in exercises, and the answer is 2π . On the other hand, if we choose C_2 to be the same path but traversed clockwise, the answer will be -2π . Finally, to emphasize the point, let C_3 be the trivial path which just stays at the point $(1, 0)$. The line integral for C_3 will clearly be zero. The answer is certainly not path independent, so the field is not conservative.

The *path independence* property for a vector field in fact ensures that it is conservative. For suppose \mathbf{F} is defined on some connected open set in \mathbf{R}^n and that $\int_C \mathbf{F} \cdot d\mathbf{r}$ depends only on the endpoints of C for every oriented path C in the domain of \mathbf{F} . We can construct a function f for \mathbf{F} as follows. Choose a base point P_0 in the domain of \mathbf{F} . For any other point P with position vector \mathbf{r} in the domain of \mathbf{F} choose any path C in the domain from P_0 to P . Define

$$f(\mathbf{r}) = \int_C \mathbf{F} \cdot d\mathbf{r} = \int_{P_0}^P \mathbf{F} \cdot d\mathbf{r}$$



where the second form of the notation emphasizes that the result depends only on the endpoints of the path. I claim that $\nabla f = \mathbf{F}$. To see this, choose a parameterizing function $\mathbf{r}(u)$, $a \leq u \leq t$ for \mathcal{C} where $\mathbf{r}(a) = \mathbf{r}_0$ and $\mathbf{r}(t) = \mathbf{r}$. Then by the fundamental theorem of calculus

$$\frac{d}{dt}f(\mathbf{r}(t)) = \frac{d}{dt} \int_a^t \mathbf{F}(\mathbf{r}(u)) \cdot \mathbf{r}'(u) du = \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t).$$

On the other hand, by the chain rule

$$\frac{d}{dt}f(\mathbf{r}(t)) = \nabla f(\mathbf{r}(t)) \cdot \mathbf{r}'(t),$$

so

$$\nabla f(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t).$$

However, since the path \mathcal{C} is entirely arbitrary, its velocity vector $\mathbf{r}'(t)$ at its endpoint is also entirely arbitrary. That means that the dot products of ∇f and \mathbf{F} with every possible vector (including \mathbf{i} , \mathbf{j} , and \mathbf{k}) are the same, and it follows that $\nabla f = \mathbf{F}$, as claimed. This is an important enough fact to state formally. **Theorem 5.4** Let

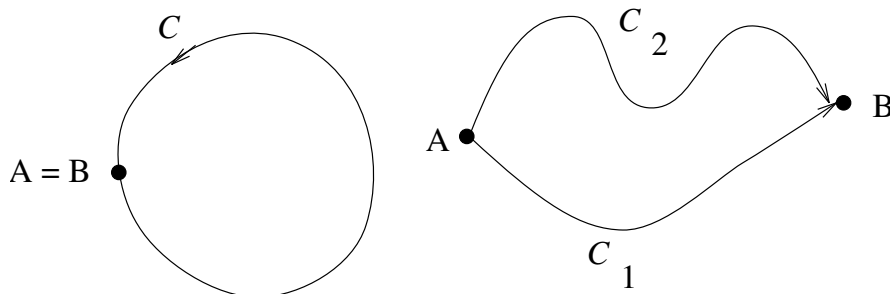
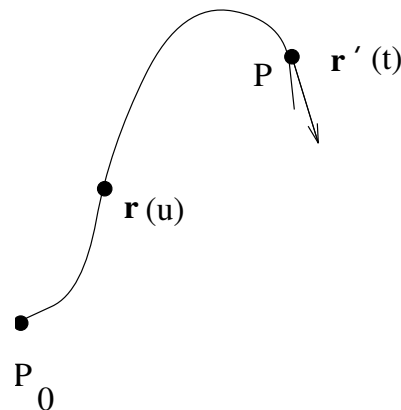
\mathbf{F} be a vector field defined on some connected open set in \mathbf{R}^n . \mathbf{F} is conservative if either of the following conditions holds: (i) $\mathbf{F} = \nabla f$ for some function f or (ii) \mathbf{F} satisfies the path independence condition for line integrals in its domain.

It should be noted that if \mathbf{F} is conservative, then *the line integral for \mathbf{F} around any closed loop is zero*. For, the result is independent of the path, and we may assume the path is the trivial path which never leaves the starting point. (See the diagram below.)

The converse is also true. If $\int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r} = 0$ for any closed loop \mathcal{C} in the domain of \mathbf{F} , then \mathbf{F} must be conservative. For, if \mathcal{C}_1 and \mathcal{C}_2 are two oriented paths which start and end at the same point, then we may consider the closed path \mathcal{C} formed by traversing first \mathcal{C}_1 , and then the opposite \mathcal{C}_2' of the second path. We have

$$0 = \int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r} = \int_{\mathcal{C}_1} \mathbf{F} \cdot d\mathbf{r} - \int_{\mathcal{C}_2} \mathbf{F} \cdot d\mathbf{r}$$

whence $\int_{\mathcal{C}_1} \mathbf{F} \cdot d\mathbf{r} = \int_{\mathcal{C}_2} \mathbf{F} \cdot d\mathbf{r}$. Thus, \mathbf{F} has the path independence property for line integrals.



Relation to the Work Energy Theorem in Physics For force fields, the notion of ‘conservative field’ is intimately tied up with the law of conservation of energy. To see why, suppose a particle moves under the influence of a conservative force $\mathbf{F} = \nabla f$. Then, by Newton’s Second Law, we have

$$m \frac{d^2 \mathbf{r}}{dt^2} = \nabla f.$$

Take the dot product of both sides with the velocity vector $\mathbf{v} = \frac{d\mathbf{r}}{dt}$. We get

$$m \frac{d\mathbf{v}}{dt} \cdot \mathbf{v} = \nabla f \cdot \mathbf{v}. \quad (62)$$

There is another way to obtain this equation. Let

$$E = \frac{1}{2} m (\mathbf{v} \cdot \mathbf{v}) - f.$$

Then

$$\begin{aligned} \frac{dE}{dt} &= \frac{1}{2} m (2 \frac{d\mathbf{v}}{dt} \cdot \mathbf{v}) - \frac{df}{dt} \\ &= m \frac{d\mathbf{v}}{dt} \cdot \mathbf{v} - \nabla f \cdot \mathbf{v}. \end{aligned}$$

Thus, equation (62) is equivalent to the assertion that $\frac{dE}{dt} = 0$, i.e., that E is constant.

This discussion should be familiar to you from physics. The quantity $T = \frac{m}{2} \mathbf{v} \cdot \mathbf{v} = \frac{m}{2} |\mathbf{v}|^2$ is called the *kinetic energy*, the quantity $V = -f$ is usually called the *potential energy*, and their *sum* E is called the *total energy*. Under the above assumptions, the total energy is conserved.

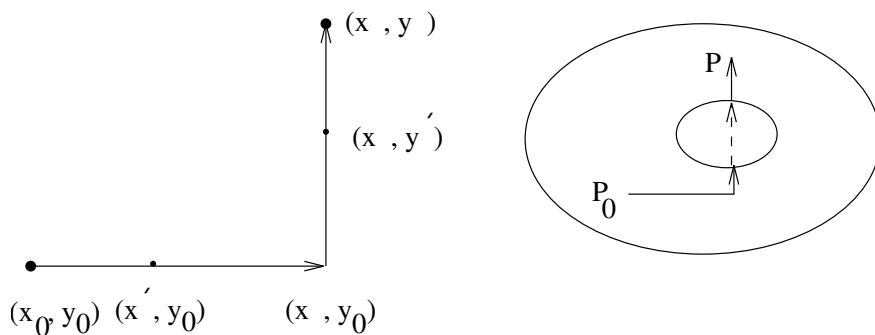
Because of the above considerations, a function $V = -f$ such that $\mathbf{F} = -\nabla V$ is often called a *potential function* for \mathbf{F} . (From a purely mathematical point of view, the minus sign is an unnecessary complication, so mathematicians sometimes leave it out when using the term ‘potential’. In this course, we shall follow the common usage in physics and keep the minus sign.)

Finding Potential Functions If we are given a conservative vector field \mathbf{F} , we want a method for finding a function f such that $\mathbf{F} = \nabla f$. ($V = -f$ will be the associated potential function.) We can of course use the line integral method described above. To this end, it is easiest to choose the path in a straightforward standard way. One common choice is to choose a polygonal path with segments parallel to the coordinate axes. For example, suppose $\mathbf{F} = \langle F_1, F_2 \rangle$ is a vector field in the plane. Fix a point (x_0, y_0) in its domain, and for any other point (x, y) in its domain, consider the path consisting of two segments, the first parallel to the

x -axis from (x_0, y_0) to (x, y_0) and the second parallel to the y -axis from (x, y_0) to (x, y) . It is easy to check that the line integral for this composite path is

$$f(x, y) = \int_{x_0}^x F_1(x', y_0) dx' + \int_{y_0}^y F_2(x, y') dy'. \quad (63)$$

(Note that we have to use *dummy variables* x', y' to avoid confusing the values of x and y at the endpoint from the values on the line segments.)



This formula may not always apply because the domain of the function can have *holes* in it which may block one or the other of the vertical line segments for some values of x or y .

Formula (63) can be a bit awkward to apply, but there is an equivalent method which uses indefinite integrals instead. We illustrate it by an example.

Example 104 Let $n = 2$ and let $\mathbf{F}(x, y) = \langle x^2 + 2xy, x^2 + 2y \rangle$. We shall find a function f as follows. If $\nabla f = \mathbf{F}$, then using components

$$\begin{aligned} \frac{\partial f}{\partial x} &= F_1(x, y) = x^2 + 2xy, \\ \frac{\partial f}{\partial y} &= F_2(x, y) = x^2 + 2y. \end{aligned}$$

Integrate the first equation with respect to x to obtain

$$f(x, y) = \frac{1}{3}x^3 + x^2y + C(y).$$

Note that the indefinite integral as usual involves an arbitrary constant, but since this was a ‘partial integration’ keeping y constant, this constant can in fact depend on y . Now differentiate with respect to y to obtain

$$\frac{\partial f}{\partial y} = x^2 + C'(y).$$

Comparing this with the previous expression for $\frac{\partial f}{\partial y}$ yields

$$x^2 + C'(y) = x^2 + 2y \quad \text{or} \quad C'(y) = 2y$$

from which we conclude $C(y) = y^2 + E$, where E is a constant independent of both x and y . Hence,

$$f(x, y) = \frac{1}{3}x^3 + x^2y + y^2 + E$$

is the most general function which will work. If we just want one such function, we can take $E = 0$, for example. Checking that this works yields

$$\begin{aligned} \frac{\partial f}{\partial x} &= x^2 + 2xy \\ \frac{\partial f}{\partial y} &= x^2 + 2y \end{aligned}$$

as required.

It is very important to check the answer since you could easily make a mistake in integrating. It might even be true that the field is not even conservative, but through an error, you have convinced yourself that you have found a function f with the right properties. There is a variation of the method which sometimes is a little faster.

As above, look for f such that

$$\frac{\partial f}{\partial x} = F_1(x, y) = x^2 + 2xy, \quad \frac{\partial f}{\partial y} = F_2(x, y) = x^2 + 2y.$$

Integrate the first equation with respect to x to obtain

$$f(x, y) = \frac{1}{3}x^3 + x^2y + C(y).$$

Next integrate the second equation with respect to y to obtain

$$f(x, y) = x^2y + y^2 + D(x).$$

If we compare these expressions, we see that we may choose $C(y) = y^2$ and $D(x) = \frac{1}{3}x^3$. Hence,

$$f(x, y) = \frac{1}{3}x^3 + x^2y + y^2$$

is the desired function.

Example 105 Let $n = 3$ and $\mathbf{F}(x, y, z) = \langle x + 2xy + z, x^2 + z, x + y \rangle$. We find a function f as follows. We want $\nabla f = \mathbf{F}$ or, in terms of components,

$$\frac{\partial f}{\partial x} = x + 2xy + z, \quad \frac{\partial f}{\partial y} = x^2 + z, \quad \frac{\partial f}{\partial z} = x + y.$$

Integrating the first equation with respect to x yields

$$f(x, y, z) = \frac{1}{2}x^2 + x^2y + zx + C(y, z)$$

where the constant of integration may depend on y and z . Differentiating with respect to y and comparing with the second equation yields

$$\frac{\partial f}{\partial y} = x^2 + \frac{\partial C}{\partial y} = x^2 + z.$$

Hence, $\frac{\partial C}{\partial y} = z$ from which we conclude

$$C(y, z) = zy + D(z) \quad \text{or} \quad f(x, y, z) = \frac{1}{2}x^2 + x^2y + zx + zy + D(z).$$

Differentiating with respect to z and comparing with the third equation yields

$$\frac{\partial f}{\partial z} = x + y + \frac{dD}{dz} = x + y$$

from which we conclude $D'(z) = 0$ or $D(z) = E$ where E is a constant independent of x, y , and z . Hence, the most general function f is given by

$$f(x, y, z) = \frac{1}{2}x^2 + x^2y + zx + zy + E.$$

For simplicity, we may take $E = 0$. You should check that the gradient of

$$f(x, y) = \frac{1}{2}x^2 + x^2y + zx + zy$$

is the original vector field \mathbf{F} .

Example 106 Let $n = 2$ and $\mathbf{F}(x, y) = \langle x + 1, y + x^2 \rangle$. A function f would have to satisfy

$$\frac{\partial f}{\partial x} = x + 1, \quad \frac{\partial f}{\partial y} = y + x^2.$$

Integrating the first equation with respect to x yields

$$f(x, y) = \frac{1}{2}x^2 + x + C(y).$$

Differentiating with respect to y and comparing with the second equation yields

$$\frac{\partial f}{\partial y} = C'(y) = y + x^2$$

which is impossible since $C(y)$ is not supposed to depend on x . What went wrong here is that the field is not conservative, i.e., *there is no such function f* , and the method for finding one breaks down.

Screening tests It would be nice to have a way to check in advance if a vector field \mathbf{F} is conservative before trying to find a function f with gradient \mathbf{F} . Fortunately, there are several ways to do this. First consider the case of a plane field $\mathbf{F} = \langle F_1, F_2 \rangle$. If $\nabla f = \mathbf{F}$ for some function f , then

$$\frac{\partial f}{\partial x} = F_1 \quad \text{and} \quad \frac{\partial f}{\partial y} = F_2.$$

If we differentiate the first equation with respect to y and the second with respect to x , we obtain

$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial F_1}{\partial y} \quad \text{and} \quad \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial F_2}{\partial x}.$$

If the function f is smooth enough (which we will normally want to be the case), the two mixed partials are equal, so we conclude that if the (smooth) field \mathbf{F} is conservative, then it must satisfy the *screening test*

$$\frac{\partial F_1}{\partial y} = \frac{\partial F_2}{\partial x}. \quad (64)$$

Example 106 again $F_1(x, y) = x + 1$ and $F_2(x, y) = y + x^2$ so

$$\frac{\partial F_1}{\partial y} = 0 \neq \frac{\partial F_2}{\partial x} = 2x$$

so \mathbf{F} does not pass the screening test and is not conservative.

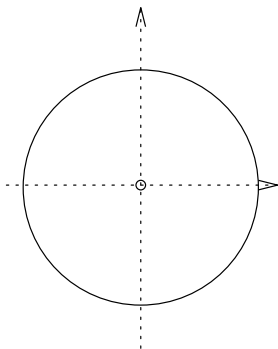
Unfortunately, it is possible for a vector field to pass the screening test but still not be conservative. In other words, the test is a bit too loose.

Example 107 Let $\mathbf{F}(x, y) = \langle \frac{-y}{x^2 + y^2}, \frac{x}{x^2 + y^2} \rangle$. \mathbf{F} may also be represented in polar form as $\mathbf{F} = \frac{1}{r} \mathbf{u}_\theta$. (Check that for yourself!) Notice that the domain of \mathbf{F} excludes the origin since the common denominator $x^2 + y^2$ vanishes there. We have

$$\begin{aligned} \frac{\partial F_1}{\partial y} &= \frac{(x^2 + y^2)(-1) - (-y)(2y)}{(x^2 + y^2)^2} = \frac{y^2 - x^2}{(x^2 + y^2)^2} \\ \frac{\partial F_2}{\partial x} &= \frac{(x^2 + y^2)(1) - x(2x)}{(x^2 + y^2)^2} = \frac{y^2 - x^2}{(x^2 + y^2)^2} \end{aligned}$$

so \mathbf{F} passes the screening test. However, \mathbf{F} is certainly not conservative. In fact, if C is a circle of any radius centered at the origin, then $\int_C \mathbf{F} \cdot d\mathbf{r} = \pm 2\pi$ with the sign depending on whether the circle is traversed counter-clockwise or clockwise. In any case, it is certainly not zero which would be the case for a conservative vector field. (The calculation of the integral is quite routine and in fact has been done previously in exercises. You should do it again now for practice.)

It would be much better if we could be sure that any vector field which passes the screening test is in fact conservative. We shall see later that this depends on



the nature of the *domain* of the vector field. In the present case, it is the ‘hole’ created by omitting the origin that is the cause of the difficulty. It is important to note, however, that the issue we raise here is not a minor point of interest only to mathematicians who insist of splitting hairs. The vector field in the example is in fact of great physical significance, and plays an important role in electromagnetic theory. There is a more complicated version of the screening test for vector fields in space. Let $\mathbf{F} = \langle F_1, F_2, F_3 \rangle$ be a vector field defined on some connected open set in \mathbf{R}^3 . If there is a function f such that $\nabla f = \mathbf{F}$, then writing this out in components yields

$$\frac{\partial f}{\partial x} = F_1, \quad \frac{\partial f}{\partial y} = F_2, \quad \text{and} \quad \frac{\partial f}{\partial z} = F_3.$$

By computing all the mixed partials (and assuming the order doesn’t matter), we obtain

$$\begin{aligned} \frac{\partial F_1}{\partial y} &= \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial F_2}{\partial x} \\ \frac{\partial F_1}{\partial z} &= \frac{\partial^2 f}{\partial x \partial z} = \frac{\partial F_3}{\partial x} \\ \frac{\partial F_2}{\partial z} &= \frac{\partial^2 f}{\partial y \partial z} = \frac{\partial F_3}{\partial y} \end{aligned}$$

so we get the more complicated screening test

$$\begin{aligned} \frac{\partial F_2}{\partial z} &= \frac{\partial F_3}{\partial y} \\ \frac{\partial F_1}{\partial z} &= \frac{\partial F_3}{\partial x} \\ \frac{\partial F_1}{\partial y} &= \frac{\partial F_2}{\partial x}. \end{aligned}$$

The reason why we list the equations in this particular order will become clear in the next section.

Example 108 Let $\mathbf{F}(x, y, z) = \langle y + z, x + z, x + y \rangle$. Find $\int_C \mathbf{F} \cdot d\mathbf{r}$ for the straight line path which goes from $(1, -1, 2)$ to $(0, 0, 3)$. We first check to see if the field might be conservative. We have

$$\begin{aligned} \frac{\partial F_2}{\partial z} &= 1 = \frac{\partial F_3}{\partial y} \\ \frac{\partial F_1}{\partial z} &= 1 = \frac{\partial F_3}{\partial x} \\ \frac{\partial F_1}{\partial y} &= 1 = \frac{\partial F_2}{\partial x} \end{aligned}$$

so it passes the screening test. We now try to find a function f . The equation $\partial f / \partial x = y + z$ yields $f(x, y, z) = yx + zx + C(y, z)$. Differentiate with respect to y to obtain $x + \partial C / \partial y = x + z$ or $\partial C / \partial y = z$. This yields $C(y, z) = zy + D(z)$.

Hence, $f(x, y, z) = yx + zx + zy + D(z)$. Differentiating with respect to z yields $x+y+D'(z) = x+y$ or $D'(z) = 0$. Hence, $D(z) = E$, and $f(x, y, z) = xy+xz+yz+E$ gives the most general function f . For convenience take $E = 0$. I leave it to you to check that $\nabla f = \mathbf{F}$ for $f(x, y, z) = xy + xz + yz$. To find the integral, we need only evaluate the function f at the endpoints of the path.

$$\int_{(1,-1,2)}^{(0,0,3)} \mathbf{F} \cdot d\mathbf{r} = f(0, 0, 3) - f(1, -1, 2) = 0 - (-4) = 4.$$

Exercises for 5.3.

- Find the force exerted by the gravitational dipole described in Example 101 by calculating ∇f for the function f derived in the example.
- Four point masses each of mass m are placed at the points $(\pm a, \pm a, 0)$ in the x, y -plane.
 - Show that the gravitational potential at a point $(0, 0, z)$ due to these four masses is given by $V(0, 0, z) = -\frac{4Gm}{\sqrt{2a^2 + z^2}}$.
 - Determine the gravitational force on a unit test mass at the point $(0, 0, z)$.
- Find the gravitational potential at a point on the z -axis due to a mass M uniformly distributed over a thin wire lying along the circle $x^2 + y^2 = a^2$ in the x, y -plane. Calculate the gravitational force on a unit test mass at a point on the z -axis.
- Suppose a scalar function defined on all of \mathbf{R}^n (with $n = 2$ or $n = 3$) satisfies $\nabla f = 0$ everywhere. Show that f is constant. (Hint: Use formula (61). Note that if the domain of f splits into two or more disconnected components, then the argument won't work because you can't find a path from a point in one component to a point in another component which stays in the domain of the function. Hence, the function might assume different (constant) values on the different components.)
- Find a function f such that $\mathbf{F} = \nabla f$ if one exists for each of the following plane vector fields.
 - $\mathbf{F}(x, y) = \langle 3x^2 + y^2, 2xy + y^2 \rangle$.
 - $\mathbf{F}(x, y) = \langle 3x^2 + y^2, 3xy + y^2 \rangle$.
 - $\mathbf{F}(x, y) = \langle e^y, xe^y \rangle$.
- Find a function f such that $\mathbf{F} = \nabla f$ if one exists for each of the following vector fields in space.
 - $\mathbf{F}(x, y, z) = \langle yz, xz, xy \rangle$.
 - $\mathbf{F}(x, y, z) = \langle x^2, y^2, z^2 \rangle$.
 - $\mathbf{F}(x, y, z) = \langle z, x, y \rangle$.

7. Evaluate each of the following path independent line integrals. (In each case, the vector field with the indicated components is conservative.)
- (a) $\int_{(1,2)}^{(3,5)} (x^2 + 4y^2)dx + (8xy)dy$.
- (b) $\int_{(1,-1,0)}^{(4,1,1)} x dx + y dy + z dz$.
8. Each of the following vector fields is not conservative. Verify that by showing that it does not satisfy the relevant screening test and also by finding a closed path for which the line integral is not zero.
- (a) $\mathbf{F}(x, y) = \langle -y, 2x \rangle$.
- (b) $\mathbf{F}(x, y, z) = \langle z, -x, e^z xy \rangle$.
9. We know that the vector field $\mathbf{F}(x, y) = \langle -y/(x^2 + y^2), x/(x^2 + y^2) \rangle, (x, y) \neq (0, 0)$ is not conservative. Ignore this fact and try to find a function f such that $\mathbf{F} = \nabla f$. You should come up with something like $f(x, y) = \tan^{-1}(y/x)$. Since the vector field is not conservative, no function exists. Explain the seeming contradiction. Hint: Consider the domains of \mathbf{F} and of f .
10. Suppose \mathbf{F} is a conservative vector field. Then we know that the formula

$$f(\mathbf{r}) = \int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{F} \cdot d\mathbf{r}$$

defines a function such that $\mathbf{F} = \nabla f$. We know that choosing a different base point \mathbf{r}'_0 results in another such function f' , and f' differs from f by a constant. How is the constant related to the base points \mathbf{r}_0 and \mathbf{r}'_0 ?

5.4 Divergence and Curl

In this section we deal only with vector fields in space.

The gradient of a scalar function

$$\nabla f = \left\langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right\rangle = \mathbf{i} \frac{\partial f}{\partial x} + \mathbf{j} \frac{\partial f}{\partial y} + \mathbf{k} \frac{\partial f}{\partial z}$$

may be viewed as the result of applying the vector operator

$$\nabla = \left\langle \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right\rangle = \mathbf{i} \frac{\partial}{\partial x} + \mathbf{j} \frac{\partial}{\partial y} + \mathbf{k} \frac{\partial}{\partial z}$$

to the scalar field f . It is natural ask then if we may apply this vector operator to a *vector* field \mathbf{F} . There are two ways to do this.

We may take the dot ‘product’

$$\begin{aligned}\nabla \cdot \mathbf{F} &= \left\langle \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right\rangle \cdot \langle F_1, F_2, F_3 \rangle \\ &= \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}.\end{aligned}$$

The result is a scalar field called the *divergence* of \mathbf{F} . It is also denoted $\operatorname{div} \mathbf{F}$.

Example 109 Let $\mathbf{F}(\mathbf{r}) = x^2\mathbf{i} + y^2\mathbf{j} + z^2\mathbf{k}$. Then

$$\nabla \cdot \mathbf{F} = 2x + 2y + 2z = 2(x + y + z).$$

Note that the result is always a scalar field.

We may also form the cross ‘product’

$$\begin{aligned}\nabla \times \mathbf{F} &= \left\langle \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right\rangle \times \langle F_1, F_2, F_3 \rangle \\ &= \left\langle \frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}, \frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}, \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right\rangle.\end{aligned}$$

The result is a vector field called the *curl* of \mathbf{F} . It is also denoted $\operatorname{curl} \mathbf{F}$.

Example 110 Let $\mathbf{F} = yz\mathbf{i} + xz\mathbf{j} = \langle yz, xz, 0 \rangle$. Then

$$\nabla \times \mathbf{F} = \langle 0 - x, -(0 - y), z - z \rangle = -x\mathbf{i} + y\mathbf{j}.$$

Note that the result is always a vector field.

If you refer back to the previous section, you will see that the three quantities which must vanish in the screening test for a conservative vector field in space are just the components of the curl. Hence, the screening test can be written more simply

$$\nabla \times \mathbf{F} = \mathbf{0}.$$

In particular, *every conservative vector field in space has zero curl*. Still another way to say the same thing is that

$$\nabla \times (\nabla f) = \mathbf{0}$$

for every scalar field f .

We shall investigate the significance of the divergence and curl in succeeding sections.

The formalism of the operator ∇ leads to some fascinating formulas and expressions, many of which have important applications. You will have an opportunity to explore some of these in the exercises. Probably the most important elaboration is the operator

$$\nabla^2 = \nabla \cdot \nabla = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

which is called the *Laplace operator*. If f is a scalar field, then $\nabla^2 f$ is called the Laplacian of f . The Laplace operator appears in many of the important partial differential equations of physics. Here are some

$$\begin{array}{ll} \nabla^2 f = 0 & \text{Laplace's Equation} \\ \nabla^2 f = \frac{1}{c^2} \frac{\partial^2 f}{\partial t^2} & \text{Wave Equation} \\ \nabla^2 f = \kappa \frac{\partial f}{\partial t} & \text{Diffusion or Heat Equation} \end{array}$$

Exercises for 5.4.

- Calculate the divergence and curl of each of the following vector fields in \mathbf{R}^3 .
 - $\mathbf{F}(x, y, z) = \langle x, y, z \rangle$.
 - $\mathbf{F}(x, y, z) = -y\mathbf{i} + x\mathbf{j}$.
 - $\mathbf{F}(x, y, z) = \frac{1}{\rho^3} \langle x, y, z \rangle$ where $\rho = \sqrt{x^2 + y^2 + z^2}$. Hint: The algebra will be easier if you use $\frac{\partial \rho}{\partial x} = \frac{x}{\rho}$ and the corresponding formulas for y and z .

- Prove that divergence and curl are linear, i.e., verify the formulas

$$\begin{aligned} \nabla \cdot (a\mathbf{F} + b\mathbf{G}) &= a\nabla \cdot \mathbf{F} + b\nabla \cdot \mathbf{G} \\ \nabla \times (a\mathbf{F} + b\mathbf{G}) &= a\nabla \times \mathbf{F} + b\nabla \times \mathbf{G} \end{aligned}$$

where \mathbf{F} and \mathbf{G} are vector fields and a and b are constant scalars.

- Verify the formula

$$\nabla \cdot (f\mathbf{F}) = f\nabla \cdot \mathbf{F} + \nabla f \cdot \mathbf{F}$$

where f is a scalar field and \mathbf{F} is a vector field.

What should the corresponding formula for curl be?

- Verify the formula

$$\nabla \cdot (\nabla \times \mathbf{F}) = 0$$

for an arbitrary vector field \mathbf{F} . This formula will play an important role later in this chapter.

- Verify the formula

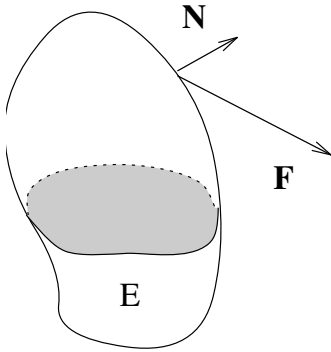
$$\nabla \times (\nabla \times \mathbf{F}) = \nabla(\nabla \cdot \mathbf{F}) - \nabla^2 \mathbf{F}$$

where $\nabla^2 \langle F_1, F_2, F_3 \rangle = \langle \nabla^2 F_1, \nabla^2 F_2, \nabla^2 F_3 \rangle$.

6. Show that the scalar field defined by $f(\mathbf{r}) = \frac{1}{\rho}$ satisfies Laplace's equation $\nabla^2 f = 0$ except at the origin where it is not defined. Hint: You can use a previous exercise in which you considered $-\nabla f$.
7. Let \mathbf{F} and \mathbf{G} be two vector fields. Guess a formula for $\nabla \cdot (\mathbf{F} \times \mathbf{G})$ and then verify that your formula is correct.

5.5 The Divergence Theorem

In the succeeding sections we shall discuss three important theorems about integrals of vector fields: the divergence theorem, Green's theorem, and Stokes's theorem. These theorems play specially important roles in the theory of electromagnetic fields, and we shall introduce the theorems in the order they are needed in that theory. We start with the divergence theorem, which is closely connected with the theories of static electric fields and gravitational fields which share many common mathematical features.



Let \mathbf{F} be a vector field defined on some open set in \mathbf{R}^3 . Moreover, suppose \mathbf{F} is *smooth* in the sense that the partial derivatives of the components of \mathbf{F} are all continuous. Let E be a solid region in the domain of \mathbf{F} which is bounded by a finite collection of graphs of smooth functions. Denote by ∂E the *surface* bounding E . *Orient* this boundary ∂E by specifying that the normals point *away* from the solid region E . Some simple examples of such regions would be rectangular boxes, solid spheres, solid hemispheres, solid cylinders, etc. However, we shall also allow more complicated regions such as a solid torus or the solid region between two spheres. In the latter case the boundary ∂E consists of two disconnected components.

Theorem 5.5 (Divergence Theorem) Let \mathbf{F} be a vector field in \mathbf{R}^3 and E a solid region as above. Then

$$\iint_{\partial E} \mathbf{F} \cdot d\mathbf{S} = \iiint_E \nabla \cdot \mathbf{F} dV.$$

Before trying to prove this theorem, we shall show how to use it and also investigate what it tells us about the physical meaning of the divergence $\nabla \cdot \mathbf{F}$.

Using the Divergence Theorem to Calculate Surface Integrals It is generally true that the triple integral on the right is easier to calculate than the surface integral on the left. That is the case first because volume integrals of scalar functions are easier to find than surface integrals and secondly because the divergence of \mathbf{F} is often simpler than \mathbf{F} .

Example 111 Consider $\iint_{\mathcal{S}} \mathbf{F} \cdot d\mathbf{S}$ where \mathcal{S} is a sphere of radius R centered at the origin, and $\mathbf{F}(x, y, z) = x\mathbf{i} + y\mathbf{j}$. If we let E be the solid sphere enclosed by \mathcal{S} , then

$\mathcal{S} = \partial E$. $\nabla \cdot \mathbf{F} = \partial x/\partial x + \partial y/\partial y = 2$. Hence,

$$\iint_{\mathcal{S}} \mathbf{F} \cdot d\mathbf{S} = \iiint_E 2 dV = 2 \frac{4}{3} \pi R^3 = \frac{8}{3} \pi R^3.$$

You were asked to do this surface integral previously by direct calculation in an exercise. You should go back and review the calculation. You will see that doing it by means of the divergence theorem is quite a bit easier.

Example 112 Let $\mathbf{F}(x, y, z) = z^2 \mathbf{k}$ and let \mathcal{S} be the hemispherical surface $x^2 + y^2 + z^2 = R^2, z \geq 0$. Use ‘upward’ normals. In this case the surface does not bound a solid region. However, that is easy enough to remedy. Let \mathcal{S}' be the disk of radius R in the x, y -plane. Then the solid hemisphere $0 \leq z \leq \sqrt{R^2 - x^2 - y^2}$ is bounded on the top by \mathcal{S} and on the bottom by \mathcal{S}' . We may write $\partial E = \mathcal{S} \cup \mathcal{S}'$ where we use ‘upward’ normals for \mathcal{S} and downward normals for \mathcal{S}' . Also, $\nabla \cdot \mathbf{F} = 2z$, and

$$\iiint_{\mathcal{S} \cup \mathcal{S}'} \mathbf{F} \cdot d\mathbf{S} = \iiint_E 2z dV.$$

The triple integral on the right is quite straightforward to do. For example, we could switch to spherical coordinates and use $z = \rho \cos \phi$ to obtain

$$\begin{aligned} \iiint_E 2z dV &= 2 \int_0^{2\pi} \int_0^{\pi/2} \int_0^R \rho^3 \cos \phi \sin \phi d\rho d\phi d\theta = 4\pi \frac{R^4}{4} \left(-\frac{\cos^2 \phi}{2} \Big|_0^{\pi/2} \right) \\ &= \frac{\pi}{2} R^4. \end{aligned}$$

The surface integral on the left may be split up as a sum

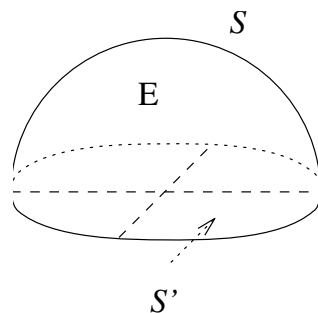
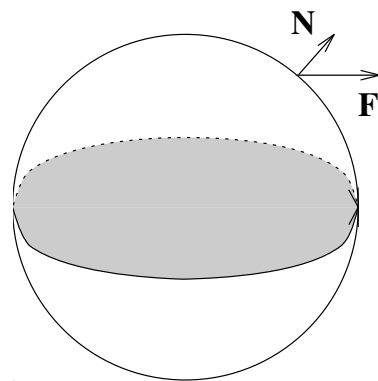
$$\iint_{\mathcal{S} \cup \mathcal{S}'} \mathbf{F} \cdot d\mathbf{S} = \iint_{\mathcal{S}} \mathbf{F} \cdot d\mathbf{S} + \iint_{\mathcal{S}'} \mathbf{F} \cdot d\mathbf{S}$$

where the first integral in the sum is the one we want to find. However, the second integral is quite easy to calculate. Namely, the field $\mathbf{F} = z\mathbf{k} = 0$ in the x, y -plane, so any surface integral over a flat surface in the x, y -plane will be zero for this field. Hence, the divergence theorem in this case comes down to

$$\iint_{\mathcal{S}} \mathbf{F} \cdot d\mathbf{S} + 0 = \frac{\pi}{2} R^4$$

$$\text{or } \iint_{\mathcal{S}} \mathbf{F} \cdot d\mathbf{S} = \frac{\pi}{2} R^4.$$

The above example illustrates a common way to apply the divergence theorem to find the flux through a surface which is not closed. Namely, one tries to add one or more additional components so as to bound a solid region. In this way one reduces the desired surface integral to a volume integral and other surface integrals which may be easier to calculate.

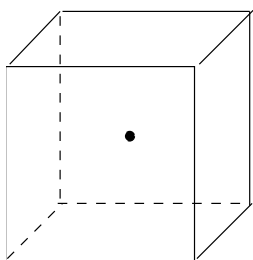


Interpretation of the divergence The divergence theorem gives us a way to interpret the divergence of a vector field in a more geometric manner. To this end consider a point P with position vector \mathbf{r} in the domain of a (smooth) vector field \mathbf{F} . Let E be a small element of volume containing the point P . To be definite, picture E as a small cube centered at P , but in principle it could be of any shape whatsoever. The quotient

$$\frac{1}{v(E)} \iint_{\partial E} \mathbf{F} \cdot d\mathbf{S}$$

(where $v(E)$ denotes the volume of the set E and the surface integral is computed using outward orientation) is called the *flux per unit volume*. I claim that the divergence of \mathbf{F} at the point P is given by

$$\nabla \cdot \mathbf{F}(\mathbf{r}) = \lim_{E \rightarrow P} \frac{1}{v(E)} \iint_{\partial E} \mathbf{F} \cdot d\mathbf{S}. \quad (65)$$



The importance of this formula is two-fold. First of all, it relates the divergence directly to the concept of flux. In effect, it allows us to think of the divergence of \mathbf{F} as being a measure of the *sources* of the field. For example, for a gravitational or electrostatic field, we can think of the lines of force emanating from mass or charge, and the divergence of the field is related to the mass or charge density. In the case of a momentum field for a fluid flow, the interpretation of divergence based on this formula is a bit more complicated, but similar considerations apply.

Secondly, the formula gives us a characterization of the divergence which does not depend on the use of a specific coordinate system. The same formula applies if we use different axes for rectangular coordinates or even if we use curvilinear coordinates such as cylindrical or spherical coordinates.

To derive formula (65), we argue as follows. By the average value property for triple integrals, we have

$$\nabla \cdot \mathbf{F}(\mathbf{r}') = \frac{1}{v(E)} \iiint_E \nabla \cdot \mathbf{F} \, dV$$

for an appropriate point \mathbf{r}' in E . (See Chapter IV, section 11 where the corresponding property for double integrals was discussed.) By the divergence theorem, the volume integral may be replaced by a surface integral, so

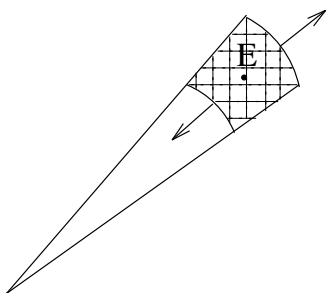
$$\nabla \cdot \mathbf{F}(\mathbf{r}') = \frac{1}{v(E)} \iint_{\partial E} \mathbf{F} \cdot d\mathbf{S}.$$

Now take the limit as $E \rightarrow P$. Since \mathbf{r} is the coordinate vector of P , it is clear that $\mathbf{r}' \rightarrow \mathbf{r}$. Hence,

$$\nabla \cdot \mathbf{F}(\mathbf{r}) = \lim_{E \rightarrow P} \nabla \cdot \mathbf{F}(\mathbf{r}') = \lim_{E \rightarrow P} \frac{1}{v(E)} \iint_{\partial E} \mathbf{F} \cdot d\mathbf{S}$$

as required.

As an application of formula (1), we show that the inverse square law vector field $\mathbf{F} = (1/|\mathbf{r}|^2)\mathbf{u}_\rho$ has zero divergence. (You should have done this directly in an



exercise by doing the messy calculation in rectangular coordinates.) To see this, consider an element of volume E centered at a point P with position vector $\mathbf{r} \neq \mathbf{0}$. Assume in particular that E is a curvilinear spherical cell, i.e., a typical element of volume for spherical coordinates. (See Chapter IV, Section 6.) E is bounded on the inside and outside by spherical surfaces of radii ρ_1 and ρ_2 respectively. It is bounded on either side by planes ($\theta = \theta_1, \theta = \theta_2$) and on ‘top’ and ‘bottom’ by conical surfaces ($\phi = \phi_1, \phi = \phi_2$), and each of those four bounding surfaces is *parallel to the field* \mathbf{F} . Hence, the flux through the boundary ∂E of E is zero except possibly for the inner and outer surfaces, each of which is a spherical rectangle. Let these two rectangles have areas A_1 and A_2 respectively. Since \mathbf{F} is parallel to the normal \mathbf{N} on each of these surfaces, and is otherwise constant, we get for the flux through the outer surface

$$\frac{1}{\rho_2^2} A_2$$

and similarly for the inner surface with 1 replacing 2 and with the sign reversed since the direction of the normal is reversed. Hence, the net flux is

$$\frac{1}{\rho_2^2} A_2 - \frac{1}{\rho_1^2} A_1$$

However, both curvilinear rectangles, can be described by the same angular limits $\phi_1 \leq \phi \leq \phi_2, \theta_1 \leq \theta \leq \theta_2$ —only the value of ρ changes. Hence,

$$A_1 = \int_{\theta_1}^{\theta_2} \int_{\phi_1}^{\phi_2} \rho_1^2 \sin \phi \, d\phi \, d\theta = \rho_1^2 \int_{\theta_1}^{\theta_2} \int_{\phi_1}^{\phi_2} \sin \phi \, d\phi \, d\theta$$

or

$$\frac{A_1}{\rho_1^2} = \int_{\theta_1}^{\theta_2} \int_{\phi_1}^{\phi_2} \sin \phi \, d\phi \, d\theta.$$

Since the same calculation would work for A_2/ρ_2^2 and give exactly the same value, we conclude that the net flux through the boundary ∂E is zero. If we take the limit as $E \rightarrow P$, we still get zero, so the divergence is zero.

Note that the upshot of this argument is that since the inverse square law has its ‘source’ at the origin, streamlines (lines of force) entering any element of volume not including the origin all leave and no new ones originate, so the net flux is zero.

We note in passing that if A is the area of any region on a sphere of radius ρ , the quantity A/ρ^2 is called the *solid angle* subtended by the region. This quantity is independent of the radius ρ of the sphere in the sense that if we consider a different (concentric) sphere and the projected region on the other sphere, the ratio of area to ρ^2 does not change.

Exercises for 5.5.

1. Let $\mathbf{F}(\mathbf{r}) = \mathbf{r}$. (Then $\nabla \cdot \mathbf{F} = 3$.) Verify that the divergence theorem is true for each of the following solid regions E by calculating the outward flux

$\iint_{\partial E} \mathbf{F} \cdot \mathbf{N} dS$ through the boundary of E and checking that it is three times the volume.

- (a) E is a solid sphere of radius a centered at the origin.
 - (b) E is a cube of side a in the first octant with opposite vertices at $(0, 0, 0)$ and (a, a, a) .
2. Let $\mathbf{F} = x^2\mathbf{i} + y^2\mathbf{j} + z^2\mathbf{k}$. Use the divergence theorem to calculate the flux out of the following solid regions.
 - (a) The solid cube of side 1 in the first octant with opposite vertices at $(0, 0, 0)$ and $(1, 1, 1)$.
 - (b) The inside of the cylinder $x^2 + y^2 = a^2$, $0 \leq z \leq h$.
 3. Find $\iint_{\mathcal{S}} \mathbf{F} \cdot d\mathbf{S}$ where $\mathbf{F}(x, y, z) = x^3\mathbf{i} + y^3\mathbf{j} + z^3\mathbf{k}$ and \mathcal{S} is a sphere of radius a centered at the origin. Use outward normals.
 4. Find $\iint_{\mathcal{S}} \mathbf{F} \cdot d\mathbf{S}$ where $\mathbf{F}(x, y, z) = \langle x + e^{y^2}, \cos(xz), \sin(x^3 + y^3) \rangle$ and \mathcal{S} is the boundary of the solid region bounded below by $z = x^2 + y^2$ and bounded above by the plane $z = 4$. Use outward normals.
 5. Let f and g be scalar fields which are both sufficiently smooth on a solid region E and its boundary. Apply the divergence theorem to $f\nabla g$ to obtain *Green's first identity*

$$\iint_{\partial E} f \nabla g \cdot \mathbf{N} dS = \iiint_E (\nabla f \cdot \nabla g + f \nabla^2 g) dV.$$

Note that $\nabla g \cdot \mathbf{N}$ is just the directional derivative of g in the direction of the normal vector \mathbf{N} . This is sometimes called the normal derivative of g on the surface $\mathcal{S} = \partial E$.

Reverse the roles of f and g and subtract to obtain *Green's second identity*

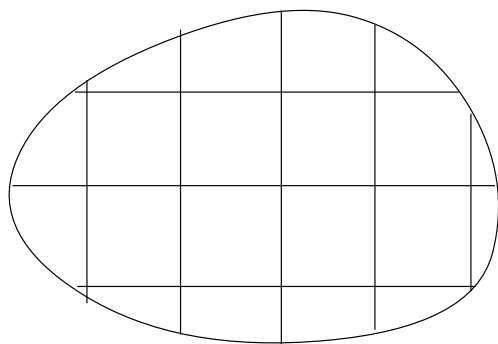
$$\iint_{\partial E} (f \nabla g \cdot \mathbf{N} - g \nabla f \cdot \mathbf{N}) dS = \iiint_E (f \nabla^2 g - g \nabla^2 f) dV.$$

6. Let $\mathbf{F} = \frac{1}{|\mathbf{r}|^2} \mathbf{u}_\rho$, and let \mathcal{S} be a disk of radius $\frac{\sqrt{3}a}{2}$ in the plane $z = \frac{a}{2}$ with center on the z -axis. Orient \mathcal{S} with upward normals. Use the divergence theorem to calculate $\iint_{\mathcal{S}} \mathbf{F} \cdot \mathbf{n} dS$. Hint: Form a closed surface with a spherical cap \mathcal{S}' of radius a . Calculate the surface integral for the cap and use the fact that $\nabla \cdot \mathbf{F} = 0$ between the spherical cap and the disk.
7. Suppose that it is known that the flux of the vector field \mathbf{F} out of every sufficiently small cube is zero. Show that $\nabla \cdot \mathbf{F} = 0$.

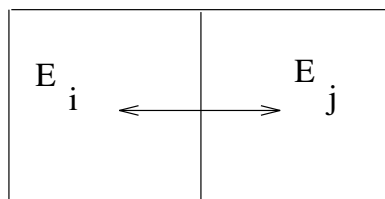
5.6 Proof of the Divergence Theorem

In this section, we shall give two proofs of the divergence theorem. The first is not really a proof but what is sometimes called a ‘plausibility argument’. It clarifies some of the ideas and helps you understand why the theorem might be true. The second proof is less enlightening but is mathematically correct. Even if you are not interested in the rigorous details, you should study the first part of this section because it introduces arguments commonly used in applications.

The first argument goes as follows. Since the right hand side of the formula in the theorem is a triple integral, consider a dissection of the solid region E into small cells formed as usual by three mutually perpendicular families of planes.



Dissection



Adjacent internal cells

Cross sectional views

Most of these cells will be small rectilinear boxes, but some, those on the boundary, will have one curved side. Let E_i be a typical cell. Using the interpretation of divergence as the *limit* of flux per unit volume, we have for E_i small enough

$$\frac{1}{\Delta V_i} \iint_{\partial E_i} \mathbf{F} \cdot d\mathbf{S} \approx \nabla \cdot \mathbf{F}(\mathbf{r}_i) \quad (66)$$

where ΔV_i is the volume of E_i and \mathbf{r}_i is the coordinate vector of a point in E_i . If we cross multiply and add, we get

$$\sum_i \iint_{\partial E_i} \mathbf{F} \cdot d\mathbf{S} \approx \sum_i \nabla \cdot \mathbf{F}(\mathbf{r}_i) \Delta V_i.$$

If we make the dissections finer and finer and take the limit, the sum on the right approaches $\iiint_E \nabla \cdot \mathbf{F} dV$. The sum on the left merits further discussion. Suppose that the numbering is such that cells E_i and E_j are adjacent, so they share a

common face. On that face, the normal relative to E_i will point opposite to the normal relative to E_j . (Each normal points away from one cell and into the other.)

As a result, the common face will appear in both terms of the sum

$$\iint_{\partial E_i} \mathbf{F} \cdot d\mathbf{S} + \iint_{\partial E_j} \mathbf{F} \cdot d\mathbf{S}$$

but with *opposite signs*. As a result, all *internal* components of the boundaries of the cells appear *twice* in the sum so as to cancel, and the only components left are those associated with the external boundary ∂E . In other words,

$$\sum_i \iint_{\partial E_i} \mathbf{F} \cdot d\mathbf{S} = \iint_{\partial E} \mathbf{F} \cdot d\mathbf{S}.$$

Putting this all together, we get

$$\iint_{\partial E} \mathbf{F} \cdot d\mathbf{S} = \iiint_E \nabla \cdot \mathbf{F} \, dV$$

as required.

The above argument is not logically valid for the following reason. The approximation (66) comes from the formula

$$\nabla \cdot \mathbf{F}(\mathbf{r}) = \lim_{E \rightarrow P} \frac{1}{v(E)} \iint_{\partial E} \mathbf{F} \cdot d\mathbf{S} \quad (67)$$

which was derived from the divergence theorem. Clearly we can't use a consequence of the divergence theorem to prove the divergence theorem or we will get in a vicious logical circle. To repair this argument, one would have to derive formula (67) *without using the divergence theorem*. This is not too hard to do if one assumes the solid E always has some specific form such as a rectangular box. However, since some of the cells in a dissection will have curved faces, that does not suffice. It is necessary to derive formula (67) for very general regions E . To the best of my knowledge, there is no simple way to do that without in essence hiding a proof of the divergence theorem in the argument.

A correct proof of the divergence theorem proceeds as follows. First write

$$\mathbf{F} = F_1 \mathbf{i} + F_2 \mathbf{j} + F_3 \mathbf{k}.$$

Suppose we can verify the three formulas

$$\begin{aligned} \iint_{\partial E} F_1 \mathbf{i} \cdot d\mathbf{S} &= \iiint_E \frac{\partial F_1}{\partial x} \, dV \\ \iint_{\partial E} F_2 \mathbf{j} \cdot d\mathbf{S} &= \iiint_E \frac{\partial F_2}{\partial y} \, dV \\ \iint_{\partial E} F_3 \mathbf{k} \cdot d\mathbf{S} &= \iiint_E \frac{\partial F_3}{\partial z} \, dV. \end{aligned}$$

Then adding these up will give the divergence theorem. Clearly, we need only verify one of the three formulas since the arguments will be basically the same in the three cases. We shall verify the third formula. In essence, that means we restrict to the case $\mathbf{F} = F_3 \mathbf{k}$, $\nabla \cdot \mathbf{F} = \partial F_3 / \partial z$. Consider a dissection of the solid region E such that each cell E_i is either a rectangular box or at worst bounded on either the top or the bottom by the graph of a smooth function. (It is believable that any region one is likely to encounter can be so decomposed, but in any case we can prove the theorem only for such regions.) As above, cancellation on internal interfaces yields

$$\sum_i \iint_{\partial E_i} \mathbf{F} \cdot d\mathbf{S} = \iint_{\partial E} \mathbf{F} \cdot d\mathbf{S}.$$

Similarly,

$$\sum_i \iiint_{E_i} \nabla \cdot \mathbf{F} dV = \iiint_E \nabla \cdot \mathbf{F} dV.$$

Hence, it will suffice to prove

$$\iint_{\partial E_i} \mathbf{F} \cdot d\mathbf{S} = \iiint_{E_i} \nabla \cdot \mathbf{F} dV$$

for each of the cells E_i . However, by assumption each of these cells is bounded on top and bottom by graphs of smooth functions $z = f(x, y)$ and $z = g(x, y)$. For the cells which are just boxes, the two functions are constant functions, for a cell on the top boundary, the bottom function is constant, and for a cell on the bottom boundary, the top function is constant. In any case, this reduces the problem to verifying the formula of the divergence theorem for a region which can be described by $g(x, y) \leq z \leq f(x, y)$ where (x, y) ranges over some domain D in the x, y -plane. (The domain D will in most cases be just a rectangle.) We now compute the surface integral for such a region under the assumption, as above, that $\mathbf{F} = F_3 \mathbf{k}$. Since the field is vertical, it is parallel to the sides of the region, so the only contributions to the flux come from the top and bottom surfaces. On the top surface, we have

$$d\mathbf{S} = \langle -f_x, -f_y, 1 \rangle dy dx$$

so the flux is

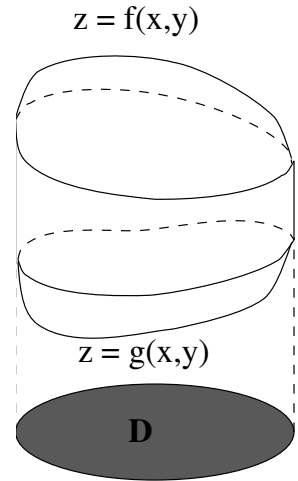
$$\iint_D F_3(x, y, f(x, y)) dy dx.$$

Similarly, on the bottom surface, we have

$$d\mathbf{S} = \langle g_x, g_y, -1 \rangle dy dx$$

where the signs are reversed because we need the downward pointing normals. Hence, the flux is

$$- \iint_D F_3(x, y, g(x, y)) dy dx,$$



and the net flux through the boundary of the region is

$$\begin{aligned} \iint_D F_3(x, y, f(x, y)) dy dx - \iint_D F_3(x, y, g(x, y)) dy dx = \\ \iint_D (F_3(x, y, f(x, y)) - F_3(x, y, g(x, y))) dy dx. \end{aligned}$$

Next we calculate the volume integral.

$$\begin{aligned} \iiint_E \frac{\partial F_3}{\partial z} dV &= \iint_D \int_{g(x, y)}^{f(x, y)} \frac{\partial F_3}{\partial z} dz dy dx \\ &= \iint_D F_3(x, y, z) \Big|_{g(x, y)}^{f(x, y)} dy dx \\ &= \iint_D (F_3(x, y, f(x, y)) - F_3(x, y, g(x, y))) dy dx. \end{aligned}$$

Comparing, we see that the answers are the same which proves that the surface integral equals the volume integral. That completes the proof of the divergence theorem.

Note that the second proof finally comes down to an application of basic integration theory and ends up using the fundamental theorem of calculus. The divergence theorem may be viewed just as a higher dimensional extension of the fundamental theorem.

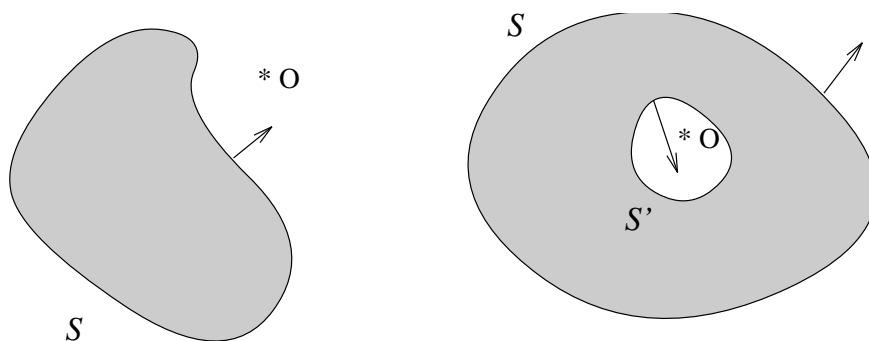
Exercises for 5.6.

1. Let \mathbf{F} be a vector field. Suppose that it is known that the flux out of every sufficiently small rectangular box is zero. Using the reasoning in the proof of the divergence theorem, show that the flux out of any rectangular box of any size whatsoever is zero. Don't use the divergence theorem itself, only the reasoning in the proof.

5.7 Gauss's Law and the Dirac Delta Function

We return to our discussion of inverse square laws. In particular, let $\mathbf{F} = (1/|\mathbf{r}|^2)\mathbf{u}_\rho$ and let \mathcal{S} be any surface bounding a solid region E in \mathbf{R}^3 . Since \mathbf{F} has a singularity at the origin, assume the surface does not pass through the origin. It is easy to calculate the flux through \mathcal{S} if the origin is not in E . For, since $\nabla \cdot \mathbf{F} = 0$ everywhere in the region E , the divergence theorem tells us

$$\iint_{\mathcal{S}} \mathbf{F} \cdot d\mathbf{S} = \iiint_E (0) dV = 0.$$



The situation is somewhat more complicated if the origin is in the solid region bounded by \mathcal{S} . Note that in this case the divergence theorem does not apply because the smoothness hypothesis on \mathbf{F} fails at a point in E . However, we can calculate the flux by the following trick. Consider a small sphere \mathcal{S}' centered at the origin but otherwise entirely inside the surface \mathcal{S} . Let E' be the solid obtained by removing the interior of the small sphere \mathcal{S}' from E . Thus, E' is the solid region *between* \mathcal{S}' and \mathcal{S} . The boundary of E' consists of the two surfaces \mathcal{S} and \mathcal{S}' . Using ‘outward’ normals (i.e., away from E'), the normals point outward on \mathcal{S} and *inward* on \mathcal{S}' . Since the field is smooth in E' , we may apply the divergence theorem to obtain

$$\iint_{\mathcal{S} \cup \mathcal{S}'} \mathbf{F} \cdot d\mathbf{S} = \iiint_{E'} \nabla \cdot \mathbf{F} \, dV = 0.$$

On the other hand,

$$\iint_{\mathcal{S} \cup \mathcal{S}'} \mathbf{F} \cdot d\mathbf{S} = \iint_{\mathcal{S}} \mathbf{F} \cdot d\mathbf{S} + \iint_{\mathcal{S}'} \mathbf{F} \cdot d\mathbf{S},$$

so

$$\iint_{\mathcal{S}} \mathbf{F} \cdot d\mathbf{S} = - \iint_{\mathcal{S}'} \mathbf{F} \cdot d\mathbf{S}.$$

This reduces the calculation of the flux through the surface \mathcal{S} to the calculation of the flux through a sphere \mathcal{S}' centered at origin. The latter calculation was basically done in Example 1 of Section 2 of this chapter. You should do it over again for practice. The answer turns out to be 4π if the sphere is given outward orientation. In the present case the orientation is reversed which changes the sign, but the above formula gives one last change of sign, so we conclude for the original surface \mathcal{S} ,

$$\iint_{\mathcal{S}} \mathbf{F} \cdot d\mathbf{S} = 4\pi.$$

To summarize, for the field $\mathbf{F} = (1/|\mathbf{r}|^2)\mathbf{u}_\rho$,

$$\iint_{\mathcal{S}} \mathbf{F} \cdot d\mathbf{S} = \begin{cases} 0 & \text{if } \mathbf{0} \text{ is outside } \mathcal{S} \\ 4\pi & \text{if } \mathbf{0} \text{ is inside } \mathcal{S} \end{cases}.$$

Gauss's Law The electric field of a point charge q located at the point \mathbf{r}_0 is given by Coulomb's law

$$\mathbf{E} = \frac{q}{|\mathbf{r} - \mathbf{r}_0|^2} \mathbf{u} \quad (68)$$

where \mathbf{u} is a unit vector pointing directly away from the point source. (We have dropped some important physical constants for the sake of mathematical simplicity.) Coulomb's law is analogous to Newton's law of gravitation for a point mass.

The calculation of flux discussed above applies just as well to such a field. The flux out of any closed surface is either 0 if the source is outside the surface and it is $4\pi q$ if the source is inside the surface. (The only change is a shift of the source from the origin to the point \mathbf{r}_0 and multiplication by the magnitude of the charge.) More generally, suppose we have many point sources with charges q_1, q_2, \dots, q_n located at position vectors $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$. Let \mathbf{E}_i be the electric field of the i th charge, and let

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 + \dots + \mathbf{E}_n$$

be the resulting field from all the charges. If \mathcal{S} is any closed surface, we will have

$$\iint_{\mathcal{S}} \mathbf{E} \cdot d\mathbf{S} = \iint_{\mathcal{S}} \mathbf{E}_1 \cdot d\mathbf{S} + \iint_{\mathcal{S}} \mathbf{E}_2 \cdot d\mathbf{S} + \dots + \iint_{\mathcal{S}} \mathbf{E}_n \cdot d\mathbf{S}.$$

The i th integral on the right will be either 0 or $4\pi q_i$ depending on whether the i th source is outside or inside \mathcal{S} . Hence, we get

$$\iint_{\mathcal{S}} \mathbf{E} \cdot d\mathbf{S} = 4\pi \sum_{q_i \text{ inside } \mathcal{S}} q_i = 4\pi Q$$

where Q is the sum of the charges inside the surface. This is a special case of Gauss's Law in electrostatics which asserts that for any electrostatic field, the flux out of a closed surface is $4\pi Q$ where Q is the total charge contained within.

(The constant 4π in Gauss's law depends on the form we adopted for Coulomb's law, which is something of an oversimplification. It is convenient for a purely mathematical discussion, but in reality Coulomb's law involves some additional constants which depend on the system of units employed. For example, for one commonly used system of units, Coulomb's law becomes

$$\mathbf{E} = \frac{1}{4\pi\epsilon} \frac{q}{|\mathbf{r} - \mathbf{r}_0|^2} \mathbf{u}$$

where ϵ is the so-called permittivity. In that case, the factor 4π disappears, and Gauss's law becomes

$$\iint_{\mathcal{S}} \mathbf{E} \cdot d\mathbf{S} = \frac{Q}{\epsilon}.$$

We leave further discussion of such matters to your physics professor.)

Gauss's law is very important in electrostatics, and is closely related to the divergence theorem, but you should avoid confusing the two.

The Dirac Delta Function There is a way to recover the divergence theorem for solid regions in which the vector field has singularities. It involves use of the Dirac Delta function which was discussed in Chapter IV, Section 11. For example, for the inverse square law, we can write formally

$$\nabla \cdot \left(\frac{1}{|\mathbf{r}|^2} \mathbf{u}_\rho \right) = 4\pi \delta(\mathbf{r})$$

where as previously $\delta(\mathbf{r})$ is 'defined' to be 0 except at $\mathbf{r} = \mathbf{0}$, and it is required to satisfy

$$\iiint_E \delta(\mathbf{r}) dV = \begin{cases} 0 & \text{if } \mathbf{0} \text{ is not in } E \\ 1 & \text{if } \mathbf{0} \text{ is in } E \end{cases}.$$

With this interpretation, the divergence theorem

$$\iint_{\partial E} \mathbf{F} \cdot d\mathbf{S} = \iiint_E \nabla \cdot \mathbf{F} dV$$

is true, since both sides are either 0 or 4π depending on whether the origin is outside or inside E . Note however that if a normal function vanishes everywhere except at one point, its triple integral is zero. Thus, it is important to remember that the Dirac Delta 'function' is not a function in the usual sense.

Exercises for 5.7.

- Let $\mathbf{F} = \frac{1}{|\mathbf{r}|^2} \mathbf{u}_\rho$. For each of the following surfaces determine the outward flux.
 - The ellipsoid $x^2/4 + y^2/16 + z^2 = 1$.
 - The surface of the cube with vertices at $(\pm 2, \pm 2, \pm 2)$.
 - The sphere $x^2 + y^2 + (z - 3)^2 = 1$.
 - (Optional) The surface of the unit cube in the first octant with opposite vertices at $(0, 0, 0)$ and $(1, 1, 1)$. Hint: \mathbf{F} blows up at $(0, 0, 0)$ so you can't apply the divergence theorem without modification. This does not create problems for the surface integrals on the three faces in the coordinate planes because these are zero anyway. (Why?) Let a be small, and consider the part of the cube outside the sphere of radius a centered at the origin. Apply the divergence theorem to that solid region and let a approach 0. The answer is $\pi/2$.
- Using the divergence theorem and Gauss's law, show that $\nabla \cdot \mathbf{E} = 0$ for an electrostatic field \mathbf{E} in empty space (where there is no charge). Hint: By Gauss's Law, the flux out of any sufficiently small sphere centered at a point where there is no charge is zero.

3. Electrostatic fields are always conservative. If $\mathbf{E} = \nabla f$ for a function f , show that f satisfies Laplace's equation $\nabla^2 f = 0$ if the charge density is zero.
4. Suppose that \mathbf{E} is a smooth vector field in \mathbf{R}^3 which is spherically symmetric about the origin and it satisfies $\nabla \cdot \mathbf{E} = \rho$. Find \mathbf{E} .
5. Suppose it is known that \mathbf{E} is a smooth vector field which is cylindrically symmetric about the \mathbf{z} -axis and it satisfies $\nabla \cdot \mathbf{E} = \frac{\sin r}{r}$. Find \mathbf{E} .

5.8 Green's Theorem

We consider the analogue of the divergence theorem in \mathbf{R}^2 .

Let $\mathbf{F} = \langle F_1, F_2 \rangle = F_1 \mathbf{i} + F_2 \mathbf{j}$ be a (smooth) vector field defined on some open set in the plane. Let D be a plane region bounded by a curve $\mathcal{C} = \partial D$. (D is analogous to E and \mathcal{C} to $\mathcal{S} = \partial E$.) The analogue of the flux is the integral over the curve \mathcal{C}

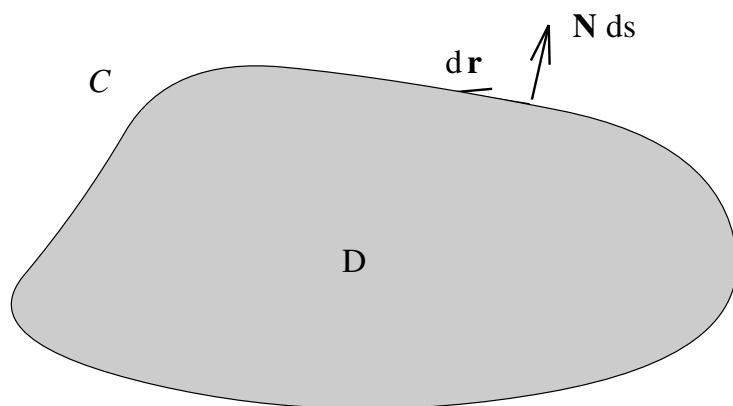
$$\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} \, ds$$

where \mathbf{N} denotes the *outward* unit normal to \mathcal{C} at each point of \mathcal{C} and ds stands for the element of arc length. This flux integral can also be expressed in rectangular coordinates as follows. If we write $d\mathbf{r} = dx \mathbf{i} + dy \mathbf{j}$, then the vector $dy \mathbf{i} - dx \mathbf{j}$ is perpendicular to $d\mathbf{r}$. Moreover, if we assume that \mathcal{C} is traversed in such a way that the region D is *always on the left*, then this vector does point away from D . (See the diagram.) Since, its magnitude is $\sqrt{(dy)^2 + (-dx)^2} = ds$, we have

$$\mathbf{N} \, ds = dy \mathbf{i} - dx \mathbf{j}$$

and

$$\int_{\mathcal{C}} \mathbf{F} \cdot \mathbf{N} \, ds = \int_{\mathcal{C}} -F_2 dx + F_1 dy.$$



Example 113 Let $\mathbf{F}(x, y) = \langle x, y \rangle = \mathbf{r}$, and let D be a disk of radius R . If the boundary $C = \partial D$ is traversed in the counter-clockwise direction, the region will always be on the left. We can see geometrically, that

$$\mathbf{F} \cdot \mathbf{N} ds = R R d\theta = R^2 d\theta$$

so

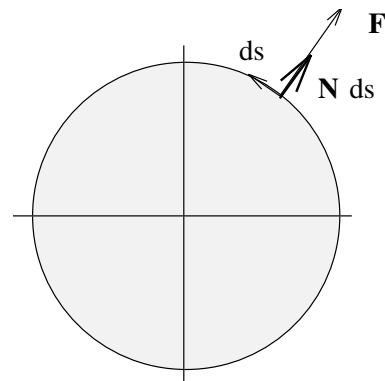
$$\int_C \mathbf{F} \cdot \mathbf{N} ds = \int_0^{2\pi} R^2 d\theta = 2\pi R^2.$$

We could also calculate this analytically as follows. Choose the parametric representation $\mathbf{r} = \langle R \cos \theta, R \sin \theta \rangle$. Thus,

$$\begin{aligned} x &= R \cos \theta & dx &= -R \sin \theta d\theta \\ y &= R \sin \theta & dy &= R \cos \theta d\theta. \end{aligned}$$

Hence,

$$\begin{aligned} \int_C -F_2 dx + F_1 dy &= \int_0^{2\pi} (-R \sin \theta)(-R \sin \theta d\theta) + (R \cos \theta)(R \cos \theta d\theta) \\ &= \int_0^{2\pi} R^2 d\theta = 2\pi R^2. \end{aligned}$$



Using the above definition for flux, the plane version of the divergence theorem looks very much like the version in space.

$$\int_{\partial D} \mathbf{F} \cdot \mathbf{N} ds = \iint_D \nabla \cdot \mathbf{F} dA. \quad (69)$$

On the left, as noted above, the flux is an integral over a curve rather than a surface, and on the right, we have a double integral rather than a triple integral. In addition, since $\mathbf{F} = \langle F_1, F_2 \rangle$, the divergence of \mathbf{F} is

$$\nabla \cdot \mathbf{F} = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y}.$$

Formula (69) is one form of *Green's Theorem* in the plane. It holds if the components of \mathbf{F} have continuous partial derivatives..

The proof of Green's Theorem is very similar to the proof of the divergence theorem in space. In fact, it is even easier since regions in the plane are easier to deal with than regions in space. We won't go over it again.

Green's Theorem is usually stated in another equivalent form. Let $\mathbf{F} = \langle F_1, F_2 \rangle$ be a vector field in \mathbf{R}^2 , and consider the *perpendicular field* $\mathbf{G} = \langle G_1, G_2 \rangle$ where $G_1 = F_2$ and $G_2 = -F_1$. Applying (69) to \mathbf{G} yields

$$\int_{\partial D} \mathbf{G} \cdot \mathbf{N} \, ds = \iint_D \nabla \cdot \mathbf{G} \, dA.$$

However,

$$\mathbf{G} \cdot \mathbf{N} \, ds = -G_2 dx + G_1 dy = -(-F_1)dx + F_2 dy = F_1 dx + F_2 dy$$

and this is nothing other than what we called $\mathbf{F} \cdot d\mathbf{r}$ when we were discussing line integrals. Hence, the integral on the left becomes the line integral

$$\int_C \mathbf{F} \cdot d\mathbf{r}.$$

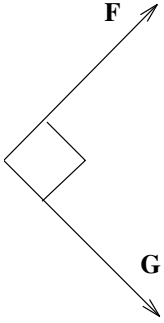
Similarly, the divergence of \mathbf{G} is

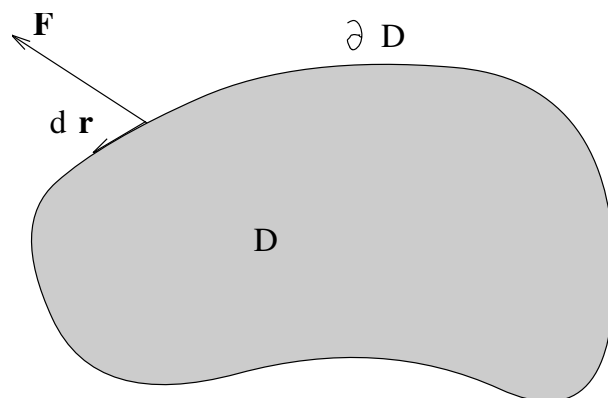
$$\begin{aligned} \frac{\partial G_1}{\partial x} + \frac{\partial G_2}{\partial y} &= \frac{\partial F_2}{\partial x} + \frac{\partial(-F_1)}{\partial y} \\ &= \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}. \end{aligned}$$

Hence, we get the following alternate form of the theorem.

Theorem 5.6 (Green's Theorem) Let \mathbf{F} be a smooth vector field in the plane. Suppose D is a region contained in the domain of \mathbf{F} which is bounded by a finite collection of smooth curves. Then

$$\int_{\partial D} \mathbf{F} \cdot d\mathbf{r} = \iint_D \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dA. \quad (70)$$





Applications of Green's Theorem Green's Theorem may be used to calculate line integrals by reducing to easier double integrals. This is analogous to using the divergence theorem to calculate surface integrals in terms of volume integrals.

Example 114 Let $\mathbf{F}(x, y) = \langle -y, x \rangle$ and let \mathcal{C} be the rectangle with vertices $(1, 2)$, $(3, 2)$, $(3, 3)$, and $(1, 3)$. Assume \mathcal{C} is traversed in the counter-clockwise direction. We can use Green's theorem to calculate $\int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r}$ as follows. Let D be the region enclosed by the rectangle, so $\mathcal{C} = \partial D$. Note that \mathcal{C} is traversed so that D is always on the left. Then,

$$\begin{aligned} \int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r} &= \iint_D \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} dA \\ &= \iint_D (1 - (-1)) dA = 2A(D) = 2 \times 2 = 4. \end{aligned}$$

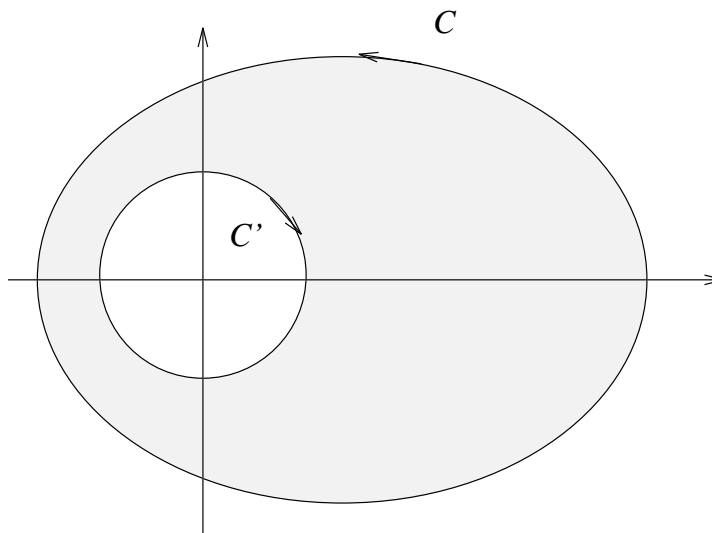


Example 115 Let

$$\mathbf{F}(x, y) = \frac{-y}{x^2 + y^2} \mathbf{i} + \frac{x}{x^2 + y^2} \mathbf{j} = \frac{1}{r} \mathbf{u}_{\theta} \quad \text{for } (x, y) \neq (0, 0),$$

and let \mathcal{C} be the ellipse $\frac{(x-1)^2}{9} + \frac{y^2}{4} = 1$. Assume \mathcal{C} is traversed counter-clockwise. One is tempted to try to use Green's theorem for the region D contained inside \mathcal{C} . Unfortunately, the vector field is not smooth at the origin, so the theorem does not apply to D . However, we can attempt the same trick we used in the case of the inverse square field for a surface enclosing the origin. (See Section 7.)

Namely, choose a circle \mathcal{C}' centered at the origin with radius small enough to fit inside the ellipse \mathcal{C} . Let D' be the region lying *between* \mathcal{C}' and \mathcal{C} . \mathbf{F} is smooth in D' , so Green's theorem does apply. The boundary of D comes in two disconnected components: $\partial D = \mathcal{C} \cup \mathcal{C}'$. Also, with \mathcal{C} traversed counter-clockwise, D' will be on its left as required, but \mathcal{C}' must be traversed *clockwise* for D' to be on its left.



With these assumptions,

$$\int_{C \cup C'} \mathbf{F} \cdot d\mathbf{r} = \iint_D \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} dA.$$

The integrand on the right was calculated in Example 7 of Section 3.

There we showed that

$$\frac{\partial F_2}{\partial x} = \frac{y^2 - x^2}{(x^2 + y^2)^2} = \frac{\partial F_1}{\partial y}$$

so that the difference is zero. Hence, the integral on the right is zero. Expanding the integral on the left, we obtain

$$\int_C \mathbf{F} \cdot d\mathbf{r} + \int_{C'} \mathbf{F} \cdot d\mathbf{r} = 0.$$

The second line integral has been done many times for this vector field in this course, and the answer is -2π . The minus sign, of course, arises because the path is traversed clockwise, which is opposite to the usual direction. Transposing, we obtain

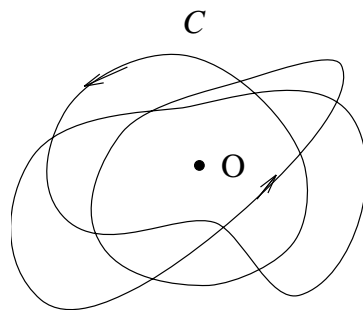
$$\int_C \mathbf{F} \cdot d\mathbf{r} = 2\pi.$$

Note that this same argument would have worked for any path C which is the boundary of a region D containing the origin. In fact, for $\mathbf{F} = (1/r)\mathbf{u}_\theta$, we have

$$\int_{\partial D} \mathbf{F} \cdot d\mathbf{r} = \begin{cases} 0 & \text{if the origin is not in } D \\ 2\pi & \text{if the origin is in } D \end{cases}.$$

The case in which D contains the origin is covered by the argument used in the example—by excising a small disk from D . The case in which D does not contain the origin follows directly from Green's theorem, since in that case the integrand on the right is continuous and is zero everywhere in D .

One sometimes need the line integral $\int_C (1/r) \mathbf{u}_\theta \cdot d\mathbf{r}$ for a closed curve C which goes around the origin more than once. In that case, the curve must intersect itself (or overlap) and it cannot be the boundary of a bounded region D . The integral is $\pm 2\pi n$, where n is the number of times the curve goes around the origin, and the sign depends on whether it is traversed counter-clockwise or clockwise. (Can you prove that?) The case $n = 0$ corresponds to the curve not going around the origin at all.

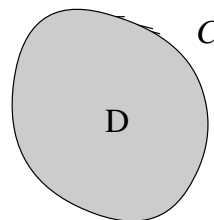


Area by Green's theorem Both Green's theorem and the divergence theorem are 'normally' used to calculate the left hand side by reducing it to the right hand side. However, there are occasions where one reverses this. For example, consider the vector field $\mathbf{F}(x, y) = \langle -y, x \rangle$. The double integral in Green's theorem is

$$\iint_D (1 - (-1)) dA = 2 \iint_D dA = 2A(D).$$

Hence, Green's theorem gives us the following formula for the area of D

$$A(D) = \frac{1}{2} \int_{\partial D} -y dx + x dy. \quad (71)$$



This seems a bizarre way to calculate an area, but it is sometimes useful.

Example 116 Let D be the area inside the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$. Parameterize the ellipse ∂D by

$$x = a \cos \theta \quad y = b \sin \theta \quad 0 \leq \theta \leq 2\pi.$$

Then

$$dx = -a \sin \theta d\theta \quad dy = b \cos \theta d\theta$$

so

$$-y dx + x dy = -(b \sin \theta)(-a \sin \theta d\theta) + (a \cos \theta)(b \cos \theta d\theta) = ab d\theta.$$

It follows from (71) that

$$A(D) = \frac{1}{2} ab(2\pi) = \pi ab.$$

You should compare this with the usual way of calculating this area, and also the method using the change of variables formula and the Jacobian.

It should be noted that the area can also be calculated using the line integral of the vector field $\mathbf{F} = \langle -y, 0 \rangle$ or $\mathbf{F} = \langle 0, x \rangle$ which give you the method you learned in your single variable calculus course. (Why?) Formula (71) is more symmetric, and so has a better chance of simplifying the calculation.

Exercises for 5.8.

1. Let $\mathbf{F} = \langle x, y \rangle$. (Then $\nabla \cdot \mathbf{F} = 2$.) Verify formula (69) (the first form of Green's Theorem) for each of the following curves by calculating the outward flux through the given curve and checking that it is twice the area enclosed.
 - (a) A circle of radius a centered at the origin.
 - (b) A square of side a in the first quadrant with opposite vertices at $(0, 0)$ and (a, a) .
2. Let $\mathbf{F} = \langle -y, x \rangle$. (Then $\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} = 2$.) Verify formula (70) (the second form of Green's Theorem) for each of the following curves by calculating the line integral for the given curve and checking that it is twice the area enclosed.
 - (a) A circle of radius a centered at the origin, traversed counterclockwise.
 - (b) A square of side a in the first quadrant with opposite vertices at $(0, 0)$ and (a, a) , traversed counterclockwise.
3. Use Green's Theorem (second form) to calculate $\int_C (y^2)dx + (x^2)dy$ where C is the path which starts at $(0, 0)$ goes to $(3, 0)$ then to $(3, 3)$ and finally back to $(0, 0)$.
4. Use Green's Theorem (second form) to calculate $\int_C -y dx + x dy$ where C is the semi-circular path from $(a, 0)$ to $(-a, 0)$ with $y \geq 0$. Hint: C is not a closed path, but you can form a closed path by adding the linear path C' from $(-a, 0)$ to $(a, 0)$. Use Green's Theorem to relate \int_C to $\int_{C'}$ through the use of $\int_{C \cup C'}$.
5. Let $\mathbf{F}(x, y) = \langle y + \sin(x^2), (1 + y^2)^{1/5} \rangle$. Calculate $\int_C \mathbf{F} \cdot d\mathbf{r}$ for C the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$, traversed counterclockwise. Hint: You should remember what the area of an ellipse is.
6. Let $\mathbf{F}(x, y) = \langle e^{x^3}, xy \rangle$. Calculate $\int_C \mathbf{F} \cdot d\mathbf{r}$ for C the path bounded below by the graph of $y = x^2$ and above by the graph of $x = y^2$ and traversed in the counterclockwise direction.
7. What is $\int_C \frac{-y}{x^2 + y^2} dx + \frac{x}{x^2 + y^2} dy$ for each of the following curves? (a) The ellipse $x^2/9 + y^2/16 = 1$ traversed counterclockwise. (b) The triangle with vertices $(1, 1)$, $(2, 3)$, and $(0, 6)$ traversed clockwise.
8. Throughout this problem, take $\mathbf{F} = \frac{-y}{r}\mathbf{i} + \frac{x}{r}\mathbf{j}$ where $r = \sqrt{x^2 + y^2}$.
 - (a) Show that $\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} = \frac{1}{r}$.
 - (b) Show by direct calculation that $\int_C \mathbf{F} \cdot d\mathbf{r} = 2\pi$ where C is the circle $x^2 + y^2 = 1$ traversed counterclockwise.
 - (c) Let C be the square with vertices at the four points $(\pm 2, \pm 2)$ and assume C is traversed counterclockwise. Use Green's Theorem for a region with a

hole to show that

$$\int_C \mathbf{F} \cdot d\mathbf{r} = 2\pi + 8 \int_0^{\pi/4} \int_1^{2 \sec \theta} dr d\theta.$$

- (d) Evaluate the double integral to complete the calculation.
- (e) Calculate $\int_C \mathbf{F} \cdot d\mathbf{r}$ directly. Hint: Calculate the line integral for one side of the square and multiply by 4.
9. Find the area bounded by each of the following curves by using the formula $\frac{1}{2} \int_C -y dx + x dy$. Compute the area by another method and compare the answers.
- (a) $y = 0, x = 1, y = x^2$. Hint: You could parameterize the parabola by $\mathbf{r} = \langle (1-t), (1-t)^2 \rangle, 0 \leq t \leq 1$.
- (b) $\mathbf{r} = \langle t^2, t^3 \rangle, -1 \leq t \leq 1$ and $x = 1$. Hint: For an alternate method note that the parametric representation describes the two curves $y = \pm x^{3/2}$.
10. Let $P_0 = (0, 0), P_1 = (x_1, y_1), P_2 = (x_2, y_2)$ be the vertices of a triangle in the plane. Derive a formula for its area using Green's Theorem. Hint: You should get the same answer as you would by taking half the cross product of the vectors from the origin to the other vertices.
11. Derive the first form of Green's theorem (69) from the divergence theorem using the 'pillbox argument' as follows. Let $\mathbf{F} = \langle F_1, F_2, 0 \rangle$ and assume it depends only on x and y . Let D be a bounded region in the x, y -plane, and let E be a 'pillbox' produced by extending it upward one unit in the z -direction. (a) Show $\iint_{\partial E} \mathbf{F} \cdot \mathbf{N} dS = \iint_{\partial D} \mathbf{F} \cdot \mathbf{N} ds$ as follows. First note that the flux through the top and bottom are zero. (Why?) Then calculate the flux through the sides using an element of area of height 1 and base ds . (b) Show $\iiint_E \nabla \cdot \mathbf{F} dV = \iint_D \nabla \cdot \mathbf{F} dA$ by considering an element of volume which is a thin 'column' of height 1 and base dA .

5.9 Stokes's Theorem

Stokes's theorem is a generalization of the second form of Green's theorem (for line integrals). To motivate it, notice that the integrand

$$\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}$$

in the double integral looks like one of the components of the curl. To make this more explicit, view the plane vector field \mathbf{F} as a vector field in space with its third component zero, i.e., put

$$\mathbf{F} = \langle F_1, F_2, 0 \rangle.$$

Then

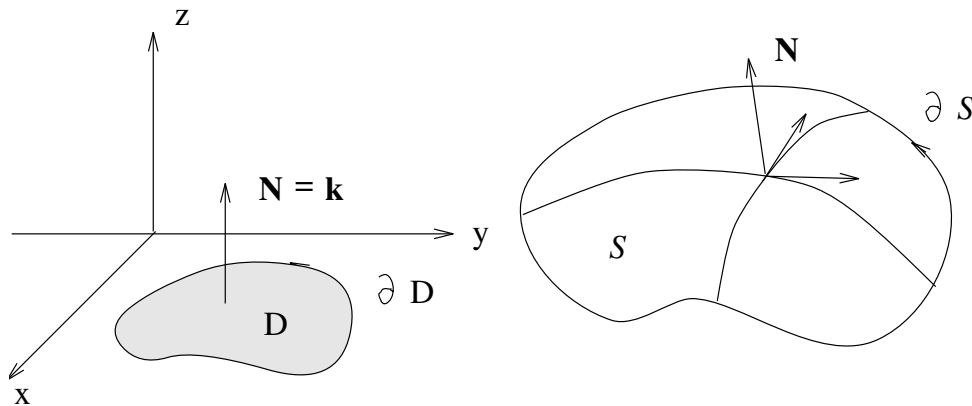
$$\begin{aligned}\nabla \times \mathbf{F} &= \left\langle \frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}, \frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}, \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right\rangle \\ &= \left\langle 0, 0, \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right\rangle.\end{aligned}$$

(The first two components are zero because $F_3 = 0$ and \mathbf{F} is a *plane* vector field so its components F_1 and F_2 are functions only of x and y . However, we don't actually have to worry about that for the moment.) If we treat D as a surface which happens to lie in the x, y -plane, and if we use the upward pointing normal $\mathbf{N} = \mathbf{k}$, we have

$$\int_D (\nabla \times \mathbf{F}) \cdot \mathbf{N} \, dS = \iint_D \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \, dA.$$

Hence, Green's theorem can be rewritten

$$\int_{\partial D} \mathbf{F} \cdot d\mathbf{r} = \iint_D (\nabla \times \mathbf{F}) \cdot \mathbf{N} \, dS.$$



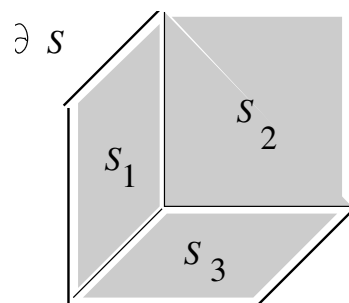
Stokes's theorem asserts that this same formula works for any surface in space and the curve bounding it. That is, it works if we are careful to arrange the orientation of the surface and the curve carefully, so we will devote some attention to that point.

Suppose then that \mathcal{S} is an oriented surface in space. That means that a unit normal vector \mathbf{N} has been specified at each point of \mathcal{S} , and that these normals are related to each other in some coherent way as we move around on the surface. For example, if the surface is given parametrically by $\mathbf{r} = \mathbf{r}(u, v)$, then the formula

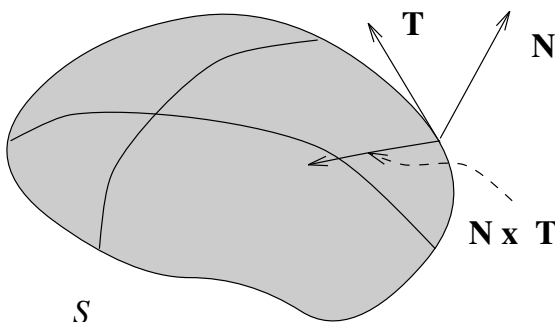
$$d\mathbf{S} = \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \, du \, dv$$

gives a preferred normal direction at each point. More generally \mathcal{S} might consist of a finite collection of such surfaces which are joined together along smooth curves along their edges. An example of such a surface would be three faces of a cube with a common vertex. In such cases, it is not so easy to explain how the unit normals should be related when you cross an edge, but in most cases it is intuitively clear what to do. (We shall discuss how to do this rigorously later.)

Generally, such a surface \mathcal{S} will have a well defined *boundary* $\partial\mathcal{S}$, which will be a smooth curve in space or a finite collection of such curves. There will be two possible ways to traverse this boundary, and we specify one, i.e., we choose an orientation for the boundary. In particular, that means that at each point, we specify a unit tangent vector \mathbf{T} . At each point on the boundary consider the cross-product $\mathbf{N} \times \mathbf{T}$ of the unit normal to the surface and the unit tangent vector to its boundary. This cross-product will be tangent to the surface and perpendicular to its boundary. (See the diagram.) Hence, it either points in towards the surface or out away from it. We shall say that the orientation of the surface and the orientation of its boundary are *consistent* if $\mathbf{N} \times \mathbf{T}$ always points toward \mathcal{S} . This can be said more graphically as follows. If you imagine yourself—suitably reduced in size—walking around the edge of the surface in the preferred direction for $\partial\mathcal{S}$, with your head in the direction of the preferred normal, then the region will always be on your *left*. As you see this is a natural generalization to a curved surface of the relationship we required for a plane region in Green's theorem.



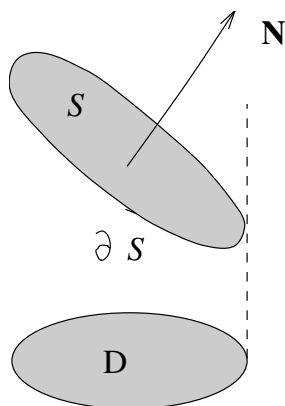
$$S = S_1 + S_2 + S_3$$



Theorem 5.7. (Stokes's Theorem) Let \mathbf{F} be a vector field in space with continuous partial derivatives. Let \mathcal{S} be a surface obtained by patching together smooth surfaces along a finite collection of smooth curves and such that $\partial\mathcal{S}$ consists of a finite collection of smooth curves. Assume finally that the orientations of \mathcal{S} and $\partial\mathcal{S}$ are consistent. Then

$$\int_{\partial\mathcal{S}} \mathbf{F} \cdot d\mathbf{r} = \iint_{\mathcal{S}} (\nabla \times \mathbf{F}) \cdot \mathbf{N} \, dS.$$

We will discuss the proof of Stokes's theorem in the next section, but first we give some applications.



Applications of Stokes's Theorem By now you should be beginning to get the idea. Stokes's theorem is normally used to calculate a line integral by setting it equal to a surface integral.

Example 117 Let $\mathbf{F}(\mathbf{r}) = \langle -y, x, 0 \rangle = r\mathbf{u}_\theta$, and let \mathcal{C} be the ellipse which is the intersection of the cylinder $x^2 + y^2 = 1$ with the plane $x + y + z = 2$. Assume \mathcal{C} is traversed in the counter-clockwise direction when viewed from above. We shall use Stokes's theorem to calculate $\int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r}$. To do this, we need to find a surface \mathcal{S} with boundary $\partial\mathcal{S} = \mathcal{C}$. Since we are in space, there are in fact infinitely many such surfaces. We shall do the problem two different ways, each of which is instructive.

First, the portion of the plane $x + y + z = 2$ contained within the ellipse is one possible \mathcal{S} . To be consistent with the orientation of the curve \mathcal{C} , we need to use the 'upward' pointing normals for \mathcal{S} . (See the diagram.)

The curl of this vector field is easily calculated: $\nabla \times \mathbf{F} = \langle 0, 0, 2 \rangle$. Hence,

$$\int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r} = \iint_{\mathcal{S}} (\nabla \times \mathbf{F}) \cdot d\mathbf{S} = \iint_{\mathcal{S}} (2\mathbf{k}) \cdot d\mathbf{S}.$$

Thus to complete the calculation, we must evaluate the surface integral on the right. The easiest way to do this is to treat the surface as the graph of the function given by $z = f(x, y) = 2 - x - y$ with domain D the disk, $x^2 + y^2 \leq 1$, in the x, y -plane. Then

$$d\mathbf{S} = \langle -f_x, -f_y, 1 \rangle dy dx = \langle -1, -1, 1 \rangle dy dx$$

and the surface integral is

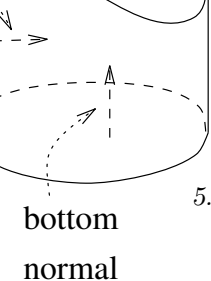
$$\iint_{\mathcal{S}} (2\mathbf{k}) \cdot d\mathbf{S} = \iint_D 2 dA = 2 A(D) = 2\pi \cdot 1^2 = 2\pi.$$

Thus,

$$\int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r} = 2\pi.$$

Here is another way to do it. Let \mathcal{S} be the surface made up by taking the part of the cylinder $x^2 + y^2 = 1$ between the ellipse and the x, y -plane together with the disk $x^2 + y^2 \leq 1$ in the x, y -plane. You can think of \mathcal{S} as obtained by taking a tin can with bottom in the x, y -plane and cutting it off by the plane $x + y + z = 2$. The result is an open can with a slanting elliptical top edge \mathcal{C} . It is a little difficult to see, but the proper direction for the normal vectors on the lateral cylindrical surface is *inward* toward the z -axis. Then, as you cross the circular edge in the x, y -plane to the disk forming the bottom component of \mathcal{S} , you need to choose the *upward* pointing normal. Since $\nabla \times \mathbf{F} = 2\mathbf{k}$ is parallel to the lateral surface, the flux through that is zero. On the bottom surface, we have $(\nabla \times \mathbf{F}) \cdot \mathbf{N} = 2\mathbf{k} \cdot \mathbf{k} = 2$. Hence, the flux through the bottom surface is 2 times the area of the disk, i.e., 2π . Thus,

$$\int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r} = \iint_{\mathcal{S}} (\nabla \times \mathbf{F}) \cdot \mathbf{N} dS = 0 + 2\pi = 2\pi.$$



There is one interesting variation of the above calculation. Let \mathcal{S} be just the lateral surface of the cylinder between the plane $x + y + z = 2$ and the x, y -plane. Then $\partial\mathcal{S}$ has two components: the top edge \mathcal{C} and the bottom edge \mathcal{C}' which is the circle $x^2 + y^2 = 1$ in the x, y -plane. Since we need to choose the inward pointing normals on the lateral surface (for consistency with the orientation of \mathcal{C}), we need to choose the *clockwise* orientation of \mathcal{C}' for consistency with that inward normal. Then

$$\int_{\mathcal{C} \cup \mathcal{C}'} \mathbf{F} \cdot d\mathbf{r} = \iint_{\mathcal{S}} (\nabla \times \mathbf{F}) \cdot \mathbf{N} dS = 0$$

since as above $\nabla \times \mathbf{F} = 2\mathbf{k}$ is parallel to \mathcal{S} . Hence,

$$\int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r} + \int_{\mathcal{C}'} \mathbf{F} \cdot d\mathbf{r} = 0$$

or

$$\int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r} = - \int_{\mathcal{C}'} \mathbf{F} \cdot d\mathbf{r}.$$

However, $\int_{\mathcal{C}'} \mathbf{F} \cdot d\mathbf{r}$ is an integral we have done several times in the past, and it is equal to -2π . (Actually, we did it with the opposite orientation and got 2π .) Hence, the final answer is 2π .

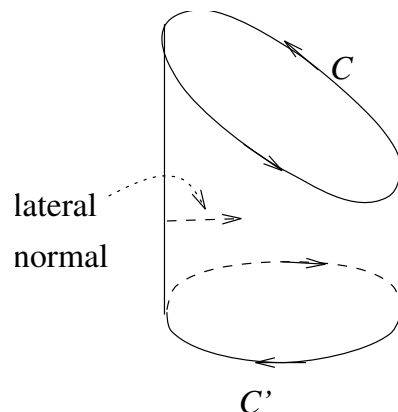
The above example illustrates the interesting ways geometric reasoning can help when applying Stokes's theorem. Sometimes you should look for a non-obvious surface which may allow you to simplify the calculation. Also, it is sometimes useful to use Stokes's theorem to reduce the desired line integral to another line integral which is easier to calculate. The second principle is also illustrated by the next example.

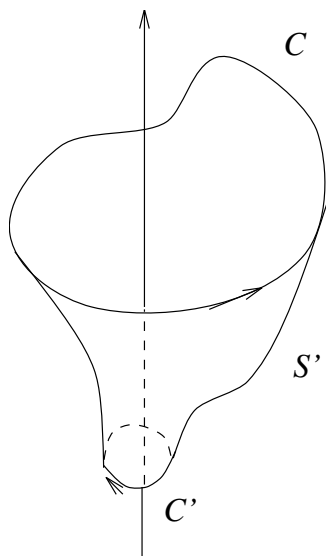
Example 118 Let

$$\mathbf{F}(\mathbf{r}) = \frac{1}{r} \mathbf{u}_\theta = \left\langle \frac{-y}{x^2 + y^2}, \frac{x}{x^2 + y^2}, 0 \right\rangle.$$

Note that \mathbf{F} blows up on the z -axis, so its domain is \mathbf{R}^3 with the z -axis deleted. This field, except for some constant, is the magnetic field intensity produced by a unit of current flowing in a thin wire along the z -axis. The lines of force are circles centered on the z -axis. Its curl is $\nabla \times \mathbf{F} = \mathbf{0}$. (You should do the calculation which is very similar to the one done for the analogous vector field in the plane in Section 7.)

Let \mathcal{C} be any closed curve which goes once around the z -axis in the counter-clockwise direction when viewed from above. Since $\nabla \times \mathbf{F} = \mathbf{0}$, you might think you could simply span the curve \mathcal{C} by any surface whatsoever and the resulting surface integral would be zero. Unfortunately, any surface spanning such a curve *must intersect the* z -axis. Since \mathbf{F} is singular on the z -axis, Stokes's theorem does not apply for such





a surface. Instead, we must proceed as indicated above. Let C' be a small circle in the x, y -plane centered at the origin. Let S' be any surface extending from C' to C . If the normals on S' are consistent with the orientation of C , then C' must be traversed in the clockwise direction. $\int_{C'} \mathbf{F} \cdot d\mathbf{r}$ has been calculated several times (e.g., once in the previous section); it equals -2π . Thus, by Stokes's theorem

$$\int_C \mathbf{F} \cdot d\mathbf{r} + \int_{C'} \mathbf{F} \cdot d\mathbf{r} = \iint_{S'} (\nabla \times \mathbf{F}) \cdot d\mathbf{S} = 0$$

or

$$\int_C \mathbf{F} \cdot d\mathbf{r} = - \int_{C'} \mathbf{F} \cdot d\mathbf{r} = -(-2\pi) = 2\pi.$$

The line integral is zero if the curve C does not go around the z -axis. (Can you prove that?)

Physical interpretation of Curl As in the case of the divergence of a vector field, we can use Stokes's theorem to give an interpretation of $\nabla \times \mathbf{F}$ which is independent of any coordinate system.

Fix a point P with position vector \mathbf{r} at which we want to calculate $\nabla \times \mathbf{F}$. Choose a normal direction \mathbf{N} at P . In the plane passing through P perpendicular to \mathbf{N} , consider a small circle C of radius R centered at P . Traverse C in the counter-clockwise direction when looked from the 'top' relative to \mathbf{N} . The line integral $\int_C \mathbf{F} \cdot d\mathbf{r}$ is called the *circulation* of the vector field along the closed curve C . If \mathbf{F} is the momentum field of a fluid flow, you can think of the circulation as indicating the average twisting effect of the field along the curve. By Stokes's theorem, we have

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \iint_S (\nabla \times \mathbf{F}) \cdot \mathbf{N} dS$$

where S is the disk spanning the circle C . However, by the average value property for integrals, we have

$$\iint_S (\nabla \times \mathbf{F}) \cdot \mathbf{N} dS = (\nabla \times \mathbf{F})(\mathbf{r}') \cdot \mathbf{N} A(S)$$

where the curl has been evaluated at an appropriate point with position vector \mathbf{r}' in the disk S . Thus,

$$(\nabla \times \mathbf{F})(\mathbf{r}') \cdot \mathbf{N} = \frac{1}{A(S)} \int_C \mathbf{F} \cdot d\mathbf{r}.$$

If we take the limit as $R \rightarrow 0$, i.e., as C shrinks to the point P , $\mathbf{r}' \rightarrow \mathbf{r}$, so

$$(\nabla \times \mathbf{F})(\mathbf{r}) \cdot \mathbf{N} = \lim_{C \rightarrow P} \frac{1}{A(S)} \int_C \mathbf{F} \cdot d\mathbf{r}.$$

The quantity on the right is called the *limiting circulation per unit area about the axis* \mathbf{N} . In the fluid flow model, it can be thought of as the twisting effect on a

small ‘paddle wheel’ with axis \mathbf{N} . As its axis is shifted, the paddle wheel will spin faster or slower, and it may even reverse direction. This all depends on the relation between the curl $(\nabla \times \mathbf{F})(\mathbf{r})$ at the point and the axis \mathbf{N} of the paddle wheel.

In the above analysis, we used a family of small circles converging to P , but the analysis would work as well for any family of curves converging to P . For example, they could be squares or triangles in the plane perpendicular to \mathbf{N} . Indeed, the curves need not be plane curves perpendicular to the specified normal direction \mathbf{N} as long as their normals all converge to \mathbf{N} .

We can use this interpretation of the curl (in the most general form) to show that $\nabla \times \mathbf{F} = \mathbf{0}$ for the field $\mathbf{F}(\mathbf{r}) = \frac{1}{r}\mathbf{u}_\theta$ considered in Example 118 above. It is probably easier to do that by direct calculation from the formula, but the geometric method is instructive. We base the calculation on cylindrical coordinates. Fix a point P not on the z -axis. To show $\nabla \times \mathbf{F}$ is zero at P , it suffices to show that its components in three mutually perpendicular directions are zero. We shall show that

$$(\nabla \times \mathbf{F}) \cdot \mathbf{N} = 0$$

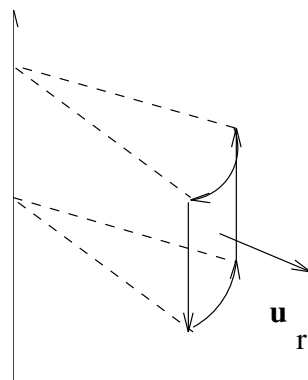
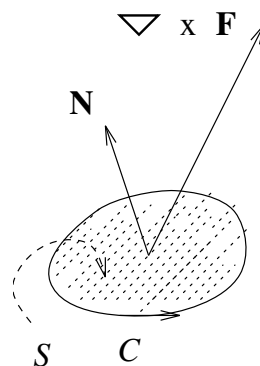
for $\mathbf{N} = \mathbf{u}_r$, (pointing radially away from the z -axis), $\mathbf{N} = \mathbf{u}_\theta$, (pointing tangent to circles centered on the z -axis), and $\mathbf{N} = \mathbf{k}$, (pointing parallel to the z -axis). For \mathbf{u}_r , consider a curvilinear rectangle centered at the point P on a cylinder (of radius r) through the point P . You should examine the diagram carefully to make sure you understand the direction in which this rectangle must be traversed to be consistent with normal direction $\mathbf{N} = \mathbf{u}_r$. In particular, note that the bottom edge is traversed in the direction of positive θ and the top edge in the direction of negative θ .

The circulation (line integral) of \mathbf{F} along this rectangle decomposes into 4 terms, one for each side of the rectangle. \mathbf{F} is perpendicular to each of the vertical edges, so the line integrals for those are zero. The line integral for the bottom edge is easy to calculate because the vector field is tangent to the edge and constant on it. (The answer is $\frac{1}{r}r\Delta\theta = \Delta\theta$ where $\Delta\theta$ is the angle subtended by the edge at the z -axis, but the actual value is not needed in the argument.) The calculation of the line integral for the top edge is exactly the same except that the sign changes because it is traversed in the opposite direction. Hence, the net circulation around the rectangle is zero. If we divide by the area of the rectangle, and take the limit as the rectangle shrinks to the point, we still get zero. It follows that $(\nabla \times \mathbf{F}) \cdot \mathbf{u}_r = 0$.

The calculations showing that $(\nabla \times \mathbf{F}) \cdot \mathbf{u}_\theta = 0$ and $(\nabla \times \mathbf{F}) \cdot \mathbf{k} = 0$ are similar, and are left to the student as exercises. (The hardest one is the one for \mathbf{k} .)

Exercises for 5.9.

1. Let $\mathbf{F}(x, y, z) = -y\mathbf{i} + x\mathbf{j} + z\mathbf{k}$. Use Stokes's Theorem to evaluate $\int_C \mathbf{F} \cdot d\mathbf{r}$ for each of the following paths.



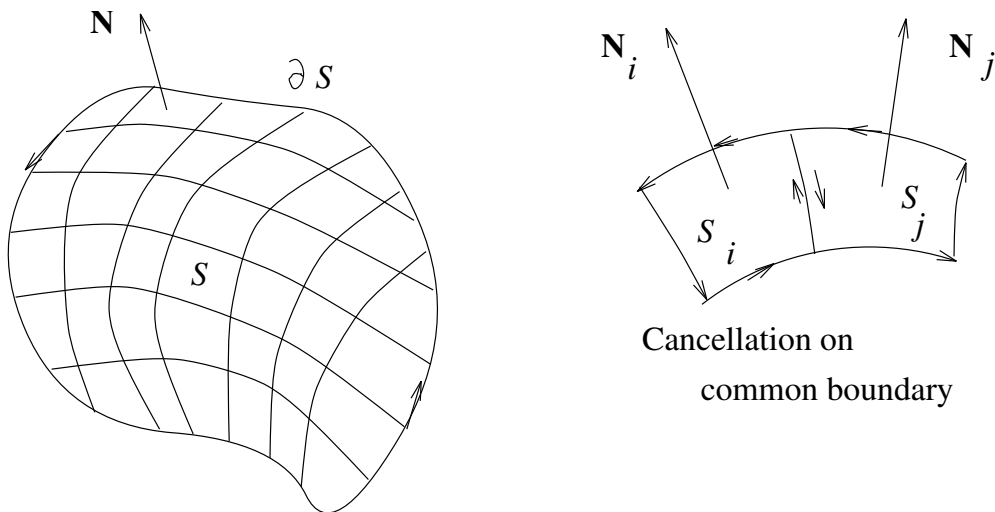
- (a) A circle of radius a in the plane $z = 5$ centered on the z axis and traversed counterclockwise when viewed from above.
- (b) A circle of radius a in the plane $x = 3$ centered on the x -axis and traversed counterclockwise when viewed from in front.
- (c) The intersection of the paraboloid $z = x^2 + y^2$ with the paraboloid $z = 8 - x^2 - y^2$ traversed counterclockwise when viewed from above.
2. For each of the following surfaces \mathcal{S} determine the proper coherent direction of the normal vector if the boundary $\mathcal{C} = \partial\mathcal{S}$ is traversed in the indicated direction.
- (a) The half open cylinder $x^2 + z^2 = 1, 0 \leq y \leq 2$, closed off on the left by the disk $x^2 + z^2 \leq 1$ in the x, z -plane. Assume the boundary $x^2 + z^2 = 1, y = 2$ is traversed counterclockwise with respect to the positive y -axis, i.e., it comes toward you and down in the first octant.
- (b) The half open cylinder as in part (a), but open on the left and closed off on the right. Assume \mathcal{C} is also traversed counterclockwise with respect to the positive y -axis.
- (c) The part of the paraboloid $z = 25 - x^2 - 2y^2$ above the plane $x + y + z = 1$. Assume the boundary \mathcal{C} is traversed counterclockwise with respect to the positive z -axis.
3. For each of the following surfaces \mathcal{S} with the indicated normal vectors, describe the proper coherent orientation for its boundary $\mathcal{C} = \partial\mathcal{S}$.
- (a) The part of the cone $z = \sqrt{x^2 + 4y^2}$ below the plane $3x + 4y + z = 10$. Use normals pointing roughly away from the z -axis.
- (b) The 3 faces of the unit cube in the first octant which lie in the coordinate planes. Use normals pointing into the first octant. Make sure you indicate the proper orientation for each of the 6 components of the boundary.
- (c) The portion of the hyperboloid of one sheet $x^2 - y^2 + z^2 = 1$ between the planes $y = 1$ and $y = -1$. Use normals pointing away from the y -axis. Hint: Note that the boundary consists of two disconnected curves.
4. Let $\mathbf{F}(x, y, z) = -y\mathbf{i} + x\mathbf{j} + z\mathbf{k}$. Use Stokes's Theorem to evaluate $\int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r}$ where \mathcal{C} is the intersection of the surface of the unit cube in the first octant with the plane $z = \frac{x}{10} + \frac{y}{12}$. Assume the curve is traversed counterclockwise when viewed from above. Hint: Think of \mathcal{C} as the boundary of the surface consisting of the vertical faces of the cube below the curve together with the bottom face.
5. Evaluate $\int_{\mathcal{C}} (\sin(x^3) + y)dx + (e^{y^2} + x)dy + \cos z dz$ for \mathcal{C} the intersection of the ellipsoid $\frac{x^2}{4} + \frac{y^2}{9} + \frac{z^2}{25} = 1$ with the paraboloid $z = x^2 + y^2$. Find the answer for both possible orientations of this curve.

6. Let \mathcal{S} be a simple closed surface in \mathbf{R}^3 and let \mathbf{F} be a smooth vector field with domain containing \mathcal{S} .
- (a) Assume \mathbf{F} is smooth on the interior E of \mathcal{S} . Use the divergence theorem and the formula $\nabla \cdot (\nabla \times \mathbf{F}) = 0$ to prove that $\iint_{\mathcal{S}} (\nabla \times \mathbf{F}) \cdot d\mathbf{S} = 0$. Use outward normals.
- (b) Do not assume that \mathbf{F} is necessarily smooth on the interior of \mathcal{S} . Prove the above formula by means of Stokes's Theorem as follows. Decompose $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ into a union of two subsurfaces which meet only on their common boundary \mathcal{C} . Note that \mathcal{C} is traversed in opposite directions for the two surfaces if its orientation is consistent with the use of the outward orientation of normals on both.
7. Let $\mathbf{F} = \langle yz, -xz, xy \rangle$. Calculate $\nabla \times \mathbf{F}$. Let C be the unit circle $x^2 + y^2 = 1$ in the x, y -plane traversed counterclockwise. (a) Calculate $\iint_S \nabla \times \mathbf{F} \cdot \mathbf{N} dS$ for S the unit disk in the x, y -plane. (b) Make the same calculation for S the hemisphere $x^2 + y^2 + z^2 = 1, z \leq 0$. (c) Make the same calculation if S is made up of the cylindrical surface $x^2 + y^2 = 1, 0 \leq z \leq 1$ capped off on top by the the plane disk $z = 1, x^2 + y^2 \leq 1$. Note that in each case the boundary of S is the curve C , so the answers should all be the same (in this case 0). Make sure you choose the proper direction for the normal vectors for each surface.
8. Let \mathbf{F} be a smooth vector field in \mathbf{R}^3 .
- (a) Assume that $\int_C \mathbf{F} \cdot d\mathbf{r} = 0$ for every sufficiently small rectangular path, whatever its orientation. Show that $\nabla \times \mathbf{F} = 0$.
- (b) Suppose we only assume that $\int_C \mathbf{F} \cdot d\mathbf{r} = 0$ for every sufficiently small rectangular path which lies in a plane parallel to one of the coordinate planes. Can we still conclude that $\nabla \times \mathbf{F} = 0$? Explain.
9. Let $\mathbf{F} = \frac{1}{r} \mathbf{u}_\theta$ where $r = \sqrt{x^2 + y^2}$.
- (a) Let C_1 be a circular arc which is part of a circle of radius r parallel to the x, y -plane, centered on the z -axis and with the arc subtending an angle α . Assume C_1 is oriented in the counterclockwise direction when viewed from above. Show that $\int_{C_1} \mathbf{F} \cdot d\mathbf{r} = \alpha$.
- (b) Consider the following path C in a plane parallel to the x, y -plane. Start at the point with cylindrical coordinates (r, θ, z) and move in the positive r -direction to $(r + \Delta r, \theta, z)$, then along an arc (as above) to $(r + \Delta r, \theta + \Delta \theta, z)$, then to $(r, \theta + \Delta \theta, z)$, and finally back on a circular arc to (r, θ, z) . Using part (a), show that $\int_C \mathbf{F} \cdot d\mathbf{r} = 0$.
- (c) Show that $(\nabla \times \mathbf{F}) \cdot \mathbf{k} = 0$.
- (d) Using a similar argument, show that $(\nabla \times \mathbf{F}) \cdot \mathbf{u}_\theta = 0$.

5.10 The Proof of Stokes's Theorem

We shall give two proofs. The first is not really a proof but a plausibility argument. You should study it because it will help you understand why the theorem is true. The second is a correct proof, but you can probably skip it unless you are specially interested in seeing a rigorous treatment.

The first argument is based on the 'physical interpretation' of the curl. Let \mathbf{F} be a smooth vector field in space and let \mathcal{S} be a surface with boundary $\partial\mathcal{S}$. Imagine \mathcal{S} decomposed into many small curvilinear parallelograms \mathcal{S}_i . For each \mathcal{S}_i , choose a point with position vector \mathbf{r}_i inside \mathcal{S}_i , and let \mathbf{N}_i be the normal vector at that point.



If \mathcal{S}_i is small enough, we can treat it as if it were an actual parallelogram passing through \mathbf{r}_i and normal to \mathbf{N}_i . Then, according to the physical interpretation of the curl, to a high degree of approximation we have

$$(\nabla \times \mathbf{F})(\mathbf{r}_i) \cdot \mathbf{N}_i \approx \frac{1}{A(\mathcal{S}_i)} \int_{\partial\mathcal{S}_i} \mathbf{F} \cdot d\mathbf{r}.$$

Hence,

$$\int_{\partial\mathcal{S}_i} \mathbf{F} \cdot d\mathbf{r} \approx (\nabla \times \mathbf{F})(\mathbf{r}_i) \cdot \mathbf{N}_i A(\mathcal{S}_i),$$

and, adding up, we obtain

$$\sum_i \int_{\partial\mathcal{S}_i} \mathbf{F} \cdot d\mathbf{r} \approx \sum_i (\nabla \times \mathbf{F})(\mathbf{r}_i) \cdot \mathbf{N}_i A(\mathcal{S}_i).$$

If we take the limit as the number of curvilinear parallelograms goes to infinity, the sum on the right approaches the surface integral $\iint_{\mathcal{S}} (\nabla \times \mathbf{F}) \cdot \mathbf{N} \, dS$. Consider the sum on the left. Assume for each \mathcal{S}_i that the orientation of the boundary $\partial\mathcal{S}_i$ is consistent with the unit normal \mathbf{N}_i . Let \mathcal{S}_i and \mathcal{S}_j be two adjacent curvilinear parallelograms which meet along a common edge \mathcal{C}_{ij} . Look at the diagram. If the normals \mathbf{N}_i and \mathbf{N}_j are 'coherently related' to one another, then the direction assigned to the common edge \mathcal{C}_{ij} by $\partial\mathcal{S}_i$ will be opposite to the direction assigned to it by $\partial\mathcal{S}_j$. Hence, the two line integrals for this common edge will cancel one another. That means that all internal segments of the boundaries of the curvilinear parallelograms will cancel, and the only portions left will be the external boundary $\partial\mathcal{S}$. Thus,

$$\sum_i \int_{\partial\mathcal{S}_i} \mathbf{F} \cdot d\mathbf{r} = \int_{\partial\mathcal{S}} \mathbf{F} \cdot d\mathbf{r}.$$

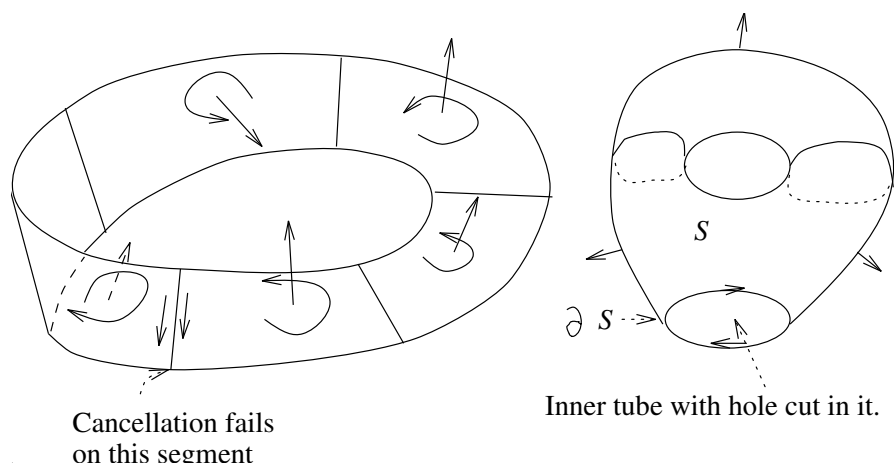
It follows that the line integral in Stokes's theorem equals the surface integral as required.

Note that this argument is not valid in the form we presented it. The reason is that we derived the physical interpretation of the curl from Stokes's theorem. Since that interpretation was a crucial step in the argument, the logic contains a vicious circle. One way around this would be to derive the physical interpretation of the curl by an independent argument. However, I know of no such argument which does not contain a hidden proof of Stokes's theorem.

A closer examination of the argument helps us understand the idea of orientation for a surface. We suppose the surface can be decomposed into patches, each small enough so that it can be assigned a coherent set of unit normals. (For example, we can assume each patch is given by a parametric representation.) For any given patch, the normal directions will impose a consistent orientation on its boundary. We say that the normals on the patches are *coherently related* if common boundary segments are traced in opposite directions for adjacent patches. (See the diagram.)

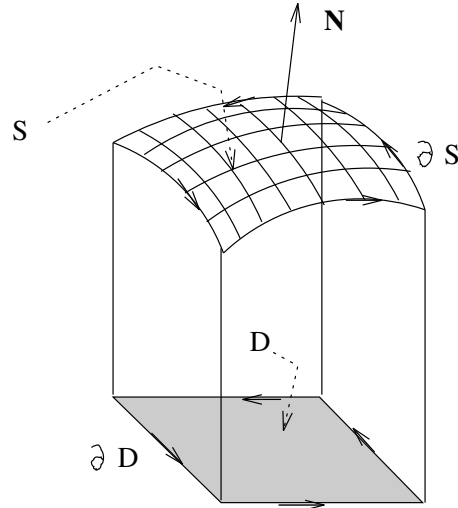
As we mentioned earlier in Section 2,

it may not in fact be possible to assign a coherent set of normals to the entire surface. The Möbius band is an example of a surface for which that is not possible, i.e., it is *non-orientable*. The diagram shows an attempt to decompose a Möbius band into patches with coherent normals and boundaries.



It should also be noted that the method of forming a surface by putting together coherently related simple patches can lead to some fairly complicated results. See the diagram for an example. Stokes's theorem still applies to the more general surfaces.

The Correct Proof (May Be Skipped) Let \mathcal{S} be an oriented surface which can be decomposed into smooth patches as described above. By further decomposition, if necessary, we may assume that each patch is the *graph of a function*. By arguing as above about mutual cancellation along common edges, the line integral along $\partial\mathcal{S}$ may be expressed as the sum of the line integrals for the boundaries of the individual patches. Similarly the surface integral may be expressed as the sum of the surface integrals for the individual patches. Thus it suffices to prove Stokes's theorem (that the line integral equals the surface integral) for each individual patch. Thus, we are reduced to proving the theorem for the graph of a function. Suppose then that \mathcal{S} is the graph of the function expressed by $z = f(x, y)$ with domain D in the x, y -plane. (Essentially the same argument will work for graphs expressible by $x = g(y, z)$ or $y = h(x, z)$.) Assume the orientation of \mathcal{S} is the one such that the z -component of \mathbf{N} is always positive. (For the reverse orientation, just reverse all the signs in the arguments below.)



First we calculate $\int_{\partial S} \mathbf{F} \cdot d\mathbf{r}$. Choose a parametric representation $x = x(t), y = y(t), a \leq t \leq b$ for ∂D , the boundary of the parameter domain. Then, $x = x(t), y = y(t), z = f(x(t), y(t))$ will be a parametric representation of ∂S . Moreover, if ∂D is traversed so that D is on the left, the resulting orientation of ∂S will be consistent with the generally upward orientation of S , as above. Then

$$\mathbf{F} \cdot d\mathbf{r} = F_1 dx + F_2 dt + F_3 dz.$$

However, $dz = f_x dx + f_y dy$ yields

$$\begin{aligned} \mathbf{F} \cdot d\mathbf{r} &= F_1 dx + F_2 dy + F_3(f_x dx + f_y dy) \\ &= (F_1 + F_3 f_x) dx + (F_2 + F_3 f_y) dy. \end{aligned}$$

Hence,

$$\int_{\partial S} \mathbf{F} \cdot d\mathbf{r} = \int_{\partial D} (F_1 + F_3 f_x) dx + (F_2 + F_3 f_y) dy = \int_{\partial D} G_1 dx + G_2 dy$$

where

$$\begin{aligned} G_1(x, y) &= F_1(x, y, f(x, y)) + F_3(x, y, f(x, y)) f_x(x, y) \\ G_2(x, y) &= F_2(x, y, f(x, y)) + F_3(x, y, f(x, y)) f_y(x, y) \end{aligned}$$

Note that $\mathbf{G} = \langle G_1, G_2 \rangle$ is a plane vector field obtained by putting $z = f(x, y)$ and so eliminating the explicit dependence on z . Now, we may use Green's theorem in the plane to obtain

$$\int_{\partial D} G_1 dx + G_2 dy = \iint_D \left(\frac{\partial G_2}{\partial x} - \frac{\partial G_1}{\partial y} \right) dy dx.$$

The calculation of the partials on the right is a little tricky. By the *chain rule*

$$\begin{aligned}\frac{\partial}{\partial x}F_2(x, y, f(x, y)) &= \frac{\partial F_2}{\partial x} \frac{\partial x}{\partial x} + \frac{\partial F_2}{\partial y} \frac{\partial y}{\partial x} + \frac{\partial F_2}{\partial z} \frac{\partial z}{\partial x} \\ &= \frac{\partial F_2}{\partial x}(1) + \frac{\partial F_2}{\partial y}(0) + \frac{\partial F_2}{\partial z} f_x \\ &= \frac{\partial F_2}{\partial x} + \frac{\partial F_2}{\partial z} f_x.\end{aligned}$$

The notation is a little confusing. On the left, we are taking the partial derivative with respect to x *after* making the substitution $z = f(x, y)$. The function being differentiated is thus a function of x and y alone. On the right, the partial derivatives are taken *before* making the substitution, so at that stage x, y and z are treated as independent variables. Similar calculations yield

$$\begin{aligned}\frac{\partial}{\partial x}F_3(x, y, f(x, y)) &= \frac{\partial F_3}{\partial x} + \frac{\partial F_3}{\partial z} f_x \\ \frac{\partial}{\partial y}F_1(x, y, f(x, y)) &= \frac{\partial F_1}{\partial y} + \frac{\partial F_1}{\partial z} f_y \\ \frac{\partial}{\partial y}F_3(x, y, f(x, y)) &= \frac{\partial F_3}{\partial y} + \frac{\partial F_3}{\partial z} f_y.\end{aligned}$$

Thus,

$$\begin{aligned}\frac{\partial G_2}{\partial x} &= \frac{\partial F_2}{\partial x} + \frac{\partial F_2}{\partial z} f_x + \left(\frac{\partial F_3}{\partial x} + \frac{\partial F_3}{\partial z} f_x \right) f_y + F_3 f_{yx} \\ \frac{\partial G_1}{\partial y} &= \frac{\partial F_1}{\partial y} + \frac{\partial F_1}{\partial z} f_y + \left(\frac{\partial F_3}{\partial y} + \frac{\partial F_3}{\partial z} f_y \right) f_x + F_3 f_{xy}.\end{aligned}$$

(The product rule has been used for the second terms in the expressions for G_1 and G_2 .) Hence, subtracting, we obtain for the integrand

$$\frac{\partial G_2}{\partial x} - \frac{\partial G_1}{\partial y} = \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} + \left(\frac{\partial F_2}{\partial z} - \frac{\partial F_3}{\partial y} \right) f_x + \left(\frac{\partial F_3}{\partial x} - \frac{\partial F_1}{\partial z} \right) f_y.$$

Next we evaluate $\iint_S \mathbf{F} \cdot d\mathbf{S}$. We have

$$d\mathbf{S} = \langle -f_x, -f_y, 1 \rangle dy dx.$$

Also,

$$\nabla \times \mathbf{F} = \left\langle \frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}, \frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}, \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right\rangle$$

so

$$(\nabla \times \mathbf{F}) \cdot d\mathbf{S} = \left[\left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z} \right) (-f_x) + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x} \right) (-f_y) + \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right] dy dx,$$

and it is not hard to check that the expression in brackets is the same integrand as above. It follows that

$$\int_{\partial S} \mathbf{F} \cdot d\mathbf{r} = \int_{\partial D} G_1 dx + G_2 dy = \iint_D \frac{\partial G_2}{\partial x} - \frac{\partial G_1}{\partial y} dy dx = \iint_S \nabla \times \mathbf{F} \cdot d\mathbf{S}.$$

That completes the proof.

Exercises for 5.10.

1. Consider the surface consisting of the three faces of the unit cube in the first octant which lie in the coordinate planes. Assume that the normals are chosen to point into the first octant. Determine the orientation of the boundary of each of the three faces and verify that cancellation occurs as described in the proof of Stokes's Theorem along common edges.

5.11 Conservative Fields, Reconsidered

Let \mathbf{F} denote a vector field in \mathbf{R}^n where $n = 2$ or $n = 3$. We shall look again at the screening tests to determine whether \mathbf{F} might be conservative. For a vector field in the plane ($n = 2$), the test is

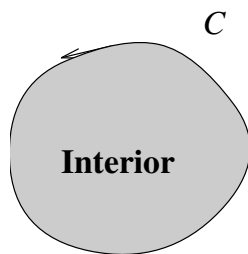
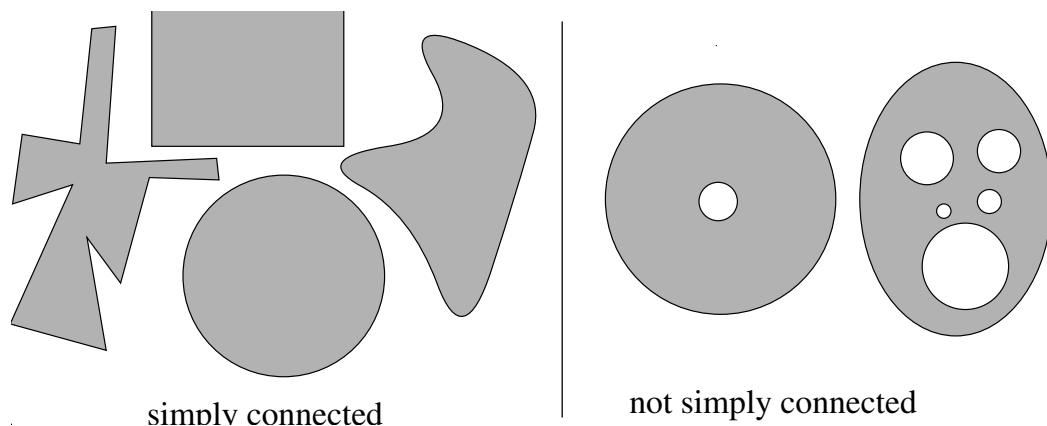
$$\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} = 0$$

(equation (2) in Section 3), and for a vector field in space ($n = 3$), the test may be written

$$\nabla \times \mathbf{F} = \mathbf{0}$$

(Section 4). We pointed out in our earlier discussion that a vector field could pass the screening test but still not be conservative. The most important example of such a field is $\mathbf{F} = \frac{1}{r}\mathbf{u}_\theta$. However, this can only happen if the geometry of the *domain* of the vector field is sufficiently complicated. In this section we shall look into the matter in more detail.

First consider the plane case ($n = 2$). We said previously that an open set D is connected if it can't be decomposed into disjoint open sets, i.e., if it comes in 'one piece'. There is a somewhat related but much subtler notion. If a connected region D in \mathbf{R}^2 does not have any 'holes' in it, it is called *simply connected*. The diagram gives some examples of regions in the plane which are simply connected and which are not simply connected.



A slightly more rigorous characterization of ‘simply connected’ is as follows. Consider a simple closed curve \mathcal{C} in the plane. By that we mean that the curve does not cross itself anywhere. It is intuitively clear that such a curve bounds a region in the plane which we call the *interior* of the curve. The region D is *simply connected* if for any simple closed curve \mathcal{C} which lies entirely in D , the interior of \mathcal{C} also lies in D . The idea is that if there were a ‘hole’ in D , even one consisting of a single missing point, then one could find a curve which goes around the hole and then the interior of that curve would contain at least one point not in D . (The actual definition of the term ‘simply connected’ is a bit more involved, but *for regions in the plane*, the above characterization is equivalent.)

Note that many common regions in the plane are simply connected. For example, all rectangles, disks, etc. are simply connected.

Theorem 5.8 Let $\mathbf{F} = \langle F_1, F_2 \rangle$ be a plane vector field which is smooth on its domain of definition D . Suppose D is simply connected. Then \mathbf{F} is conservative if and only if it passes the screening test

$$\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} = 0.$$

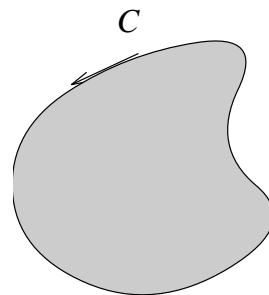
Proof. If \mathbf{F} is conservative, then we know it passes the screening test.

Suppose conversely that $\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} = 0$. One way to show that \mathbf{F} is conservative is to show that $\int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r} = 0$ for every closed curve \mathcal{C} in the domain D . In principle, we must show this for every closed curve, but in fact it is enough to show it for every simple closed curve, i.e., every closed curve which does not cross itself. Suppose then that \mathcal{C} is a *simple* closed curve, and let D' denote the interior of \mathcal{C} . By hypothesis, D' is entirely contained within D where the vector field \mathbf{F} is assumed to be smooth.

Hence, Green's theorem applies and we may conclude that

$$\begin{aligned}\int_C \mathbf{F} \cdot d\mathbf{r} &= \iint_{D'} \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} dA \\ &= \iint_{D'} (0) dA = 0\end{aligned}$$

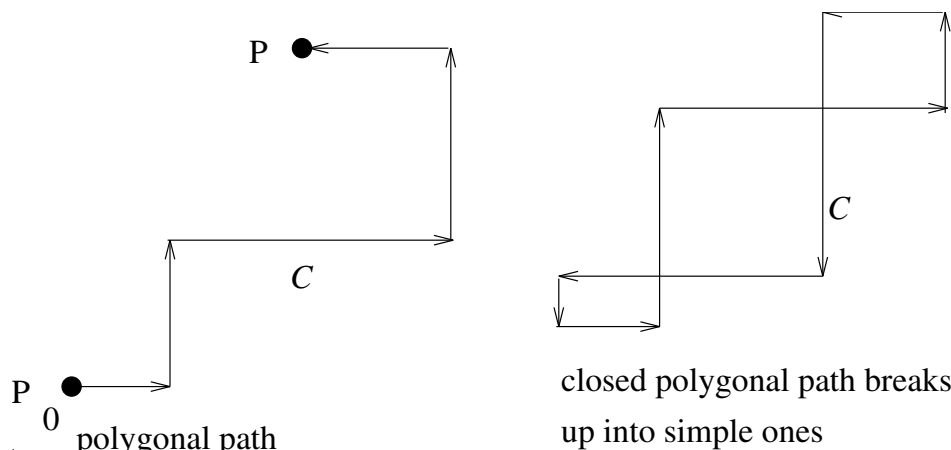
as required. \square



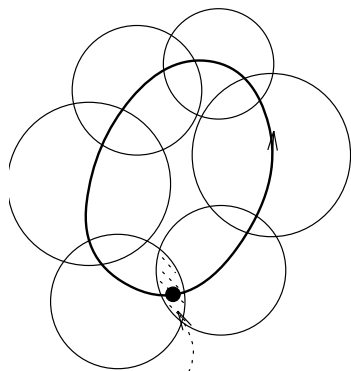
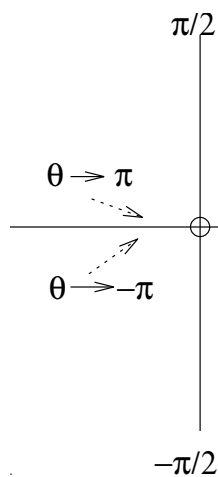
Remarks for those curious about the details. There are some tricky points which were glossed over in the above discussion. First, the assertion that every simple closed curve bounds a region in the plane is actually a deep theorem called the *Jordan Curve Theorem*. That theorem is quite difficult to prove in full generality. Fortunately, there are tricky ways to get around that for what we want here. Thus, to show \mathbf{F} is conservative, it suffices to choose a base point \mathbf{r}_0 and to define a function f with gradient \mathbf{F} using the formula

$$f(\mathbf{r}) = \int_C \mathbf{F} \cdot d\mathbf{r}$$

where C is any path from \mathbf{r}_0 to \mathbf{r} . Since this can be any path, it might as well be a polygonal path with edges parallel to one of the coordinate axes. The Jordan Curve theorem is much easier to verify for such paths. Similarly, the reduction from curves which do cross themselves to curves which do not is not hard to justify for such polygonal paths. See the diagram for a hint about how to do it.



Example Let $\mathbf{F}(x, y) = \frac{1}{r} \mathbf{u}_\theta = \frac{1}{x^2 + y^2} \langle -y, x \rangle$. The domain of this function is the plane \mathbf{R}^2 with the origin deleted. Hence, it is not simply connected. Thus the theorem does not apply. Indeed, the field is not conservative but does pass

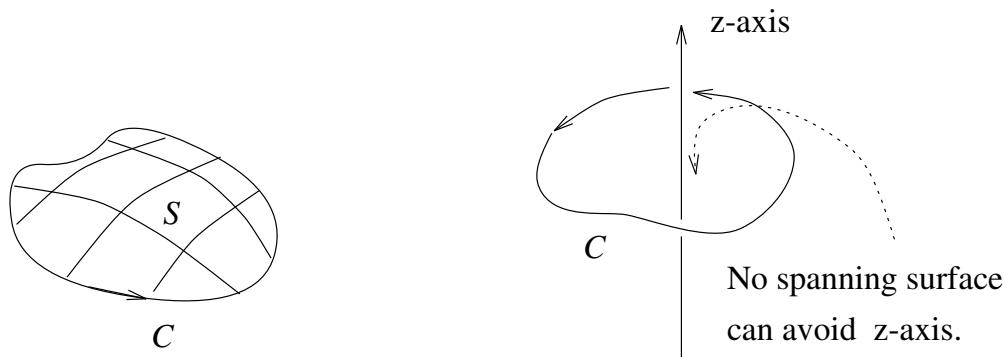


Solutions obtained in small neighborhoods may conflict after traversing a closed loop.

the screening test. However, we may choose simply connected subdomains and consider the vector field on such a subdomain. For example, let D be the part of the plane to the right of the y -axis and not including the y -axis. This region is simply connected—i.e., there are no holes—and \mathbf{F} does pass the screening test, so the theorem tells us there must be a function f such that $\mathbf{F} = \nabla f$ everywhere on D . In fact, you showed previously in an exercise that the function defined by $f(x, y) = \tan^{-1}(y/x)$ is just such a function. (Check again that $\nabla f = \mathbf{F}$ in case you don't remember the exercise.) The natural question then—posed in the exercise—is why can't we use this same function for the original domain of \mathbf{F} ? Let's see what happens if we try. Note first that in the right half plane, we have $\theta = \tan^{-1}(y/x)$ so there is an obvious way to try to extend the definition of the function f . Let $f(x, y)$ be the polar coordinate θ of the point (x, y) . Unfortunately, θ is not uniquely defined. If we start working forward from the positive y -axis, θ starts at $\pi/2$ and increases. If we start working backward from the negative y -axis, θ starts at $-\pi/2$ and gets progressively more negative. On the negative x -axis, these two attempts to extend the definition of $f(x, y)$ will run into a problem. If we approach from above the proper value will be π , but if we approach from below the proper value will be $-\pi$. *There is no way around this difficulty.* We know that because the field is not conservative. Hence, there is no continuous function f such that $\mathbf{F} = \nabla f$ on the entire domain of \mathbf{F} .

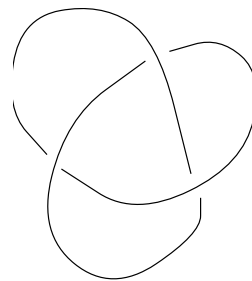
The above example illustrates a very important principle in mathematics and its applications. We may have a problem (for example a system of differential equations) which, for every point where the problem makes sense, has a solution which works in a sufficiently small neighborhood of that point. In that case, we say that the problem is solvable *locally*. That does not guarantee, however, that the problem can be solved *globally*, i.e., that there is a single continuous solution which works everywhere. Since every point has a simply connected neighborhood (a small rectangle or small disk), any vector field which passes the screening test is locally a gradient, but may not be globally a gradient. The issue of 'local solutions' versus 'global solution' has only recently entered the awareness of physicists and other scientists. The reason is that much of physics was concerned with solving differential equations or systems of differential equations, and the solution methods tend to give local solutions. However, the importance of global solutions has become increasingly clear in recent years. The above ideas can be extended to vector fields

in space, but the geometry is more complicated because the concept 'hole' can be more complicated. Let D be a connected open set in \mathbf{R}^3 . Suppose that any simple closed curve C in D is the boundary of an orientable surface S which also lies entirely in D . In this case we shall say that D *spans curves*. Most common regions, e.g., solid spheres or rectangular boxes, have this property. The region consisting of \mathbf{R}^3 with a single point deleted also has this property. For, if a curve C goes around the missing point, it can be spanned by a surface which avoids the point. On the other hand, the region obtained by deleting the entire z -axis from \mathbf{R}^3 does not have the property since any curve going around the z -axis cannot be spanned by a surface which is not pierced by the z -axis.



The analogue of the above theorem holds in space: *If \mathbf{F} is a smooth vector field on some open set D in \mathbf{R}^3 and D spans curves, then \mathbf{F} is conservative if and only if it satisfies the screening test $\nabla \times \mathbf{F} = \mathbf{0}$.* The proof is very similar, but it uses Stokes's theorem instead of Green's theorem. You should work it out for yourself. This notion is a bit more complicated than you might think at first. A curve in

\mathbf{R}^3 might not intersect itself but could still have quite complicated geometry. For example, it might be knotted in such a way that it can not be straightened out without being cut. Also, we didn't go into how complicated the 'surfaces' spanning such a curve may be.



The term 'simply connected' may also be defined for regions E in space, but the definition is not the same as that for ' E spans curves' given above. However, the notions are fairly closely related, and a region which is simply connected does span curves (but not necessarily vice versa).

Exercises for 5.11.

- Tell if each of the following regions is simply connected. If not, give a reason for your conclusion
 - The set of all points in \mathbf{R}^2 except for the semi-circle $x^2 + y^2 = 1, y > 0$.
 - The set of all points in \mathbf{R}^2 except for $(1, 0)$ and $(-1, 0)$.
 - The region in \mathbf{R}^2 between $x^2 + y^2 = 1$ and $x^2 + y^2 = 4$.
- Tell if each of the following regions in \mathbf{R}^3 spans curves.
 - The set of all points in \mathbf{R}^3 except for the line segment from $(0, 0, 0)$ to $(1, 0, 0)$.
 - The interior of the torus in \mathbf{R}^3 obtained by rotating the circle $(x-2)^2 + y^2 = 1$ about the z -axis.

(c) The region in \mathbf{R}^3 between the sphere $x^2 + y^2 + z^2 = 1$ and the sphere $x^2 + y^2 + z^2 = 4$.

(d) The set of all points in \mathbf{R}^3 except for those on the circle $x^2 + y^2 = 1, z = 0$ in the x, y -plane.

3. You showed previously by a messy direct calculation that $\nabla \times (\nabla f) = \mathbf{0}$ for any scalar field f in space. Here is an alternate argument with avoids the messy calculation. (However, it depends on Stokes's theorem, the proof of which is messy enough in its own right!)

Use the 'physical interpretation of the curl' to show that $\nabla \times \mathbf{F} \cdot \mathbf{N} = 0$ for any conservative field $\mathbf{F} = \nabla f$ and any vector \mathbf{N} . (Use the path independence property for conservative fields.) Then conclude that $\nabla \times \mathbf{F} = \mathbf{0}$.

5.12 Vector Potential

Let \mathbf{F} be a vector field in \mathbf{R}^3 . We said that \mathbf{F} is conservative if $\mathbf{F} = \nabla f$ for some scalar field f . Electrostatic fields are conservative, so the previous mathematics is good to know if you are studying such fields. Magnetic fields, however, are not conservative, so a different but related kind of mathematics is appropriate for them. We go into that now.

Let \mathbf{F} be a vector field in space. Another vector field \mathbf{A} is called a *vector potential* for \mathbf{F} if

$$\nabla \times \mathbf{A} = \mathbf{F}.$$

Note that such an \mathbf{A} is not unique. Namely, if \mathbf{U} is any field satisfying $\nabla \times \mathbf{U} = \mathbf{0}$, then

$$\nabla \times (\mathbf{A} + \mathbf{U}) = \nabla \times \mathbf{A} + \nabla \times \mathbf{U} = \mathbf{F} + \mathbf{0} = \mathbf{F}.$$

Hence, a vector potential for \mathbf{F} , if there is one, may always be modified by such a \mathbf{U} . Since any conservative \mathbf{U} satisfies $\nabla \times \mathbf{U} = \mathbf{0}$, there is a very wide choice of possible vector potentials. This is to be distinguished from the case of scalar potentials which are unique up to an additive constant.

There is also a screening test to determine if \mathbf{F} might have a vector potential \mathbf{A} . Namely, you should have checked the identity

$$\nabla \cdot (\nabla \times \mathbf{A}) = 0,$$

i.e., *divergence following curl is always 0*. (If you didn't do that exercise, do it now!) That means, if $\nabla \cdot \mathbf{F} \neq 0$, then there is no point looking for a vector potential. If it does vanish, then it is worth a try.

Example 119 Let $\mathbf{F}(x, y, z) = \langle x, y, -2z \rangle$. We have

$$\nabla \cdot \mathbf{F} = 1 + 1 - 2 = 0$$

so \mathbf{F} passes the screening test. We look for a vector potential \mathbf{A} as follows. We try to find one with $A_3 = 0$. This is plausible for the following reason. Suppose we found a vector potential with $A_3 \neq 0$. In that case, we could certainly find a scalar field f such that $A_3 = \partial f / \partial z$. Then,

$$\nabla \times (\mathbf{A} - \nabla f) = \nabla \times \mathbf{A} - \nabla \times (\nabla f) = \nabla \times \mathbf{A}.$$

However, the third component of $\mathbf{A} - \nabla f$ is $A_3 - \partial f / \partial z = 0$.

Assume now that $\nabla \times \mathbf{A} = \mathbf{F}$ where $A_3 = 0$. Then

$$-\frac{\partial A_2}{\partial z} = x, \quad \frac{\partial A_1}{\partial z} = y, \quad \frac{\partial A_2}{\partial x} - \frac{\partial A_1}{\partial y} = -2z.$$

Thus, from the first two equations

$$\begin{aligned} A_2 &= -xz + C(x, y) \\ A_1 &= yz + D(x, y). \end{aligned}$$

These can be substituted in the third equation to obtain

$$-z + \frac{\partial C}{\partial x} - z - \frac{\partial D}{\partial y} = -2z$$

or

$$\frac{\partial C}{\partial x} - \frac{\partial D}{\partial y} = 0.$$

There are no unique solutions C, D to this equation. That reflects the great freedom we have in choosing a vector potential. There are a variety of methods one could use to come up with explicit solutions, but in the present case, it is clear by inspection that

$$C(x, y) = D(x, y) = 0$$

will work. Hence, $A_1 = yz, A_2 = -xz, A_3 = 0$ is a solution and

$$\mathbf{A} = \langle yz, -xz, 0 \rangle$$

is a vector potential. You should check that it works.

Not every field \mathbf{F} which passes the screening test $\nabla \cdot \mathbf{F} = 0$ has a vector potential.

Example 120 Let $\mathbf{F} = \frac{1}{\rho^2} \mathbf{u}_\rho$ be our friend the inverse square law. We have remarked several times that $\nabla \cdot \mathbf{F} = 0$ for an inverse square law, so \mathbf{F} passes the screening test. However, $\mathbf{F} \neq \nabla \times \mathbf{A}$ for any possible \mathbf{A} . Namely, we know from flux calculations made previously that for this inverse square law

$$\iint_S \mathbf{F} \cdot d\mathbf{S} = 4\pi$$

for any sphere \mathcal{S} centered at the origin. On the other hand, it is not too hard to see that

$$\int_{\mathcal{S}} \nabla \times \mathbf{A} \cdot d\mathbf{S} = 0$$

for any vector field \mathbf{A} . To see this, divide the sphere into an upper hemisphere \mathcal{S}_1 and a lower hemisphere \mathcal{S}_2 which meet at the equator, which is a common boundary for both. Note that $\partial\mathcal{S}_1$ is traversed in the *opposite direction* from $\partial\mathcal{S}_2$, although they are the same curve if orientation is ignored. By Stokes's theorem, we have

$$\begin{aligned} \int_{\partial\mathcal{S}_1} \mathbf{A} \cdot d\mathbf{r} &= \iint_{\mathcal{S}_1} \nabla \times \mathbf{A} \cdot d\mathbf{S}, \\ \int_{\partial\mathcal{S}_2} \mathbf{A} \cdot d\mathbf{r} &= \iint_{\mathcal{S}_2} \nabla \times \mathbf{A} \cdot d\mathbf{S}. \end{aligned}$$

If we add these two equations, we get zero on the left because the line integrals are for the same curve traced in opposite directions. On the right, we get

$$\iint_{\mathcal{S}_1} \nabla \times \mathbf{A} \cdot d\mathbf{S} + \iint_{\mathcal{S}_2} \nabla \times \mathbf{A} \cdot d\mathbf{S} = \iint_{\mathcal{S}} \nabla \times \mathbf{A} \cdot d\mathbf{S}$$

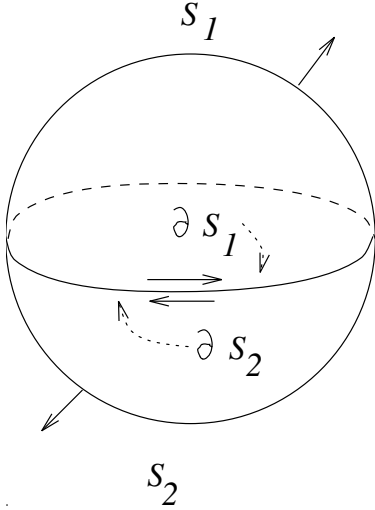
so it is zero.

The property analogous to ' E spans curves' is E *spans surfaces*. A connected open set E in \mathbf{R}^3 has this property if, for every simple closed surface \mathcal{S} in E , the interior of \mathcal{S} is also in E . The set E consisting of \mathbf{R}^3 with the origin deleted does not span surfaces since the interior of any sphere centered at the origin is not entirely in E . On the other hand, the set E consisting of \mathbf{R}^3 with the entire z -axis deleted does span surfaces. Do you see why?

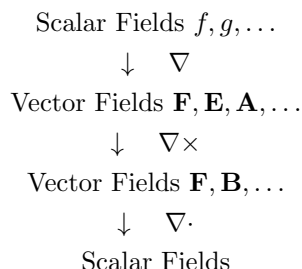
Theorem 5.9 Let \mathbf{F} be a smooth vector field on an open set E in \mathbf{R}^3 which spans surfaces. Then \mathbf{F} has a vector potential if and only if $\nabla \cdot \mathbf{F} = 0$.

The proof of this theorem is much too difficult to attempt here. *Warning!* The author has not been completely honest about the concept ' E spans surfaces' because the phrase 'simple closed surface' was not defined. This phrase evokes the idea of something straightforward like the surface of a sphere or a cube. However, it is necessary for the above theorem to hold that we consider much more general surfaces. For example, a toral surface, i.e., the surface of a doughnut shaped solid, is an example of a 'simple closed surface'. In general, any surface which may be viewed as the boundary of a plausible solid region (and to which the divergence theorem may be applied) is a possible candidate. (If you want to learn more about this subject, you should plan someday to study *algebraic topology* and in particular a famous result called *DeRham's Theorem*.) Probably the only cases you ever need to apply the theorem to are where E is equal to all of \mathbf{R}^3 or at worst the interior of a *really* simple closed surface like a sphere or a cube.

Summary of Relations among Fields in \mathbf{R}^3 Some the relations among scalar and vector fields in \mathbf{R}^3 described in the previous sections can be summarized in the



following table.



The way to interpret this table is as follows. The vertical arrows indicate operations. The first, the gradient operation (∇), takes scalar fields into vector fields. The second, the curl operation ($\nabla \times$), takes vector fields into vector fields. The third, the divergence operation ($\nabla \cdot$), takes vector fields into scalar fields.

The result of doing two successive operations is zero. Thus, $\nabla \times (\nabla f) = \mathbf{0}$ and $\nabla \cdot (\nabla \times \mathbf{A}) = 0$. Moreover, at the second and third levels we have screening tests. If a vector field at the given level passes the test (i.e., the operation at that level takes it to zero), one may ask if the vector field comes from a ‘potential’ at the previous level. Under suitable connectedness assumptions, the answer will be yes.

Exercises for 5.12.

1. In each case check if the given field has divergence zero. If so, look for a vector potential. (In each case, the domain is all of \mathbf{R}^3 which spans surfaces, so there must be a vector potential if the divergence is zero.)

(a) $\langle x^2, y^2, -2(x+y)z \rangle$.

(b) $\langle y, x, z \rangle$

(c) $\langle z, x, y \rangle$.

2. Let

$$\mathbf{F}(x, y, z) = \left\langle \frac{x}{x^2 + y^2}, \frac{y}{x^2 + y^2}, 0 \right\rangle = \frac{1}{r} \mathbf{u}_r.$$

(a) Show that $\nabla \cdot \mathbf{F} = 0$ except on the z -axis where \mathbf{F} is undefined. Hint: You can save yourself some time by referring to the formula for divergence in cylindrical coordinates in Section 13.

(b) Find a vector potential for \mathbf{F} . (The domain of \mathbf{F} *does* span surfaces. Can you see that?)

3. You showed in a previous exercise by a messy direct calculation that $\nabla \cdot (\nabla \times \mathbf{A}) = 0$ for any vector field \mathbf{A} in space. Derive this fact by the following

conceptual argument which relies on both the divergence theorem and Stokes's theorem. (However, recall that the proofs of those theorems are not so easy.) Use the argument in Example 120, to show that $\iint_{\mathcal{S}} (\nabla \times \mathbf{A}) \cdot d\mathbf{S} = 0$ for any spherical surface in the domain of \mathbf{A} . Then use the 'physical interpretation of the divergence' to conclude that $\nabla \cdot (\nabla \times \mathbf{A}) = 0$.

4. Let $\mathbf{F} = \langle F_1, F_2, F_3 \rangle$ be a vector field in space satisfying $\nabla \cdot \mathbf{F} = 0$. Suppose the domain of \mathbf{F} is all of \mathbf{R}^3 . Fix values y_0, z_0 , and define

$$\begin{aligned} A_1 &= \int_{z_0}^z F_2(x, y, z') dz' - \int_{y_0}^y F_3(x, y', z_0) dy', \\ A_2 &= - \int_{z_0}^z F_1(x, y, z') dz', \\ A_3 &= 0. \end{aligned}$$

Show that $\nabla \times \mathbf{A} = \mathbf{F}$. Hint: You will have to use the general rule for differentiating integrals

$$\frac{\partial}{\partial t} \int_a^b f(s, t, \dots) ds = \int_a^b \frac{\partial f}{\partial t}(s, t, \dots) ds$$

which asserts that you can interchange differentiation and integration with respect to *different* variables. This rule applies for functions with continuous partials on a bounded interval. Note however that to differentiate with respect to the upper limit of an integral, you need to use the fundamental theorem of calculus.

This is a special case of the theorem stated in the section since \mathbf{R}^3 certainly spans surfaces.

5.13 Vector Operators in Curvilinear Coordinates

This section is included here because it depends on what we have just done. However, most of you should probably skip it for now and come back to it when you need the formulas given here, which is likely to be the case at some point in your studies. You should glance at some of the formulas, for future reference. Of course, if you really want a challenging test of your understanding of the material in the previous sections, you should study this section in detail.

Gradient in Cylindrical and Spherical Coordinates For a scalar field f in the plane, we found in Chapter III, Section 7 that

$$\nabla f = \mathbf{u}_r \frac{\partial f}{\partial r} + \mathbf{u}_\theta \frac{1}{r} \frac{\partial f}{\partial \theta}. \quad (72)$$

where on the right g is the function of r, θ obtained by substituting $x = r \cos \theta, y = r \sin \theta$ in f , i.e., $f(x, y) = g(r, \theta)$.

The corresponding formula in cylindrical coordinates is

$$\nabla f = \mathbf{u}_r \frac{\partial g}{\partial r} + \mathbf{u}_\theta \frac{1}{r} \frac{\partial g}{\partial \theta} + \mathbf{k} \frac{\partial g}{\partial z} \quad (73)$$

where $f(x, y, z) = g(r, \theta, z)$. That is fairly clear because the change from rectangular coordinates x, y, z to cylindrical coordinates r, θ, z involves only the first two coordinates.

There is a corresponding formula for spherical coordinates.

$$\nabla f = \mathbf{u}_\rho \frac{\partial g}{\partial \rho} + \mathbf{u}_\phi \frac{1}{\rho} \frac{\partial g}{\partial \phi} + \mathbf{u}_\theta \frac{1}{\rho \sin \phi} \frac{\partial g}{\partial \theta}. \quad (74)$$

where $f(x, y, z) = g(\rho, \phi, \theta)$. $\mathbf{u}_\rho, \mathbf{u}_\phi$, and \mathbf{u}_θ are unit vectors in the ρ, ϕ , and θ directions respectively. At any point in space, \mathbf{u}_ρ points directly away from the origin, \mathbf{u}_ϕ is tangent to the circle of longitude through the point, and points from north to south, and \mathbf{u}_θ is tangent to the circle of latitude through the point and points from west to east.

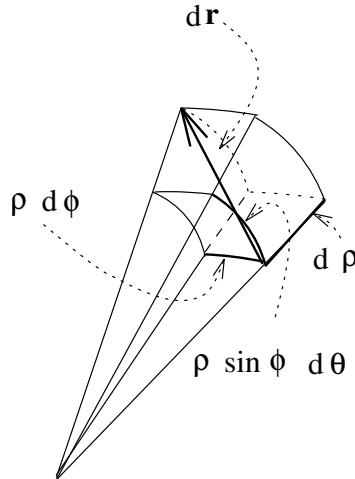
To derive this formula, we argue as follows.

$$df = \nabla f \cdot d\mathbf{r}.$$

Assume

$$\nabla f = \mathbf{u}_\rho A_\rho + \mathbf{u}_\phi A_\phi + \mathbf{u}_\theta A_\theta.$$

The trick is to express $d\mathbf{r}$ in terms of these same unit vectors. Consider a spherical cell with one corner at the point with spherical coordinates (ρ, ϕ, θ) and dimensions $d\rho, \rho d\phi$, and $\rho \sin \phi d\theta$.



Then ignoring small errors due to the curvature of the sides, we have

$$d\mathbf{r} = \mathbf{u}_\rho d\rho + \mathbf{u}_\phi \rho d\phi + \mathbf{u}_\theta \rho \sin \phi d\theta.$$

Hence,

$$\nabla f \cdot d\mathbf{r} = A_\rho d\rho + A_\phi \rho d\phi + A_\theta \rho \sin \phi d\theta.$$

However,

$$dg = \frac{\partial g}{\partial \rho} d\rho + \frac{\partial g}{\partial \phi} d\phi + \frac{\partial g}{\partial \theta} d\theta$$

so putting $df = dg$ and comparing coefficients of $d\rho, d\phi$, and $d\theta$ gives $A_\rho = \frac{\partial g}{\partial \rho}$, $A_\phi = \frac{1}{\rho} \frac{\partial g}{\partial \phi}$, and $A_\theta = \frac{1}{\rho \sin \phi} \frac{\partial g}{\partial \theta}$.

All this can be summarized by saying that the gradient operation has the following form in each coordinate system:

Polar Coordinates in the plane:

$$\nabla = \mathbf{u}_r \frac{\partial}{\partial r} + \mathbf{u}_\theta \frac{1}{r} \frac{\partial}{\partial \theta}.$$

Cylindrical Coordinates in space:

$$\nabla = \mathbf{u}_r \frac{\partial}{\partial r} + \mathbf{u}_\theta \frac{1}{r} \frac{\partial}{\partial \theta} + \mathbf{k} \frac{\partial}{\partial z}.$$

Spherical Coordinates in space:

$$\nabla = \mathbf{u}_\rho \frac{\partial}{\partial \rho} + \mathbf{u}_\phi \frac{1}{\rho} \frac{\partial}{\partial \phi} + \mathbf{u}_\theta \frac{1}{\rho \sin \phi} \frac{\partial}{\partial \theta}.$$

Divergence in Cylindrical Coordinates Let $\mathbf{F} = \mathbf{u}_r F_r + \mathbf{u}_\theta F_\theta + \mathbf{k} F_z$ be the representation of the vector field in terms of cylindrical coordinates. Then we have

$$\nabla \cdot \mathbf{F} = \left(\mathbf{u}_r \frac{\partial}{\partial r} + \mathbf{u}_\theta \frac{1}{r} \frac{\partial}{\partial \theta} + \mathbf{k} \frac{\partial}{\partial z} \right) \cdot (\mathbf{u}_r F_r + \mathbf{u}_\theta F_\theta + \mathbf{k} F_z).$$

One may calculate this *formally* using the fact that the three vectors $\mathbf{u}_r, \mathbf{u}_\theta$, and \mathbf{k} are mutually perpendicular, but one must be careful to use the product rule and apply the differentiation operators to these unit vectors. We have

$$\begin{array}{lll} \frac{\partial}{\partial r} \mathbf{u}_r = \mathbf{0} & \frac{\partial}{\partial \theta} \mathbf{u}_r = \mathbf{u}_\theta & \frac{\partial}{\partial z} \mathbf{u}_r = \mathbf{0} \\ \frac{\partial}{\partial r} \mathbf{u}_\theta = \mathbf{0} & \frac{\partial}{\partial \theta} \mathbf{u}_\theta = -\mathbf{u}_r & \frac{\partial}{\partial z} \mathbf{u}_\theta = \mathbf{0} \\ \frac{\partial}{\partial r} \mathbf{k} = \mathbf{0} & \frac{\partial}{\partial \theta} \mathbf{k} = \mathbf{0} & \frac{\partial}{\partial z} \mathbf{k} = \mathbf{0}. \end{array}$$

(These rules can be derived by geometric visualization or by writing $\mathbf{u}_r = \cos\theta\mathbf{i} + \sin\theta\mathbf{j}$ and $\mathbf{u}_\theta = -\sin\theta\mathbf{i} + \cos\theta\mathbf{j}$ and taking the indicated partial derivatives.) The result of the calculation, which you should check, is

$$\begin{aligned}\nabla \cdot \mathbf{F} &= \frac{1}{r} \left(\frac{\partial(rF_r)}{\partial r} + \frac{\partial F_\theta}{\partial \theta} + \frac{\partial(rF_z)}{\partial z} \right) \\ &= \frac{1}{r} \frac{\partial(rF_r)}{\partial r} + \frac{1}{r} \frac{\partial F_\theta}{\partial \theta} + \frac{\partial F_z}{\partial z}.\end{aligned}$$

The above calculation is done by following a plausible but arbitrary formal scheme. Hence, there is no particular reason to believe that it gives the correct answer. There are, after all, other formal schemes that one could employ. To show that the formula is correct, we need another argument, and we give that in what follows.

The basic method is to calculate the divergence as *flux per unit volume* by picking an appropriate family of curvilinear boxes which shrink to a point.

Consider a curvilinear box centered at the point with cylindrical coordinates (r, θ, z) . It has the following 6 faces.

The top and bottom faces are circular wedges centered at $(r, \theta, z + dz)$ and $(r, \theta, z - dz)$; their common area is $r(2d\theta)(2dr)$, and the normals are $\pm\mathbf{k}$.

The far and near side faces are rectangles centered at $(r, \theta + d\theta, z)$ and $(r, \theta - d\theta, z)$; their common area is $(2dr)(2dz)$, and the normals are $\pm\mathbf{u}_\theta$.

The outer and inner faces are cylindrical rectangles centered at $(r + dr, \theta, z)$ and $(r - dr, \theta, z)$; their areas are respectively $(r + dr)(2d\theta)(2dz)$ and $(r - dr)(2d\theta)(2dz)$, and the normals are $\pm\mathbf{u}_r$.

To calculate the flux out of the surface S of the box we argue as follows. First, for any face only the component of \mathbf{F} perpendicular to that face is relevant: F_z for the top and bottom faces, F_θ for the side faces, and F_r for the outer and inner faces. Secondly, to a first approximation, the flux through a face equals the value of the relevant component *at the center of the face* multiplied by its area and a *sign* depending on the direction of the normal. (The reason why it suffices to use the value of the component at the center of the face, rather than attempting to integrate over the face, is that to a first approximation we may assume the component is a *linear* function of the coordinates so, if we did integrate, each positive variation from the value at the center would be canceled by a corresponding negative variation.)

We now perform this calculation for the faces of the box.

For the top face the flux is

$$F_z(r, \theta, z + dz)(r2d\theta)(2dr).$$

However, to a first degree of approximation

$$F_z(r, \theta, z + dz) = F_z(r, \theta, z) + \frac{\partial F_z}{\partial z} dz$$

so the flux is

$$(F_z(r, \theta, z) + \frac{\partial F_z}{\partial z} dz)r(2d\theta)(2dr).$$

Similarly, the flux through the bottom face is

$$-(F_z(r, \theta, z) - \frac{\partial F_z}{\partial z} dz)r(2d\theta)(2dr).$$

(Note the normal in that case is $-\mathbf{k}$.) Hence, the total flux for the two faces is, after cancellation

$$(2\frac{\partial F_z}{\partial z} dz)r(2d\theta)(2dr) = \frac{\partial F_z}{\partial z}(2dz)r(2d\theta)(2dr) = \frac{\partial F_z}{\partial z} dV.$$

A comparable computation for the two side faces yields

$$\begin{aligned} (2\frac{\partial F_\theta}{\partial \theta} d\theta)(2dr)(2dz) &= \frac{\partial F_\theta}{\partial \theta}(2d\theta)(2dr)(2dz) \\ &= \frac{1}{r}(\frac{\partial F_\theta}{\partial \theta})(2rd\theta)(2dr)(2dz) = \frac{1}{r}(\frac{\partial F_\theta}{\partial \theta})dV. \end{aligned}$$

Note that we had to multiply (and divide) by the extra factor of r to change the θ increment to a distance.

The flux computation for the outer and inner faces is a bit different because the area as well as the component F_r is a function of the radial variable r . Thus for the outer face, the flux would be

$$F_r(r + dr, \theta, z)((r + dr)2d\theta)(2dz).$$

It is useful to rewrite this

$$((r + dr)F_r(r + dr, \theta, z))(2d\theta)(2dz)$$

and consider the quantity in parentheses as a function of r . Then making the linear approximation, the flux is

$$(rF_r(r, \theta, z) + \frac{\partial(rF_r)}{\partial r} dr)(2d\theta)(2dz)$$

and similarly for the inner face it is

$$-(rF_r(r, \theta, z) - \frac{\partial(rF_r)}{\partial r} dr)(2d\theta)(2dz).$$

Thus the net flux for the outer and inner faces is

$$(2\frac{\partial(rF_r)}{\partial r} dr)(2d\theta)(2dz) = \frac{1}{r}(\frac{\partial(rF_r)}{\partial r})(2dr)(r2d\theta)(2dz) = \frac{1}{r}(\frac{\partial(rF_r)}{\partial r})dV.$$

If we add up the three net fluxes, we get

$$\frac{1}{r} \left(\frac{\partial(rF_r)}{\partial r} \right) dV + \frac{1}{r} \left(\frac{\partial F_\theta}{\partial \theta} \right) dV + \left(\frac{\partial F_z}{\partial z} \right) dV = \left(\frac{1}{r} \frac{\partial(rF_r)}{\partial r} + \frac{1}{r} \left(\frac{\partial F_\theta}{\partial \theta} \right) + \frac{\partial F_z}{\partial z} \right) dV.$$

If we now divide by dV to get the *flux per unit volume* we get for the divergence

$$\nabla \cdot \mathbf{F} = \frac{1}{r} \frac{\partial(rF_r)}{\partial r} + \frac{1}{r} \left(\frac{\partial F_\theta}{\partial \theta} \right) + \frac{\partial F_z}{\partial z}$$

as required. *Remark.* The formula (74) for the divergence may be described in words

as follows. For each coordinate there is a multiplier which changes the coordinate to a distance. In this case the r and z multipliers are 1 but the θ multiplier is r . To obtain the divergence, multiply each component by the *other two multipliers* and then take the partial with respect to the relevant coordinate. Add up the results and then divide by the product of the multipliers. **Divergence in Spherical**

Coordinates Let $\mathbf{F} = F_\rho \mathbf{u}_\rho + F_\phi \mathbf{u}_\phi + F_\theta \mathbf{u}_\theta$ be a resolution of the vector field \mathbf{F} in terms of unit vectors appropriate for spherical coordinates. Then formally,

$$\nabla \cdot \mathbf{F} = (\mathbf{u}_\rho \frac{\partial}{\partial \rho} + \mathbf{u}_\phi \frac{1}{\rho} \frac{\partial}{\partial \phi} + \mathbf{u}_\theta \frac{1}{\rho \sin \phi} \frac{\partial}{\partial \theta}) \cdot (\mathbf{u}_\rho F_\rho + \mathbf{u}_\phi F_\phi + \mathbf{u}_\theta F_\theta).$$

Again this should be computed formally using the product rule and the rules

$$\begin{array}{lll} \frac{\partial}{\partial \rho} \mathbf{u}_\rho = \mathbf{0} & \frac{\partial}{\partial \phi} \mathbf{u}_\rho = \mathbf{u}_\phi & \frac{\partial}{\partial \theta} \mathbf{u}_\rho = \sin \phi \mathbf{u}_\theta \\ \frac{\partial}{\partial \rho} \mathbf{u}_\phi = \mathbf{0} & \frac{\partial}{\partial \phi} \mathbf{u}_\phi = -\mathbf{u}_\rho & \frac{\partial}{\partial \theta} \mathbf{u}_\phi = \cos \phi \mathbf{u}_\theta \\ \frac{\partial}{\partial \rho} \mathbf{u}_\theta = \mathbf{0} & \frac{\partial}{\partial \phi} \mathbf{u}_\theta = \mathbf{0} & \frac{\partial}{\partial \theta} \mathbf{u}_\theta = -\sin \phi \mathbf{u}_\rho - \cos \phi \mathbf{u}_\phi \end{array}$$

These formulas can be derived geometrically or by using

$$\begin{aligned} \mathbf{u}_\rho &= \langle \sin \phi \cos \theta, \sin \phi \sin \theta, \cos \phi \rangle \\ \mathbf{u}_\phi &= \langle \cos \phi \cos \theta, \cos \phi \sin \theta, -\sin \phi \rangle \\ \mathbf{u}_\theta &= \langle -\sin \theta, \cos \theta, 0 \rangle. \end{aligned}$$

The resulting formula for the divergence, which you should check, is

$$\nabla \cdot \mathbf{F} = \frac{1}{\rho^2 \sin \phi} \left(\frac{\partial(\rho^2 \sin \phi F_\rho)}{\partial \rho} + \frac{\partial(\rho \sin \phi F_\phi)}{\partial \phi} + \frac{\partial(\rho F_\theta)}{\partial \theta} \right).$$

Again, this must be verified by an independent argument. The reasoning for that is pretty much the same as in the case of cylindrical coordinates, but the curvilinear box is the one appropriate for spherical coordinates and hence somewhat more complicated. I leave it as a challenge for you to do the appropriate flux calculations.

The same rule works for interpreting this. Use multipliers 1 for ρ , ρ for ϕ , and $\rho \sin \phi$ for θ for the coordinates, and then multiply each component by the *other*

two multipliers and then take the partial with respect to the relevant coordinate. Add up the results and then divide by the product of the multipliers. **Curl in**

Cylindrical Coordinates We have

$$\nabla \times \mathbf{F} = (\mathbf{u}_r \frac{\partial}{\partial r} + \mathbf{u}_\theta \frac{1}{r} \frac{\partial}{\partial \theta} + \mathbf{k} \frac{\partial}{\partial z}) \times (\mathbf{u}_r F_r + \mathbf{u}_\theta F_\theta + \mathbf{k} F_z).$$

Again, this can be calculated formally using the appropriate rules, and the result is

$$\nabla \times \mathbf{F} = \frac{1}{r} \left(\frac{\partial(F_z)}{\partial \theta} - \frac{\partial(r F_\theta)}{\partial z} \right) \mathbf{u}_r + \left(\frac{\partial F_r}{\partial z} - \frac{\partial F_z}{\partial r} \right) \mathbf{u}_\theta + \frac{1}{r} \left(\frac{\partial(r F_\theta)}{\partial r} - \frac{\partial F_r}{\partial \theta} \right) \mathbf{k}.$$

This may be thought of as the determinant of the matrix

$$\begin{bmatrix} (1/r)\mathbf{u}_r & \mathbf{u}_\theta & (1/r)\mathbf{k} \\ \partial/\partial r & \partial/\partial \theta & \partial/\partial z \\ F_r & r F_\theta & F_z \end{bmatrix}.$$

Again, the above formulas have been derived purely formally, so one must justify them by another argument. To do this we calculate the components of the curl as *circulation per unit area* by picking an appropriate family of curvilinear rectangles which shrink to a point. We first calculate $\nabla \times \mathbf{F} \cdot \mathbf{u}_r$.

Consider the curvilinear rectangle which starts at the point with cylindrical coordinates $(r, \theta - d\theta, z - dz)$, goes to $(r, \theta + d\theta, z - dz)$, then to $(r, \theta + d\theta, z + dz)$, then to $(r, \theta - d\theta, z + dz)$, and then finally back to $(r, \theta - d\theta, z - dz)$. This curvilinear rectangle is traced on a cylinder of radius r and is centered at (r, θ, z) . Its dimensions are $2dz$ and $r(2d\theta)$. We calculate the circulation for this path which you should note has the proper orientation with respect to the outward normal \mathbf{u}_r to the cylinder.

On any side of this curvilinear rectangle we need only consider the component of \mathbf{F} parallel to that side: F_θ for the segments in the θ -direction and F_z for the segments in the z -direction.

To a first approximation, we may calculate the line integral $\int \mathbf{F} \cdot d\mathbf{r}$ for any side of the curvilinear rectangle by taking its value at the center of the side and multiplying by the length of the side. The reason this works is that to a first approximation there is as much positive variation on one side of the center point as there is negative variation on the other side, so the two cancel out.

This circulation is

$$\begin{aligned} & F_\theta(r, \theta, z - dz)r(2d\theta) + F_z(r, \theta + d\theta, z)(2dz) \\ & - F_\theta(r, \theta, z + dz)r(2d\theta) - F_z(r, \theta - d\theta, z)(2dz) \end{aligned}$$

We have the first order approximations

$$\begin{aligned} F_\theta(r, \theta, z - dz) &= F_\theta(r, \theta, z) - \frac{\partial F_\theta}{\partial z} dz, \\ F_\theta(r, \theta, z + dz) &= F_\theta(r, \theta, z) + \frac{\partial F_\theta}{\partial z} dz, \end{aligned}$$

and putting these in the first and third terms for the circulation yields the net result

$$-\frac{\partial F_\theta}{\partial z}(2dz)r(2d\theta) = -\frac{\partial F_\theta}{\partial z}dA.$$

Similarly, the first order approximations

$$\begin{aligned} F_z(r, \theta + d\theta, z) &= F_z(r, \theta, z) + \frac{\partial F_z}{\partial \theta} d\theta \\ F_z(r, \theta - d\theta, z) &= F_z(r, \theta, z) - \frac{\partial F_z}{\partial \theta} d\theta \end{aligned}$$

put in the second and fourth terms of the circulation yield the net result

$$\frac{\partial F_z}{\partial \theta}(2d\theta)(2dz) = \frac{1}{r} \frac{\partial F_z}{\partial \theta} r(2d\theta)(2dz) = \frac{1}{r} \frac{\partial F_z}{\partial \theta} dA.$$

Combining these terms yields

$$-\frac{\partial F_\theta}{\partial z}dA + \frac{1}{r} \frac{\partial F_z}{\partial \theta} dA = \left(\frac{1}{r} \frac{\partial F_z}{\partial \theta} - \frac{\partial F_\theta}{\partial z} \right) dA.$$

Now divide by the area dA of the curvilinear rectangle to obtain

$$\nabla \times \mathbf{F} \cdot \mathbf{u}_r = \frac{1}{r} \frac{\partial F_z}{\partial \theta} - \frac{\partial F_\theta}{\partial z} = \frac{1}{r} \left(\frac{\partial F_z}{\partial \theta} - \frac{\partial(rF_\theta)}{\partial z} \right).$$

The calculation of $\nabla \times \mathbf{F} \cdot \mathbf{u}_\theta$ is very similar. Use the rectangle centered at (r, θ, z) which starts at $(r - dr, \theta, z + dz)$, goes to $(r + dr, \theta, z + dz)$, then to $(r + dr, \theta, z - dz)$, then to $(r - dr, \theta, z - dz)$ and finally back to $(r - dr, \theta, z + dz)$. The net result is

$$\nabla \times \mathbf{F} \cdot \mathbf{u}_\theta = \frac{\partial F_r}{\partial z} - \frac{\partial F_z}{\partial r}.$$

The calculation of $\nabla \times \mathbf{F} \cdot \mathbf{k}$ is a bit more complicated. Consider the curvilinear ‘rectangle’ centered at (r, θ, z) which starts at $(r - dr, \theta - d\theta, z)$, goes to $(r + dr, \theta - d\theta, z)$, then to $(r + dr, \theta + d\theta, z)$, then to $(r - dr, \theta + d\theta, z)$ and finally back to $(r - dr, \theta - d\theta, z)$. Note that this is oriented properly with respect to the normal vector \mathbf{k} .

To a first approximation, the circulation $\int \mathbf{F} \cdot d\mathbf{r}$ is

$$\begin{aligned} &F_r(r, \theta - d\theta, z)(2dr) + F_\theta(r + dr, \theta, z)(r + dr)(2d\theta) \\ &- F_r(r, \theta + d\theta, z)(2dr) - F_\theta(r - dr, \theta, z)(r - dr)(2d\theta). \end{aligned}$$

We have the first order approximations

$$\begin{aligned} F_r(r, \theta - d\theta, z) &= F_r(r, \theta, z) - \frac{\partial F_r}{\partial \theta} d\theta, \\ F_r(r, \theta + d\theta, z) &= F_r(r, \theta, z) + \frac{\partial F_r}{\partial \theta} d\theta, \end{aligned}$$

and putting these in the first and third terms for the circulation yields the net result

$$-\frac{\partial F_r}{\partial \theta} 2d\theta (2dr) = -\frac{1}{r} \frac{\partial F_r}{\partial \theta} r(2d\theta)(2dr) = -\frac{1}{r} \frac{\partial F_r}{\partial \theta} dA.$$

For the second and fourth terms, the reasoning is a bit more complicated. Since both r and F_θ change, we also need to consider the variation of r . Put $H(r, \theta, z) = rF_\theta$. Then the second and fourth terms in the circulation become

$$H(r + dr, \theta, z)(2d\theta) - H(r - dr, \theta, z)(2d\theta).$$

The relevant first order approximations are

$$\begin{aligned} H(r + dr, \theta, z)(r + dr) &= rF_\theta(r, \theta, z) + \frac{\partial(rF_\theta)}{\partial r} dr, \\ H(r - dr, \theta, z)(r - dr) &= rF_\theta(r, \theta, z)r - \frac{\partial(rF_\theta)}{\partial r} dr. \end{aligned}$$

Put these in the second and fourth terms for the net result

$$\frac{\partial(rF_\theta)}{\partial r} 2dr(2d\theta) = \frac{1}{r} \frac{\partial(rF_\theta)}{\partial r} (2dr)r(2d\theta) = \frac{1}{r} \frac{\partial(rF_\theta)}{\partial r} dA.$$

Combining yields the following first order approximation for the circulation

$$-\frac{1}{r} \frac{\partial F_\theta}{\partial \theta} dA + \frac{1}{r} \frac{\partial(rF_\theta)}{\partial r} dA = \frac{1}{r} \left(\frac{\partial(rF_\theta)}{\partial r} - \frac{\partial F_r}{\partial \theta} \right) dA,$$

and dividing by dA gives

$$\nabla \times \mathbf{F} \cdot \mathbf{k} = \frac{1}{r} \left(\frac{\partial(rF_\theta)}{\partial r} - \frac{\partial F_r}{\partial \theta} \right).$$

We may summarize this information finally by writing

$$\nabla \times \mathbf{F} = \frac{1}{r} \left(\frac{\partial(F_z)}{\partial \theta} - \frac{\partial(rF_\theta)}{\partial z} \right) \mathbf{u}_r + \left(\frac{\partial F_r}{\partial z} - \frac{\partial F_z}{\partial r} \right) \mathbf{u}_\theta + \frac{1}{r} \left(\frac{\partial(rF_\theta)}{\partial r} - \frac{\partial F_r}{\partial \theta} \right) \mathbf{k}$$

as we claimed above. **Curl in Spherical Coordinates** Formally, we have

$$\nabla \times \mathbf{F} = (\mathbf{u}_\rho \frac{\partial}{\partial \rho} + \mathbf{u}_\phi \frac{1}{\rho} \frac{\partial}{\partial \phi} + \mathbf{u}_\theta \frac{1}{\rho \sin \phi} \frac{\partial}{\partial \theta}) \times (\mathbf{u}_\rho F_\rho + \mathbf{u}_\phi F_\phi + \mathbf{u}_\theta F_\theta).$$

The result of working this out is

$$\begin{aligned}\nabla \times \mathbf{F} = & \frac{1}{\rho^2 \sin \phi} \left(\frac{\partial(\rho \sin \phi F_\theta)}{\partial \phi} - \frac{\partial(\rho F_\phi)}{\partial \theta} \right) \mathbf{u}_\rho \\ & + \frac{1}{\rho \sin \phi} \left(\frac{\partial F_\rho}{\partial \theta} - \frac{\partial(\rho \sin \phi F_\theta)}{\partial \rho} \right) \mathbf{u}_\phi + \frac{1}{\rho} \left(\frac{\partial(\rho F_\phi)}{\partial \rho} - \frac{\partial F_\phi}{\partial \phi} \right) \mathbf{u}_\theta\end{aligned}$$

which may also be expressed as the determinant of the matrix

$$\begin{bmatrix} (1/(\rho^2 \sin \phi))\mathbf{u}_\rho & (1/(\rho \sin \phi))\mathbf{u}_\theta & (1/\rho)\mathbf{u}_\phi \\ \partial/\partial \rho & \partial/\partial \phi & \partial/\partial \theta \\ F_\rho & \rho F_\phi & \rho \sin \phi F_\theta \end{bmatrix}.$$

The analysis is pretty much the same as in the case for cylindrical coordinates. The justification uses curvilinear rectangles appropriate for spherical coordinates.

Exercises for 5.13.

1. Determine the form of the Laplacian operator $\nabla^2 = \nabla \cdot \nabla$ in polar coordinates.
2. Do the same for cylindrical coordinates.
3. Do the same for spherical coordinates.

Part II

Differential Equations

Chapter 6

First Order Differential Equations

6.1 Differential Forms

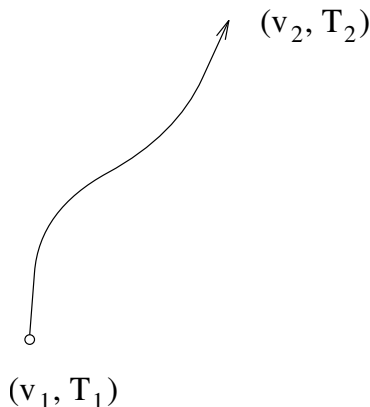
Differential forms are an alternate way to talk about vector fields, and for some applications they are the preferred way. They are used frequently in thermodynamics and also in the study of first order differential equations.

We shall concentrate on the case of forms in \mathbf{R}^2 . There is a corresponding theory for \mathbf{R}^n with $n > 2$, but it is much more difficult.

Let $\mathbf{F} = \langle F_1, F_2 \rangle$ be a plane vector field, and let \mathcal{C} be a path in the plane. Unless \mathbf{F} is conservative, the line integral $W = \int_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{r}$ is a path dependent quantity, so we might write symbolically $W = W(\mathcal{C})$. You may also have encountered path dependent quantities in your chemistry course when studying thermodynamics. Namely, if a gas satisfies an equation of state of the form $f(p, v, T) = 0$, it is possible to choose two of the three variables to be independent variables and (at least locally) solve for the third in terms of those. Suppose, for example, that v, T are the independent variables.

In this context, we can think of the path \mathcal{C} as representing a sequence of changes of state of the gas leading from the state characterized by (v_1, T_1) to the state characterized by (v_2, T_2) . Such a sequence of changes of state will yield a change in the *heat* q in the system, and this change will generally depend on the path \mathcal{C} . For ‘infinitesimal changes’ dv, dT , the *first law of thermodynamics* asserts that the *change in q* is given by

$$du - p dv$$



where $u = u(p, v, T)$ is a function of the state of the system called its *internal energy*. du can be expressed in terms of dp, dv and dT , and since dp can be expressed in terms of dv and dT , the above expression for the change in q can be put ultimately in the form

$$M(v, T)dv + N(v, T)dT.$$

Then the total change in the heat q along the path \mathcal{C} will be the line integral

$$\int_{\mathcal{C}} du - p dv = \int_{\mathcal{C}} M dv + N dT.$$

As mentioned above, it depends on the path \mathcal{C} .

Leaving the thermodynamics to your chemistry professors, let us consider the basic mathematical situation. Given a pair of functions $M(x, y), N(x, y)$, the expression

$$M(x, y) dx + N(x, y) dy$$

is called a *differential form*. We shall always assume that the component functions M and N are as smooth as we need them to be on the domain of the form. You can think of the form as giving the change of some quantity for a displacement $d\mathbf{r} = \langle dx, dy \rangle$ in \mathbf{R}^2 . Associated to such a differential form is the vector field $\mathbf{F} = \langle M, N \rangle$, and the form is just the expression $\mathbf{F} \cdot d\mathbf{r}$ appearing in line integrals for \mathbf{F} . Similarly, given a vector field $\mathbf{F} = \langle M, N \rangle$, we may consider the differential form $M dx + N dy$. Thus, the two formalisms, vector fields and differential forms, are really just alternate notations for the same concept. This may seem to add needless complication, but there are situations, e.g., in thermodynamics, where it is easier to think about a problem in the language of forms than it is in the equivalent language of vector fields.

We can translate many of the notions we encountered in studying plane vector fields to the language of differential forms.

Recall first that a plane vector field $\mathbf{F} = \langle M, N \rangle$ is conservative if and only if it is the gradient of a scalar function $\mathbf{F} = \nabla f$. Then

$$M dx + N dy = \mathbf{F} \cdot d\mathbf{r} = \nabla f \cdot d\mathbf{r}.$$

You should recognize the expression on the right as the *differential of the function*

$$df = \nabla f \cdot d\mathbf{r} = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy.$$

Thus, $\langle M, N \rangle$ is conservative if and only if the associated form $M dx + N dy$ equals the differential df of a function. Such forms are called *exact*. For exact forms, the path independence property looks particularly simple:

$$\int_{\mathcal{C}} M dx + N dy = \int_{\mathcal{C}} df = f(\text{end of } \mathcal{C}) - f(\text{start of } \mathcal{C}).$$

Recall the screening test for the field $\mathbf{F} = \langle M, N \rangle$

$$\frac{\partial N}{\partial x} = \frac{\partial M}{\partial y} \quad (75)$$

which in some (but not all) circumstances will tell us if the field is conservative. We say that the corresponding differential form $M dx + N dy$ is *closed* if its components satisfy equation (75).

We now translate some of the things we learned about vector fields to the language of forms.

(a) “Every conservative field passes the screening test (75)” becomes “Every exact form is closed”.

(b) “The field $\frac{1}{r}\mathbf{u}_\theta$ satisfies the screening test but is not conservative” becomes “ $\frac{-y}{x^2+y^2}dx + \frac{x}{x^2+y^2}dy$ is closed but not exact”.

(c) “If the domain of a field is simply connected, then the field is conservative if and only if it passes the screening test” becomes “If the domain of a form is simply connected, then the field is exact if and only if it is closed”.

Note that because of (c), a closed form $M dx + N dy$ defined on some open set in the plane is always *locally* exact. That is, for any point in its domain, we can always choose a small neighborhood, i.e., a disk or a rectangle, and a function f defined on that neighborhood such that

$$M dx + N dy = df$$

on that neighborhood. However, we can’t necessarily find a single function which will work everywhere. If we can find an f which works everywhere in the domain of the form, the form would be exact, but we might say ‘globally exact’ to emphasize the distinction.

Exercises for 6.1.

1. Tell whether or not each of the differential forms $M(x, y)dx + N(x, y)dy$ below is exact. If it is exact, find a function f such that $df = M dx + N dy$. (In each case the domain of the form is \mathbf{R}^2 , so every closed form is exact.)
 - (a) $(2x \sin y + y^3 e^x)dx + (x^2 \cos y + 3y^2 e^x)dy$
 - (b) $(\frac{1}{2}y^2 - 2ye^x)dx + (y - e^x)dy$
 - (c) $(2xy^3)dx + (3x^2y^2)dy$

2. The variables need not be called x and y . In succeeding chapters, we will generally use t and y . In each of the following cases, check to see if the given form is exact, and if it is, find a function f of which it is the differential.

(a) $(3ty + y^2)dt + (t^2 + ty)dy$

(b) $x^2 e^{tx} dt + (1 + xt)e^{tx} dx$.

3. A differential form $P(x, y, z) dx + Q(x, y, z) dy + R(x, y, z) dz$ in \mathbf{R}^3 is called closed if

$$\frac{\partial R}{\partial y} = \frac{\partial Q}{\partial z}, \quad \frac{\partial R}{\partial x} = \frac{\partial P}{\partial z}, \quad \text{and} \quad \frac{\partial Q}{\partial x} = \frac{\partial P}{\partial y}.$$

It is called exact if it is of the form df for an appropriate function $f(x, y, z)$. What do each of these statements say about the corresponding vector field $\mathbf{F}(x, y, z) = P\mathbf{i} + Q\mathbf{j} + R\mathbf{k}$? Give a condition on a domain E in \mathbf{R}^3 such that every closed form on E will be exact.

4. Check that each of the following differential forms is closed and find a function f such that it equals df .

(a) $yz dx + xz dy + xy dz$.

(b) $(y + x \cos(xt)) dt + t dy + t \cos(xt) dx$.

5. Show that every differential form of the form $M(t) dt + N(y) dy$ is closed. Show that it is also exact because it is the differential of the function $f(t, y) = A(t) + B(y)$ where $A(t)$ and $B(y)$ are indefinite integrals

$$A(t) = \int M(t) dt \quad \text{and} \quad B(y) = \int N(y) dy.$$

6.2 Using Differential Forms

We want to develop general methods for solving first order differential equations. In most applications, the independent variable represents time, so instead of calling the variables x and y , we shall call them t and y with t being the independent variable and y being the dependent variable. Then a first order differential equation can *usually* be put in the form

$$\frac{dy}{dt} = f(t, y) \tag{76}$$

where f is some specified function. A solution of (76) is a function $y = y(t)$ defined on some real t -interval such that

$$y'(t) = f(t, y(t))$$

is valid for each t in the domain of the solution function $y(t)$.

Example 121 Consider $\frac{dy}{dt} = -\frac{2t+y}{t+2y}$. We shall ‘solve’ this equation by some *formal* calculations using differential forms, and later we shall try to see why the calculations should be expected to work.

First, cross multiply to obtain

$$(t+2y)dy = -(2t+y)dt$$

or

$$(2t+y)dt + (t+2y)dy = 0. \quad (77)$$

Consider the differential form appearing on the left of this equation. It is closed since

$$\frac{\partial}{\partial t}(t+2y) = 1 = \frac{\partial}{\partial y}(2t+y).$$

(Don’t get confused by the fact that what was previously called x is now called t .) Since the domain of this form is all of \mathbf{R}^2 (the t, y -plane), and that is simply connected, the form is exact. Thus, it is the differential of a function, $df = (2t+y)dt + (t+2y)dy$. We may find that f by the same method we used to find a function with a specified conservative vector field as gradient. After all, it is just the same theory in other notation. Thus, since $df = \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial y}dy$, we want

$$\frac{\partial f}{\partial t} = 2t+y \quad \frac{\partial f}{\partial y} = t+2y.$$

Integrating each of these, we obtain

$$f(t, y) = t^2 + yt + C(y) \quad f(t, y) = ty + y^2 + D(t),$$

and comparing, we see that we may take $C(y) = y^2, D(t) = t^2$. Thus,

$$f(t, y) = t^2 + ty + y^2$$

will work. (Check it by taking its differential!) Thus, equation (77) may be rewritten

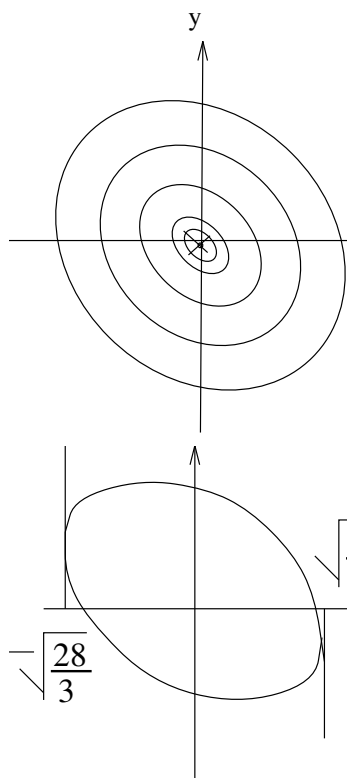
$$df = d(t^2 + ty + y^2) = 0$$

from which we conclude

$$f(t, y) = t^2 + ty + y^2 = c$$

for some constant c . We obtain this way an infinite collection of level curves of the function f as a “solution” to the original differential equation. In order to determine a unique solution, we must specify a point (t_0, y_0) lying on the level curve or impose some other equivalent condition. This corresponds to requiring that the solution of the differential equation satisfy an *initial condition* $y(t_0) = y_0$. (Refer back to Chapter II where the issue of initial conditions was first discussed.) For example, if we know that $y = 2$ when $t = 1$, we have

$$1^2 + (1)(2) + 2^2 = c \quad \text{or} \quad c = 7.$$



Hence, the corresponding “solution” is $t^2 + ty + y^2 = 7$. We can solve this for y in terms of t by applying the quadratic formula to the equation $y^2 + ty + (t^2 - 7) = 0$. This yields

$$y(t) = \frac{-t \pm \sqrt{t^2 - 4(t^2 - 7)}}{2} = \frac{-t \pm \sqrt{28 - 3t^2}}{2}$$

but only the plus sign gives a solution satisfying the condition $y = 2$ when $t = 1$.

Hence the solution we end up with is $y(t) = \frac{1}{2}(-t + \sqrt{28 - 3t^2})$.

Note that not every value of c yields a solution which makes sense. Thus, requiring $y = 0$ when $t = 0$ yields $c = 0$ or $t^2 + ty + y^2 = 0$ and the locus of this equation in \mathbf{R}^2 is the point $(0, 0)$. (Why?) This is not a total surprise, since the right hand side of the original differential equation $\frac{dy}{dt} = -\frac{2t + y}{t + 2y}$ is undefined at $(0, 0)$.

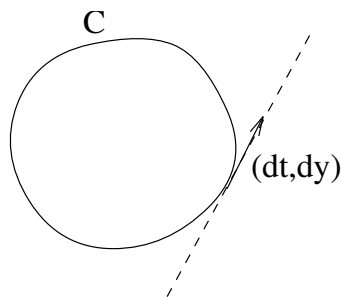
Analysis of the Solution Process We went from an equation of the form

$$\frac{dy}{dt} = f(t, y) \quad (78)$$

to one of the form

$$M dt + N dy = 0 \quad (79)$$

where $f(t, y) = -\frac{M(t, y)}{N(t, y)}$. (We could take $M = f$ and $N = 1$, but in many cases, as in Example 1, f is actually a quotient.) The two equations are not quite the same. A solution of (78) is a collection of functions $y = y(t)$, one for each initial condition, satisfying the differential equation, while a solution of (79) is a family of level curves $f(t, y) = c$ in \mathbf{R}^2 . The relation between the two is that the *graph* of each $y(t)$ is a subset of some one of the level curves.



In general, I shall say that a smooth curve C in the plane is a solution curve to equation (79) if at each point on the curve, the equation is true for each displacement vector $\langle dt, dy \rangle$ which is *tangent* to the curve. Note that at any point in the domain of the differential form, the equation (79) determines the ratio dy/dt , so it determines the vector $\langle dt, dy \rangle$ up to multiplication by a scalar. That means that the *line* along which the vector points is determined. (There is some fiddling you have to do if one or the other of the coefficients M, N vanishes. However, if they both vanish at a point, the argument fails and $\langle dt, dy \rangle$ is not restricted at all.) A general solution of (79), then, will be a family of such curves, one passing through each point of the domain of the differential form—except possibly where both coefficients vanish.

If $y = y(t)$ is a solution of the differential equation (78), then, at each point of its graph, we also have the relation

$$dy = y'(t)dt = f(t, y)dt = -\frac{M(t, y)}{N(t, y)}dt$$

so the graph is at least part of a solution curve to

$$M(t, y)dt + N(t, y)dy = 0. \quad (79)$$

Conversely, the reasoning can be reversed, so that any section of a solution curve to (79) which happens to be the graph of a function will be a solution to the original differential equation.

Example 121 illustrates this quite well. The curve $t^2 + ty + y^2 = 7$ is in fact an ellipse in the t, y -plane, but we can pick out a segment of that ellipse by solving $y(t) = \frac{1}{2}(-t + \sqrt{28 - 3t^2})$ for $28 - 3t^2 > 0$ (i.e., $-\sqrt{28/3} < t < \sqrt{28/3}$), and that yields a solution of the original differential equation satisfying $y = 2$ when $t = 1$.

There are differences between curves satisfying $M dt + N dy = 0$ and graphs of functions satisfying $dy/dt = f(t, y)$. First of all, a solution curve for the differential form could be a vertical line, and that certainly can't be the graph of a function. Secondly, at a point where $N = 0$ but $M \neq 0$, we must have $dt = 0$. At such a point the proposed tangent line to the curve is vertical, and the curve might not look like the graph of a function. However, this is quite reasonable, since at such a point the differential equation $dy/dt = -M/N$ has a singularity on the right. Finally, at points at which $M = N = 0$, all bets are off since no unique tangent line is specified. Such points are called *critical points* of the differential form.

Integrating Factors The strategy suggested in Example 1 for solving an equation of the form

$$M dt + N dy = 0$$

is to look for a function f so that $df = M dt + N dy$, and then to take the level curves of that function. Of course, this will only be feasible if the differential form is exact, so you may wonder what to try if it is *not* exact. The answer is to try to make it exact by multiplying by an appropriate function. Such a function is called an *integrating factor*. Thus, we want to look for a function $\mu(t, y)$ such that

$$\mu(M dt + N dy) = (\mu M)dt + (\mu N)dy$$

is exact.

Example 122 We shall try to solve $y dt - t dy = 0$. Note that

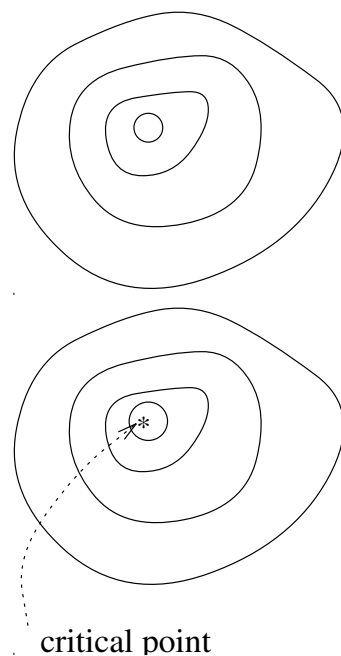
$$\frac{\partial(-t)}{\partial t} = -1 \neq 1 = \frac{\partial y}{\partial y}$$

so the form is not closed and hence it cannot be exact. We shall try to find $\mu(t, y)$ so that

$$(\mu y)dt - (\mu t)dy$$

is exact. Since every exact form is closed, we must have

$$\begin{aligned} \frac{\partial(-\mu t)}{\partial t} &= \frac{\partial(\mu y)}{\partial y} & \text{or} \\ -\frac{\partial \mu}{\partial t}t - \mu &= \frac{\partial \mu}{\partial y}y + \mu & \text{or} \\ \frac{\partial \mu}{\partial t}t + \frac{\partial \mu}{\partial y}y &= -2\mu. \end{aligned} \tag{80}$$



critical point

At this point, it is not clear how to proceed. Equation (80) alone does not completely determine μ , and indeed that is no surprise because in general there may be many possible integrating factors. Of course, we need only one. To find a μ we proceed in effect by making educated guesses. In particular, we shall look for an integrating factor which depends only on t , i.e., we assume $\mu = \mu(t)$. Of course, this may not work, but we have nothing to lose by trying. With this assumption, equation (80) becomes

$$\begin{aligned}\frac{d\mu}{dt}t &= -2\mu & \text{or} \\ \frac{d\mu}{\mu} &= -2\frac{dt}{t} & \text{which yields} \\ \ln |\mu| &= -2\ln |t| + c.\end{aligned}$$

Since we only need *one* integrating factor, we may take $c = 0$, so we get

$$\begin{aligned}\ln |\mu| &= \ln |t|^{-2} & \text{or} \\ \mu &= \pm \frac{1}{t^2}.\end{aligned}$$

Again, since we only need one integrating factor, we may as well take $\mu = 1/t^2$.

Having done all that work, you may think you are done, but of course, all we now know is that

$$\frac{1}{t^2}(y dt - t dy) = \frac{y}{t^2}dt - \frac{1}{t}dy$$

is probably exact. Hence, it now makes sense to look for a function $f(t, y)$ such that $df = \frac{1}{t^2}(y dt - t dy)$. If you remember the quotient rule, you will see immediately that $f(t, y) = -y/t$ is just such a function. However, let's be really dumb and proceed as if we hadn't noticed that. We use the usual method to find f . Integrate the equations

$$\frac{\partial f}{\partial t} = \frac{y}{t^2} \quad \frac{\partial f}{\partial y} = -\frac{1}{t}$$

to obtain

$$f = -\frac{y}{t} + C(y) \quad f = -\frac{y}{t} + D(t).$$

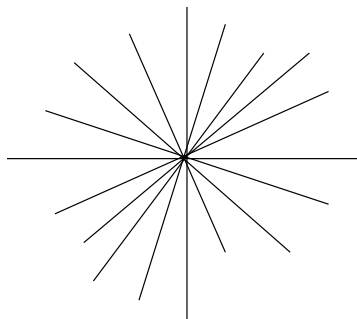
Clearly, we can take $C = D = 0$, so we do obtain $f(t, y) = -y/t$ as expected. Hence, the general solution curves are just the level curves of this function

$$f(t, y) = -y/t = c \quad \text{or} \quad y = -ct.$$

Clearly, it doesn't matter what we call the constant, so we can drop the sign and write the general solution as

$$y = ct.$$

This is the family of all straight lines through the origin with the exception of the y -axis.



There is one problem with the above analysis. What we actually obtained were solution curves for the equation

$$\mu(M dt + N dy) = 0.$$

rather than the original equation $M dt + N dy = 0$. If for example, $\mu(t, y) = 0$ has some solutions, that locus would be added as an additional *extraneous* solution which is not a solution of the original equation. In this example, $\mu = 1/t^2$ never vanishes, so we don't have any extraneous solutions. However, it has a singularity at $t = 0$, so its domain is smaller than that of the original form $y dt - t dy$ (which is defined everywhere in \mathbf{R}^2 .) Thus, it is possible that the curve $t = 0$, i.e., the y -axis is a solution curve for the original problem, which got lost when the integrating factor was introduced. In fact that is the case. If $t = 0$, so is $dt = 0$, so $y dt - t dy = 0$.

Hence, the general solution is the set of all straight lines through the origin. Note also that the origin itself is a critical point of $y dt - t dy = 0$.

In general, there are many possible integrating factors for a given differential form. Thus, in this example, $\mu = \frac{1}{yt}$ is also an integrating factor and results in the equation

$$\begin{aligned} \frac{1}{yt}(y dt - t dy) &= \frac{dt}{t} - \frac{dy}{y} = 0 \\ \text{or} \quad \frac{dt}{t} &= \frac{dy}{y} \end{aligned}$$

which is what we would have obtained by trying separation of variables. That would be the most appropriate method for this particular problem, but we did it the other way in order to illustrate the method in a simple case.

If you thoroughly understand the above example, you will be able to work many problems, but it is worthwhile also describing the *general case*. To see if the form

$$(\mu M)dt + (\mu N)dy$$

is closed, apply the screening test

$$\frac{\partial(\mu N)}{\partial t} = \frac{\partial(\mu M)}{\partial y}$$

which yields

$$\begin{aligned} \frac{\partial \mu}{\partial t} N + \mu \frac{\partial N}{\partial t} &= \frac{\partial \mu}{\partial y} M + \mu \frac{\partial M}{\partial y} \quad \text{or} \\ \frac{1}{\mu} \left[\frac{\partial \mu}{\partial t} N - \frac{\partial \mu}{\partial y} M \right] &= \frac{\partial M}{\partial y} - \frac{\partial N}{\partial t}. \end{aligned} \tag{81}$$

Unfortunately, *it is not generally possible to solve this equation!* This is quite disappointing because it also means that there is no general method to solve explicitly

the equation $M dt + N dy = 0$ or the related equation $dy/dt = f(t, y)$. There are however many special cases in which the method does work. These are discussed in great detail in books devoted to differential equations.

Typically one tries integrating factors depending only on one or the other of the variables or perhaps on their sum. This is basically a matter of trial and error.

Example 123 Consider

$$(t + e^y)dt + (\frac{1}{2}t^2 + 2te^y)dy = 0.$$

You should apply the screening test to check that the differential form on the left is not closed. We look for an integrating factor by considering equation (81) which in this case becomes

$$\frac{1}{\mu} \left[\frac{\partial \mu}{\partial t} (\frac{1}{2}t^2 + 2te^y) - \frac{\partial \mu}{\partial y} (t + e^y) \right] = \frac{\partial(t + e^y)}{\partial y} - \frac{\partial(\frac{1}{2}t^2 + 2te^y)}{\partial t}$$

or

$$\frac{1}{\mu} \left[\frac{\partial \mu}{\partial t} (\frac{1}{2}t^2 + 2te^y) - \frac{\partial \mu}{\partial y} (t + e^y) \right] = e^y - (t + 2e^y) = -(t + e^y).$$

(This was derived by substituting in equation (81), but you would have been just as well off deriving it from scratch by applying the screening test directly to the form $\mu(t + e^y)dt + \mu(\frac{1}{2}t^2 + 2te^y)dy$. There is no need to memorize equation (81).) It is apparent that assuming that $\partial\mu/\partial y = 0$ (i.e., μ depends only on t) leads nowhere. Try instead assuming $\partial\mu/\partial t = 0$. In that case, the equation becomes

$$-\frac{1}{\mu} \frac{d\mu}{dy} (t + e^y) = -(t + e^y) \quad \text{or} \quad \frac{1}{\mu} \frac{d\mu}{dy} = 1.$$

(Note that the partial derivative becomes an ordinary derivative since we have assumed μ is a function only of y .) This can be solved easily to obtain as a possible integrating factor

$$\mu = e^y.$$

I leave the details to you.

To continue, we expect that the form

$$e^y(t + e^y)dt + e^y(\frac{1}{2}t^2 + 2te^y)dy$$

is exact. Hence, we look for $f(t, y)$ such that

$$\frac{\partial f}{\partial t} = te^y + e^{2y} \quad \frac{\partial f}{\partial y} = \frac{1}{2}t^2 e^y + 2te^{2y}.$$

Integrating the first equation with respect to t and the second with respect to y yields

$$f(t, y) = \frac{1}{2}t^2e^y + te^{2y} + C(y) \quad f(t, y) = \frac{1}{2}t^2e^y + te^{2y} + D(t).$$

Comparing, we see that $C = D = 0$ will work, so we obtain $f(t, y) = \frac{1}{2}t^2e^y + te^{2y}$. Hence, the general solution is

$$\frac{1}{2}t^2e^y + te^{2y} = C.$$

Notice that in this case the integrating factor neither vanishes nor has any singularities, so we don't have to worry about adding extraneous solutions or losing solutions.

Note that in the above analysis we did not worry too much about the geometry of the domain. We know that it is not generally true that a closed form is exact, so even if we find an integrating factor, we may still not be able to find a f defined on the entire domain of the original differential form. However, this issue is not usually worth worrying about because there are so many other things which can go wrong in applying the method. In any case, we know that *locally* closed forms are exact, so that in any sufficiently small neighborhood of a point, we can always find a f which works in that neighborhood. Since, in general, solution methods for differential equations only give local solutions, this may be the best we can hope for. Extending local solutions to a global solution is often a matter of great interest and great difficulty.

Exercises for 6.2.

- Find a general solution of each of the following equations. (The differential forms are closed.)
 - $e^t \cos y \, dt + (2y - e^t \sin y) \, dy = 0$.
 - $(x^2 - y^2)dx + (y^2 - 2xy)dy = 0$.
- (a) By rewriting as a problem about differential forms, find a general solution of the differential equation

$$\frac{dy}{dt} = \frac{2t - 6y}{6t + 3y}.$$

You need not express the answer in the form $y = y(t)$. A relation between t and y will suffice.

- What is the form of the solution in (a) if $y(1) = 2$?
- Express the solution found in part (b) in the form $y = y(t)$. What should the domain of this function be?

3. Find a general solution of each of the following equations. (You will need to find integrating factors.)
- (a) $(1 + 6\frac{t}{y^2})dt + 2\frac{t}{y}dy = 0$.
- (b) $(y + e^{-x})dx + dy = 0$.
- (c) $(1 + t + y)dt - (1 + t)dy = 0$.
4. Solve $\frac{dx}{dt} = t - x$ given $x = 2$ when $t = 0$. Do this by rewriting as a problem in differential forms and finding an appropriate integrating factor.

6.3 First Order Linear Equations

The simplest first order equations are the *linear equations*, those of the form

$$\frac{dy}{dt} + a(t)y = b(t), \quad (82)$$

where $a(t)$ and $b(t)$ are known functions. Equation (82) may also be written in the form $\frac{dy}{dt} = f(t, y)$ by taking $f(t, y) = b(t) - a(t)y$.

Examples

$$\begin{aligned} \frac{dy}{dt} &= ky \\ \frac{dy}{dt} - ky &= j \\ \frac{dy}{dt} + \frac{1}{t}y &= \frac{1}{t^2} \end{aligned}$$

The first equation is that governing exponential growth or decay, and it was solved in Chapter II. The general solution is $y = Ce^{kt}$.

An example of a non-linear first order equation is

$$\frac{dy}{dt} + y^2 = t.$$

Equation (82) can be put in the form

$$(a(t)y - b(t))dt + dy = 0$$

and solved by the methods of the previous section. You should try it as an exercise to see if you understand those methods. However, we shall use another method which is easier to generalize to second order linear equations.

First consider the so-called *homogeneous* case where $b(t) = 0$, i.e., we want to solve

$$\frac{dy}{dt} + a(t)y = 0.$$

This may be done fairly easily by separation of variables.

$$\begin{aligned}\frac{dy}{y} &= -a(t)dt \\ \ln |y| &= -\int a(t)dt + c\end{aligned}$$

where $\int a(t)dt$ denotes any antiderivative

$$|y| = e^{-\int a(t)dt+c} = e^{-\int a(t)dt} e^c.$$

Put $C = \pm e^c$, depending on the sign of y , so the *general solution* takes the form

$$y = Ce^{-\int a(t)dt}, \quad (83)$$

where the constant C is determined as usual by specifying an initial condition. Of course, you could rederive this in each specific case if you remember the method, but you will probably save yourself considerable time by memorizing formula (83).

Example 124 Consider

$$\frac{dy}{dt} + \frac{1}{t}y = 0 \quad \text{for } t > 0.$$

Note that in this case $t = 0$ is a singularity of the coefficient $a(t)$. Hence, we would not expect the solution for $t < 0$ to have anything to do with the solution for $t > 0$.

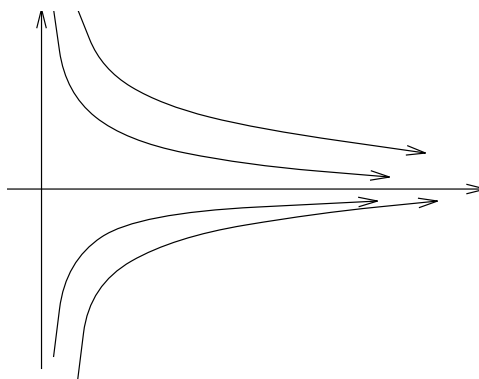
We have

$$\int a(t)dt = \int \frac{dt}{t} = \ln t.$$

(Of course, you could find the most general indefinite integral or antiderivative by adding ‘+c’, but we only need *one* antiderivative.) Hence, according to formula (83), the general solution is

$$y = Ce^{-\ln t} = \frac{C}{t}.$$

The graphs for some values of C are sketched below



Note how one specific solution may be picked out by specifying an initial condition $y(t_0) = y_0$

Consider next the *inhomogeneous case* where $b(t) \neq 0$, i.e.,

$$\frac{dy}{dt} + a(t)y = b(t).$$

Denote by

$$h(t) = e^{-\int a(t)dt}$$

the solution of the corresponding homogeneous equation with $C = 1$. We try to find a solution of the inhomogeneous equation of the form $y = u(t)h(t)$ where $u(t)$ is a function to be determined. (This method is called “variation of parameters”.) We have

$$\begin{aligned} \frac{d(uh)}{dt} + auh &= b, \\ \frac{du}{dt}h + u\frac{dh}{dt} + auh &= b, \\ \frac{du}{dt}h + u\left(\frac{dh}{dt} + ah\right) &= b. \end{aligned}$$

Since h was chosen to be a solution of the homogeneous equation, the quantity in parentheses vanishes. Hence, the above equation reduces to

$$\frac{du}{dt} = \frac{b}{h}.$$

This is easy to solve. The general solution is

$$u = \int \frac{b(t)}{h(t)} dt + C$$

where again the indefinite integral notation means that one specific antiderivative is selected. Since $y = hu$, we obtain the following general solution of the inhomogeneous equation

$$y = h(t) \int \frac{b(t)}{h(t)} dt + C h(t). \quad (84)$$

Note the form of this equation. The term $Ch(t) = Ce^{-\int a(t)dt}$ is a general solution of the homogeneous equation. The first term is the *particular solution* of the inhomogeneous equation obtained by setting $C = 0$. This is a theme which will be repeated often in what follows

A general solution of the inhomogeneous equation is obtained by adding a general solution of the homogeneous equation to a particular solution of the inhomogeneous equation.

Example 125 Consider

$$\frac{dy}{dt} + \frac{1}{t}y = \frac{1}{t^2} \quad \text{for } t > 0.$$

We determined $h(t) = \frac{1}{t}$ in the previous example where we solved the homogeneous equation. Moreover,

$$\int \frac{1/t^2}{1/t} dt = \int \frac{1}{t} dt = \ln t$$

where as suggested we pick one antiderivative. Then from equation (84), the general solution is

$$y = \frac{1}{t} \ln t + \frac{C}{t}.$$

It is important to note that equation (84) provides us, at least in principle, with a *complete solution* to the problem, a situation which will not arise often in our study of differential equations,

Sometimes Guessing is Okay In general, the solution of a first order differential equation is uniquely determined if an initial condition is specified. (We have seen how this works in several examples, and we shall discuss the theory behind it later.) Hence, if you are able to guess a solution that works, that is a perfectly acceptable way to proceed. We can adapt that principle to finding the general solution of a first order linear equation as follows. We may write the general solution (84) in the form

$$y = p(t) + Ch(t)$$

where $p(t) = h(t) \int (b(t)/h(t))dt$ is the particular solution obtained by setting $C = 0$. Suppose we can guess some particular solution $p_1(t)$ (which might be different from $p(t)$.) Then, for some value of the constant C_1 ,

$$p_1(t) = p(t) + C_1h(t) \quad \text{or} \quad p(t) = p_1(t) - C_1h(t).$$

Hence, the general solution may be written

$$\begin{aligned} y &= p_1(t) - C_1h(t) + Ch(t) = p_1(t) + (C - C_1)h(t) \\ &= p_1(t) + C'h(t) \end{aligned}$$

where $C' = C - C_1$ is also an arbitrary constant. What this says is that the general solution of the inhomogeneous equation is the sum of *any* particular solution and the general solution of the homogeneous equation.

Example 126 Consider the equation

$$\frac{dy}{dt} - ky = I \quad (85)$$

where k and I are constants. The homogeneous equation

$$\frac{dy}{dt} - ky = 0 \quad \text{or} \quad \frac{dy}{dt} = ky$$

has the general solution $y = Ce^{kt}$. We try to find a particular solution of the inhomogeneous equation by guessing. The simplest thing to try is a constant solution $y = A$. Substituting in (85) yields

$$0 - kA = I \quad \text{or} \quad A = -\frac{I}{k}.$$

Hence,

$$y = -\frac{I}{k} + Ce^{kt}$$

is a general solution of the inhomogeneous equation. Note that this is a bit simpler than using equation (84).

Using Definite Integrals in the Formulas The notation $\int a(t) dt$ stands for *any* antiderivative of the function $a(t)$. Thus

$$\int t dt = \frac{1}{2}t^2 \quad \text{and} \quad \int t dt = \frac{1}{2}t^2 + 1$$

are both true statements. This ambiguity can lead to confusion. To avoid this, we may on occasion use definite integrals with variable upper limits. Thus

$$\int_{t_0}^t a(s) ds$$

is definitely an antiderivative for $a(t)$ because the fundamental theorem of calculus tells us that

$$\frac{d}{dt} \int_{t_0}^t a(s) ds = a(t).$$

With this notation, we may write

$$h(t) = e^{-\int_{t_0}^t a(s) ds}.$$

Note that $h(t_0) = e^0 = 1$. In addition, we may write the general solution of the inhomogeneous equation

$$y = h(t) \int_{t_0}^t \frac{b(s)}{h(s)} ds + Ch(t).$$

Note that $y(t_0) = 0 + Ch(t_0) = C$, so it may also be written

$$y = h(t) \int_{t_0}^t \frac{b(s)}{h(s)} ds + y_0 h(t)$$

where $y_0 = y(t_0)$.

Note that we have been careful to use a ‘dummy variable’ s in the integrand. Sometimes people are a bit sloppy and just use t both for the integrand and the upper limit, but that is of course wrong.

Exercises for 6.3.

1. Find a general solution for each of the following linear differential equations. Either use the general formula, or find the general solution of the corresponding homogeneous solution and guess a particular solution.

(a) $\frac{dy}{dt} - (\sin t)y = 0$.

(b) $\frac{dx}{dt} - \frac{2t}{1+t^2}x = 1$.

(c) $\frac{dy}{dx} = \frac{1}{x}y + x$.

2. Solve each of the following initial value problems. Use the general formula or guess as appropriate.

(a) $\frac{dp}{dt} + 2p = 6$ given $p(0) = 4$.

(b) $t \frac{dy}{dt} + 2y = 3t$ given $y(1) = 4$.

(c) $(1+t^2) \frac{dx}{dt} + 2tx = e^t$ given $x(0) = 0$.

3. The charge q on a capacitor of capacitance C in series with a resistor of resistance R and battery with voltage V satisfies the differential equation

$$R \frac{dq}{dt} + \frac{q}{C} = V.$$

Assume $q(0) = 0$. Find $\lim_{t \rightarrow \infty} q(t)$.

4. In a certain chemical reaction, the rate of production of substance X follows the rule

$$\frac{dx}{dt} = -.001x + .01y$$

where x is the amount of substance X and y is the amount a catalyst Y . If the catalyst also decomposes according to the rule $y = 2e^{-.002t}$, find $x(t)$ assuming $x(0) = 0$.

5. Using the methods of the previous section, find an integrating factor $\mu = \mu(t)$ for

$$(b(t) - a(t)y)dt - dy = 0$$

and then solve the resulting equation. You should get the general solution of the inhomogeneous equation derived in this section.

6.4 Applications to Population Problems

With the little you now know, it is possible to solve a large number of problems both in natural and social sciences. To do so, you need to construct a *model* of reality which can be expressed in terms of a differential equation. While the mathematics that follows may be impeccable, the conclusions will still only be as good as the model. If the model is based on established physical law, as for example the theory of radioactive decay, the predictions will be quite accurate. On the other hand, in some other applications, there is little or no reason to accept the model, so the results will be of questionable value.

Population Growth with Immigration Let $p = p(t)$ denote the number of individuals in a population. (The individuals could be people, bacteria, radioactive atoms, etc.) One common assumption about population growth is that the number of births per unit time is proportional to the size of the population, i.e., the rate of change of $p(t)$ due to births is of the form $bp(t)$ for some constant b called the *birth rate*. Similarly, it is assumed that the number of deaths per unit time is of the form $dp(t)$ where d is another constant called the *death rate*. We may combine these and write symbolically

$$\frac{dp}{dt} = bp - dp = (b - d)p = kp$$

where $k = b - d$ is a constant combining both the birth and death rates. The solution to this equation is $p = p_0 e^{kt}$ where $p_0 = p(0)$. Thus according to this model, population grows exponentially if $k > 0$ and declines exponentially if $k < 0$. Suppose in addition to natural growth due to births and deaths, we also have immigration taking place at a constant rate I . (We can assume I is the net effect of immigration—into the population—and emigration—out of the population. It could even be negative in the case of net outflow.) The differential equation becomes

$$\begin{aligned} \frac{dp}{dt} &= kp + I \\ \text{or} \quad \frac{dp}{dt} - kp &= I. \end{aligned}$$

We solved this problem (with slightly different notation) in the previous section. The general solution is

$$p = -\frac{I}{k} + Ce^{kt}.$$

If $p = p_0$ at $t = 0$, we have

$$p_0 = -\frac{I}{k} + C \quad \text{or} \quad C = p_0 + \frac{I}{k}.$$

Hence, the solution can also be written

$$p = -\frac{I}{k} + (p_0 + \frac{I}{k})e^{kt}.$$

Note that *in this model* if $k > 0$ and $p_0 + I/k > 0$, then population grows exponentially, even if $I < 0$, i.e., even if there is net emigration.

The idea that natural populations grow exponentially was first popularized by Thomas Malthus (1766-1834), and it is a basic element of modern discussions of population growth. Of course, the differential equation model ignores many complicated factors. Populations come in discrete chunks and cannot be governed by a differential equation in a continuous variable p . In addition, most biological populations involve two sexes, only one of which produces births, and individuals vary in fertility due to age and many other factors. However, even when all these factors are taken into consideration, the Malthusian model seems quite accurate as long as the birth rate less the death rate is constant.

The Logistic Law The Malthusian model (with $k > 0$) leads to population growth which fills all available space within a fairly short time. That clearly doesn't happen, so limiting factors somehow come into play and change the birth rate so it no longer is constant. Various models have been proposed to describe real populations more accurately. One is the so-called *logistic law*. Assume $I = 0$ —i.e., there is no net immigration or emigration—and that the 'birth rate' has the form $k = a - bp$ where a and b are positive constants. Note that there is very little thought behind this about how populations behave. We are just choosing the simplest non-constant mathematical form for k . About the only thought involved is the decision to make the constant term positive and the coefficient of p negative. That is, we assume that as population grows, the birth rate goes down, although no mechanism is suggested for why that might occur. With these assumptions, the differential equation becomes

$$\frac{dp}{dt} = (a - bp)p.$$

This is not a linear equation, but it can be solved by separation of variables.

$$\frac{dp}{(a - bp)p} = dt \quad \text{or} \quad \int \frac{dp}{(a - bp)p} = t + c.$$

Assume $p > 0$ and $a - bp > 0$. In effect, that means we are looking for solution curves in the t, p -plane between the lines $p = 0$ and $p = a/b$. If we find any solutions which start in this region and leave it, we will be in trouble, so we will have to go back and redo the analysis. Under these assumptions, the integration on the left yields (by the method of partial fractions)

$$-\frac{1}{a} \ln(a - bp) + \frac{1}{a} \ln p = \frac{1}{a} \ln \left(\frac{p}{a - bp} \right).$$

(I tried to do this using Mathematica, but it kept getting the sign of one of the factors wrong. You should review the method of partial fractions and make sure you understand where the answer came from! If you can't do it, ask for help from your instructor.) Hence, we obtain

$$\ln\left(\frac{p}{a-bp}\right) = at + ac = at + c'$$

which after exponentiating yields

$$\frac{p}{a-bp} = e^{at}e^{c'} = Ce^{at}. \quad (86)$$

If $p(0) = p_0$, we have

$$\frac{p_0}{a-bp_0} = C.$$

Equation (86) may be solved with some simple algebra to obtain

$$p = \frac{Ca e^{at}}{1 + C b e^{at}} = \frac{\frac{p_0}{a-bp_0} a e^{at}}{1 + \frac{p_0 b}{a-bp_0} e^{at}}$$

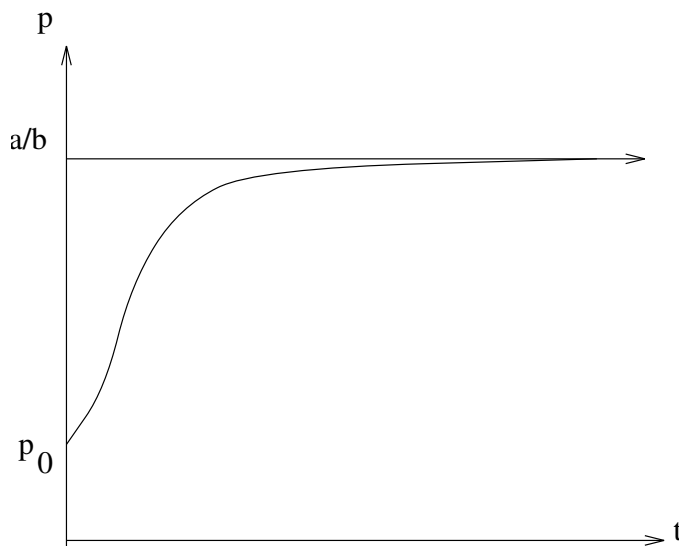
which simplifies to

$$\begin{aligned} p &= \frac{ap_0 e^{at}}{a - bp_0 + bp_0 e^{at}} \\ &= \frac{p_0 e^{at}}{1 + \frac{b}{a} p_0 (e^{at} - 1)}. \end{aligned}$$

To see what happens for large t , let $t \rightarrow \infty$. For this divide both numerator and denominator by e^{at} as in

$$p = \frac{p_0}{e^{-at} + \frac{b}{a} p_0 (1 - e^{-at})} \rightarrow \frac{p_0}{\frac{b}{a} p_0} = \frac{a}{b}.$$

Hence, the solution approaches the line $p = a/b$ asymptotically from below, but it never crosses that line. See the graph below which illustrates the population growth under the logistic law. Initially, it does appear to grow exponentially, but eventually population growth slows down and population approaches the equilibrium value $p_e = a/b$ asymptotically.



If you want to read a bit more about population models, see Section 1.5 of *Differential Equations and Their Applications* by M. Braun. However, you should take this with a grain of salt, since it is all based on assuming a particular model and then fiddling with parameters in the model to fit observation.

Note that the graphs of the solutions are also asymptotic to the line $p = 0$ as $t \rightarrow -\infty$. It may not be clear how to interpret this in terms of populations, but there is no problem with the mathematics. Thus, each of the solutions obtained above is defined for $-\infty < t < \infty$, and its graph is contained entirely in the strip $0 < p < a/b$. Moreover, the lines $p = 0$ and $p = a/b$ are also solutions, and there are solutions above and below these lines. (See the Exercises.) Indeed, by suitable choice of the arbitrary constants, you can find a solution curve passing through every point in the plane, and moreover these solution curves *never cross*.

The fact that solution curves don't cross (except in very special circumstances) is a consequence of the basic uniqueness theorem and will be discussed in the next section. However, we can use it here to justify our contention that that $0 < p(t) < a/b$ if $0 < p_0 < a/b$. What we have to worry about is the possibility that some other solution curve (not one arrived at above) could start inside the desired strip and later leave it. However, to do so, it would have to cross one of the bounding lines, and we saw that they also are solution curves, so that never happens.

Exercises for 6.4.

1. (a) The US population was approximately 4×10^6 in 1790 and 92×10^6 in 1910.

Assuming (incorrectly) Malthusian growth without immigration, estimate the ‘birth rate’ k .

- (b) If the population in 1980 was approximately 250×10^6 , would this be consistent with a Malthusian model without immigration?
2. The ‘birth rate’ of a certain population is .02 per year. Because of overcrowding, this induces a constant emigration of 10^6 per year. What does this predict about population in t years if the initial population is 10×10^6 .
3. (a) Show that $p = a/b$ and $p = 0$ are also solutions of the logistic equation. Which values of p_0 yield these solutions?
- (b) Find the solutions of the logistic equation arising from the assumption $p_0 > a/b$. Show that their solution curves approach the line $p = a/b$ asymptotically from above as $t \rightarrow \infty$.
- (c) Find the solutions of the logistic equation arising from the assumption $p_0 < 0$. Show their solution curves approach the line $p = 0$ asymptotically as $t \rightarrow -\infty$.
4. Consider the population model described by the equation

$$\frac{dp}{dt} = .04p - .01p^2 - .03.$$

(This is a logistic model with some emigration.) Assume $p(0) = 10^3$ and find $p(t)$.

5. In a chemical reaction a substance X is formed from two substances A and B . The reaction rate is governed by the rule

$$\frac{dx}{dt} = k(a - x)(b - x)$$

where $x = x(t)$ is the amount of the substance X present at time t . Here, a and b are constants related to the amounts of substances A and B initially present and k is a constant of proportionality. Assume $x(0) = 0$ and find $x(t)$. What happens as $t \rightarrow \infty$.

6.5 Existence and Uniqueness of Solutions

I mentioned earlier that the first order equations

$$\frac{dy}{dt} = f(t, y)$$

often cannot be solved explicitly. In such cases, one must resort to graphical or numerical techniques in order to get approximate solutions. Before attempting

that, however, one should be confident that there actually is a solution. Clearly, if we can't find the solution, we have to use other methods to convince ourselves it exists. In this section, we describe some of what is known in this area.

Consider the basic *initial value problem*: find $y = y(t)$ such that

$$\frac{dy}{dt} = f(t, y) \quad \text{where } y(t_0) = y_0.$$

There are two basic questions one may ask.

1. *Existence of a solution.* Under what circumstances can we be sure there is a solution $y = y(t)$ defined for some t -interval containing t_0 ? In this connection, we may also ask what is the largest possible domain for such a solution.
2. *Uniqueness of solutions.* Under what circumstances is the solution unique? That is, can there be two or more solutions satisfying the same initial condition?

It would be foolish to devote a lot of time and energy to finding a solution without first knowing the answers to these questions.

Example 127 Consider

$$\frac{dy}{dt} = y^2 \quad \text{where } y(0) = y_0,$$

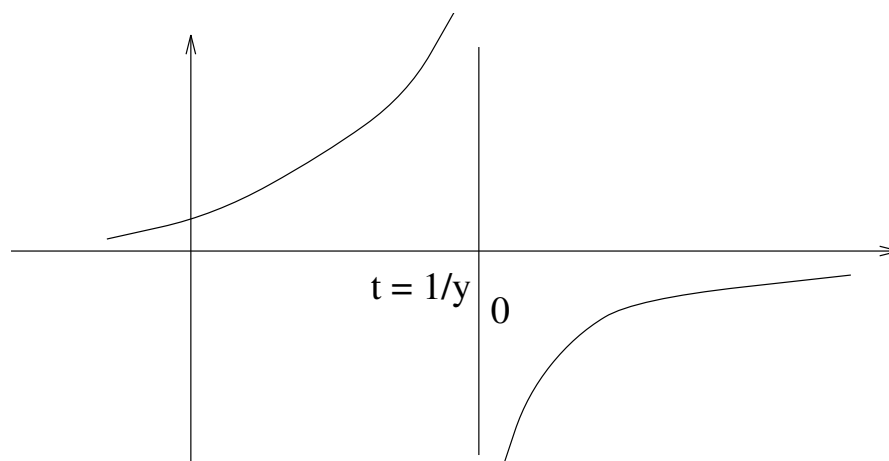
and we assume $y_0 > 0$. This equation is easy to solve by separation of variables.

$$\begin{aligned} \frac{dy}{y^2} &= dt \\ -\frac{1}{y} &= t + C \end{aligned}$$

so for $t = 0$

$$\begin{aligned} -\frac{1}{y_0} &= C \\ \text{and} \quad -\frac{1}{y} &= t - \frac{1}{y_0} \\ \text{so } y &= \frac{y_0}{1 - y_0 t}. \end{aligned}$$

The graph of this solution is a hyperbola asymptotic horizontally to the t -axis and asymptotic vertically to the line $t = 1/y_0$.



Thus the solution has a singularity at $t = 1/y_0$, and there is no coherent relationship between the solution to the right and left of the singularity. We conclude that there is a solution defined on the interval $-\infty < t < 1/y_0$, but that solution can't be extended smoothly to a solution on a larger interval. Note that the differential equation itself provides no hint that the solution has to have a singularity. The function $f(t, y) = y^2$ is smooth in the entire t, y -plane. The lesson to be learned from this is that at best we can be sure that the initial value problem has a solution *locally* in the vicinity of the initial point (t_0, y_0) .

The above solution is unique. The reasoning for that is a little tricky. Basically, the solution was deduced by a valid argument from the differential equation, so if there is a solution, it must be the one we found. Unfortunately, the solution process involved division by y^2 , so there is a minor complication. We have to worry about the possibility that there might be a solution with $y(t) = 0$ for some t . To eliminate that we argue as follows. If $y(t_1) \neq 0$ for a given t_1 , by continuity $y(t) \neq 0$ for all t in an interval centered at t_1 . Hence, the above reasoning is valid and the solution satisfies $-1/y(t) = t + C$ in that interval, which is to say its graph is part of a hyperbola. However, none of these hyperbolas intersect the t -axis, so no solution which is non-zero at one point can vanish at another. Notice, however, that $y(t) = 0$ for all t does define a solution. It satisfies the initial condition $y(0) = 0$.

Example 128 Consider

$$\frac{dy}{dt} = \frac{y}{t}.$$

I leave it to you to work out the general solution by separation of variables. It is $y = Ct$. For every $t_0 \neq 0$ and any y_0 , there is a unique such solution satisfying the initial condition $y(t_0) = y_0$. (Take $C = y_0/t_0$.) If $t_0 = 0$, there is no such solution satisfying $y(0) = y_0$ if $y_0 \neq 0$. For $y_0 = 0$, every one of these solutions satisfies the initial condition $y(0) = 0$, so the solution is not unique. This is not entirely surprising since $f(t, y) = y/t$ is not continuous for $t = 0$. Note however, that all the solutions $y = Ct$ are defined and continuous at $t = 0$.

Example 129 Consider

$$\frac{dy}{dt} = 3y^{2/3} \quad \text{where } y(0) = 0.$$

We can solve this by separation of variables

$$\begin{aligned} \frac{dy}{3y^{2/3}} &= dt \\ \frac{y^{1/3}}{3/3} &= t + C \end{aligned}$$

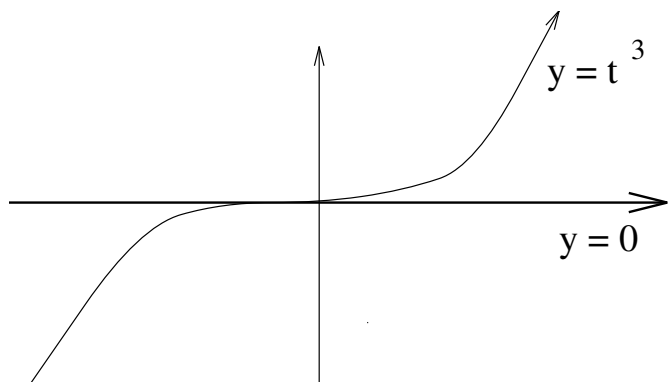
so putting $y = 0$ when $t = 0$ yields

$$\begin{aligned} 0 &= 0 + C \quad \text{or} \quad C = 0. \\ \text{Hence, } y^{1/3} &= t \quad \text{or } y = t^3 \end{aligned}$$

is a solution. (You should check that it works!) Unfortunately, the above analysis would exclude solutions which vanish, and it is easy to see that

$$y(t) = 0 \quad \text{for all } t$$

also defines a solution. Hence, there are (at least) two solutions satisfying the initial condition $y(0) = 0$.



Note that in this case $f(t, y) = 3y^{2/3}$ is continuous for all t, y . However, it is not terribly smooth since $f_y(t, y) = 2y^{-1/3}$ blows up for $y = 0$.

Propagation of Small Errors in the Solution The above example suggests that the uniqueness of solutions may depend on how smooth the function $f(t, y)$ is. We shall present an argument showing how uniqueness may be related to the behavior of the partial derivative $f_y(t, y)$. You might want to skip this discussion your first time through the subject since the reasoning is quite intricate.

Suppose $y_1(t)$ and $y_2(t)$ are two solutions of $dy/dt = f(t, y)$ which happen to be close for one particular value $t = t_0$. (The extreme case of ‘being close’ would be ‘being equal’.) Let $\epsilon(t) = y_1(t) - y_2(t)$. We want to see how far apart the solutions can get as t moves away from t_0 , so we try to determine an upper bound on the function $\epsilon(t)$. We have

$$\begin{aligned}\frac{dy_1}{dt} &= f(t, y_1(t)) \\ \frac{dy_2}{dt} &= f(t, y_2(t))\end{aligned}$$

Subtracting yields

$$\frac{d(y_1 - y_2)}{dt} = \frac{dy_1}{dt} - \frac{dy_2}{dt} = f(t, y_1(t)) - f(t, y_2(t)).$$

However, if the difference $\epsilon = y_1 - y_2$ is *small enough* we have approximately

$$f(t, y_1) - f(t, y_2) \approx f_y(t, y_2(t))(y_1 - y_2) = f_y(t, y_2)\epsilon.$$

That means that $\epsilon(t)$ *approximately* satisfies the differential equation

$$\frac{d\epsilon}{dt} = f_y(t, y_2(t))\epsilon.$$

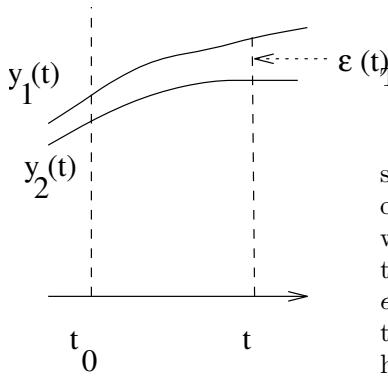
This equation is linear, and it has the solution

$$\epsilon = C e^{\int f_y(t, y_2(t)) dt}, \quad (87)$$

so we can take the right hand side as an *estimate* of the size of the error as a function of t . Of course, the above reasoning is a bit weak. First of all, we haven’t defined what we mean by an ‘approximate’ solution of a differential equation. Moreover, the approximation is only valid for ϵ small, but we want to use the result to see if ϵ is small. That is certainly circular reasoning. However, it is possible to make all this precise by judicious use of inequalities, so the method can give you an idea of how the difference of two solutions propagates.

There are two important conclusions to draw from the estimate in (87). First, assume the function f_y is continuous, so the exponential on the right will be well behaved. In that case if the difference between two solutions ϵ is initially 0, the constant C will be 0, so the difference ϵ will vanish for all t , at least in the range where all the approximations are valid. Second, assume the function $f_y(t, y_2(t))$ has a singularity at t_0 , and the exponent in (87) diverges to $-\infty$ as $t \rightarrow t_0$. In that case, the exponential would approach zero, meaning that we could have $\epsilon = y_1(t) - y_2(t) \neq 0$ for $t \neq t_0$ but have it approach 0 as $t \rightarrow t_0$. In other words, the graphs of the two solutions might cross at (t_0, y_0) . In Example 129, that is exactly what occurs. We have $f_y = 2y^{-1/3}$, so if we take $y_2 = t^3$ and $y_1 = 0$, we have

$$\begin{aligned}\int f_y(t, y_2(t)) dt &= \int 2(t^3)^{-1/3} dt = 2 \int t^{-1} dt = \ln t^2 \\ \text{so } e^{\int f_y(t, y_2(t)) dt} &= e^{\ln t^2} = t^2,\end{aligned}$$

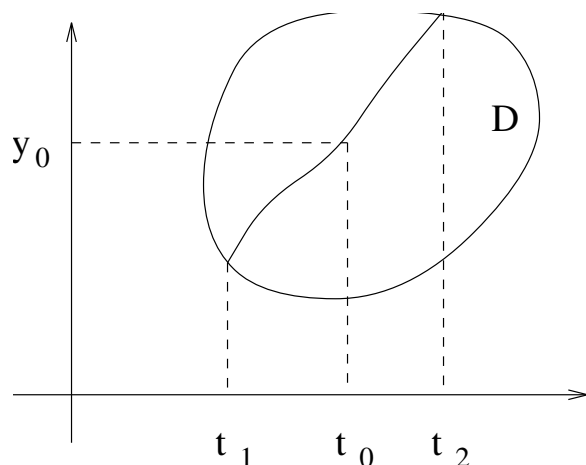
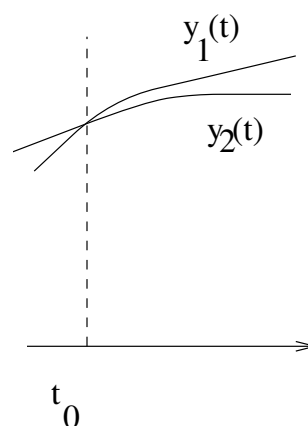


and this does in fact approach 0 as $t \rightarrow 0$. **The Basic Existence and Uniqueness Theorem**

Theorem 6.10 Assume $f(t, y)$ is continuous on an open set D in the t, y -plane. Let (t_0, y_0) be a point in D . Then there is a function $y(t)$ defined on some interval (t_1, t_2) containing the point t_0 such that

$$\begin{aligned} \frac{dy(t)}{dt} &= f(t, y(t)) && \text{for every } t \text{ in } (t_1, t_2) \\ \text{and} &&& y(t_0) = y_0. \end{aligned}$$

Moreover, if $f_y(t, y)$ exists and is continuous on D , then there is at most one such solution defined on any interval containing t_0 .

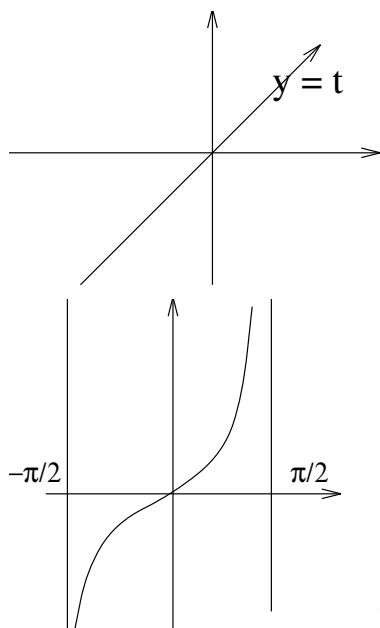


The proof of this theorem is quite involved, so it is better left for a course in Real Analysis. However, most good books on differential equations include a proof of some version of this theorem. See for example, Section 1.10 of *Braun* which uses the method of *Picard iteration*.

Example 130 Consider the logistic equation

$$\frac{dp}{dt} = ap - bp^2.$$

Here $f(t, p) = ap - bp^2$ is continuous for all (t, p) , so a solution exists satisfying any possible initial condition $p(t_0) = p_0$. Similarly $f_p(t, p) = a - 2bp$ is continuous everywhere, so every solution is unique. Thus, graphs of solutions $p = p(t)$ never cross.



Example 131 Consider

$$\frac{dy}{dt} = (y - t)^{1/3}.$$

Here $f(t, y) = (y - t)^{1/3}$ is continuous for all (t, y) so a solution always exists satisfying any given initial condition. However, $f_y(t, y) = \frac{1}{3} \frac{1}{(y - t)^{2/3}}$ has singularities along the line $y = t$. Hence, we cannot be sure of the uniqueness of solutions satisfying initial conditions of the form $y_0 = t_0$.

Let us return to the second part of the *existence* question: how large a t -interval can we choose for the domain of definition of the solution. We saw in Example 1

that this cannot be determined simply by examining the domain of $f(t, y)$. Here is another example.

Example 132 Consider the initial value problem $\frac{dy}{dt} = 1 + y^2$ with $y(0) = 0$. The equation can be solved by separation of variables, and the solution is $y = \tan t$. (Work it out yourself.) The domain of the continuous branch of this function passing through $(0, 0)$ is the interval $-\pi/2 < t < \pi/2$.

Even in cases where the differential equation can't be solved explicitly, it is *sometimes* possible to determine an interval on which there is a solution. We illustrate how to do this in the above example (without solving the equation), but the reasoning is quite subtle, and you may not want to go through it at this time.

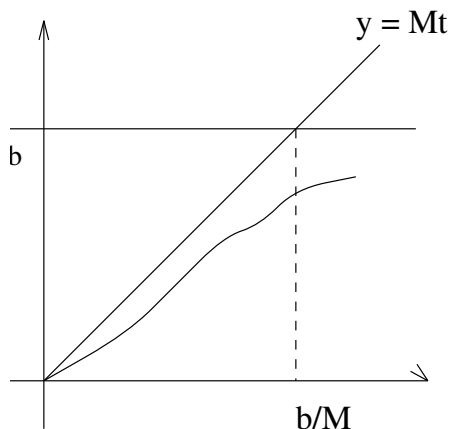
Example 132, continued We consider the case $0 \leq t$. (The reasoning for $t < 0$ is similar.) The problem is to find a value a such that when $0 \leq t \leq a$, the solution won't grow so fast that it will become asymptotic to some line $t = t_1$ with $0 < t_1 < a$. Since $y' = 1 + y^2 \geq 1$, there is certainly no problem with the solution going to $-\infty$ in such an interval, so we only concern ourselves with positive asymptotes. Suppose that we want to be sure that $0 \leq y \leq b$. Then we would have $y' = 1 + y^2 \leq 1 + b^2$, i.e., the graph would lie under the a line (starting at the initial point $(0, 0)$) with slope $1 + b^2$. That line has equation $y = (1 + b^2)t$ so it intersects the horizontal line $y = b$ in the point $(\frac{b}{1 + b^2}, b)$. Thus, for a given b , we could take $a = \frac{b}{1 + b^2}$, and the solution could not go to ∞ for $0 < t < a$.

The trick is to maximize a by finding the maximum value of $b/(1 + b^2)$ over all possible $b > 0$. This may be done by elementary calculus. Setting

$$\frac{d}{db} \frac{b}{1 + b^2} = \frac{(1 + b^2) - b(2b)}{(1 + b^2)^2} = \frac{1 - b^2}{(1 + b^2)^2} = 0$$

yields $b = 1$ (since we know $b > 0$). It is easy to see that this is indeed a maximum point and the maximum value is $1/(1 + 1^2) = 1/2$. It follows that we can be sure there is a solution for $0 \leq t < 1/2$. Similar reasoning (or a symmetry argument) shows that there is also a solution for $-1/2 < t \leq 0$.

Note that this reasoning gave us an interval $(-1/2, 1/2)$ somewhat smaller than that obtained above for the exact solution.



The analysis in the above example involves ‘lifting oneself by one’s bootstraps’ in that one uses ‘ b ’ (a bound on y) to find ‘ a ’ (a bound on t), whereas what we really want is the reverse. It only works well if $f(t, y)$ is independent of t or is bounded by functions independent of t . In general, there may be no effective way to determine an interval on which we can be sure there is a solution, although the existence theorem assures there is such an interval.

Exercises for 6.5.

- (a) Solve the initial value problem $\frac{dy}{dt} = t(y - 1)$ given $y(0) = 1$ by separation of variables.

(b) How can you conclude that this is the only solution?
- (a) Solve the initial value problem $\frac{dy}{dt} = y^{4/5}$ given $y(0) = 0$ by separation of variables.

(b) Note that $y = 0$ is also a solution to this initial value problem. Why doesn’t this contradict the uniqueness theorem?
- (a) Solve the initial value problem $\frac{dy}{dt} = \sqrt{1 - y^2}$ given $y(0) = 1$ by separation of variables.

(b) Note that $y = 1$ is also a solution to this initial value problem. Why doesn’t this contradict the uniqueness theorem?

4. (a) Solve $\frac{dy}{dt} = 2t(1 + y^2)$, $y(0) = 0$ by separation of variables.
 (b) What is the largest possible domain of definition for the solution obtained in part (a)?
5. Consider the initial value problem $y' = t^2 + y^2$ given $y(0) = 0$. This cannot be solved explicitly by any of the methods we have discussed, but the existence and uniqueness theorem tells us there is a unique solution defined on some t -interval containing $t = 0$.

Show that $(-a, a)$ is such an interval if $a \leq \frac{1}{\sqrt{2}}$ as follows. Fix a value $b > 0$.

Since $y' = t^2 + y^2 \leq a^2 + b^2$ for $0 \leq t \leq a$, $0 \leq y \leq b$ (a rectangle with opposite corners at $(0, 0)$ and (a, b)), it follows that a solution curve starting at $(0, 0)$ will pass out of the right hand side of the rectangle provided $a^2 + b^2 \leq \frac{b}{a}$. This may be rewritten $ab^2 - b + a^3 \leq 0$. Fix a and determine a condition such that the resulting quadratic expression in b does assume negative values, i.e., such that the equation $ab^2 - b + a^3 = 0$ has two unequal real roots, at least one of which is positive.

Do you think this initial value problem has a solution defined for $-\infty < t < \infty$?

6.6 Graphical and Numerical Methods

What if the initial value problem for a first order differential equation cannot be solved explicitly? In that case one may try to use graphical or numerical techniques to find an approximate solution. Even if one can't get very precise values, it is sometimes possible to get a qualitative understanding of the solution.

The Direction Field For each point (t, y) in \mathbf{R}^2 , $f(t, y)$ specifies the slope that the graph of a solution passing through that point should have. One may draw small line segments at selected points, and it is sometimes possible to sketch in a good approximation to the solution.

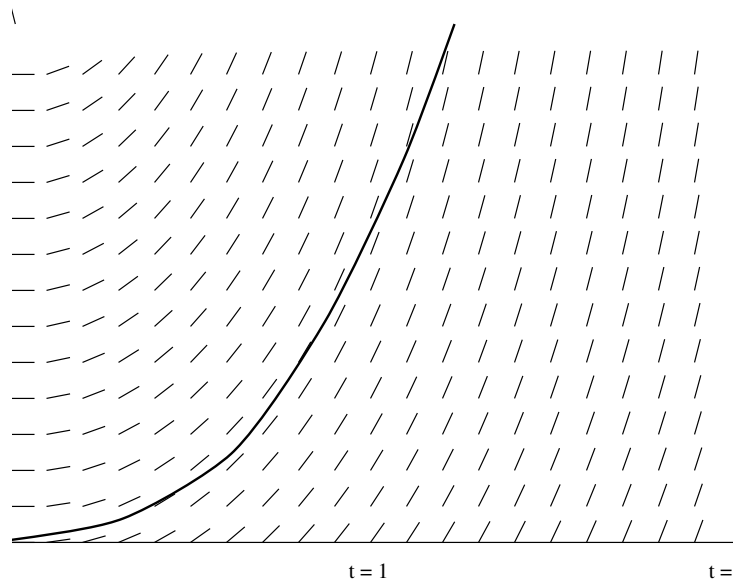
Example 133 Consider

$$\frac{dy}{dt} = t\sqrt{1 + y^3} \quad \text{where } y(0) = 1.$$

We can make an attempt to solve this by separation of variables to obtain

$$\begin{aligned} \frac{dy}{\sqrt{1 + y^3}} &= t \, dt \\ \text{so } \int \frac{dy}{\sqrt{1 + y^3}} &= \frac{1}{2}t^2 + c. \end{aligned}$$

Unfortunately, the left hand side cannot be integrated explicitly in terms of known functions, so there is no way to get an explicit solution $y = y(t)$. In the diagram below, I sketched some line segments with the proper slope $t\sqrt{1+y^3}$ at points in the first quadrant. Also, starting at the point $(0, 1)$, I attempted to sketch a curve which is tangent to the line segment at each point it passes through. I certainly didn't get it exactly right, but the general characteristics of the solution seem clear. It increases quite rapidly and it may even have an asymptote around $t = 2$.



We may verify analytically that the graph has an asymptote as follows. We have

$$\frac{dy}{dt} = t\sqrt{1+y^3} > t\sqrt{y^3} = ty^{3/2} \quad \text{for } t > 0, y > 0.$$

Hence, in the first quadrant

$$\frac{dy}{y^{3/2}} > t \, dt$$

so

$$\int_1^y \frac{du}{(u)^{3/2}} > \int_0^t t' \, dt' = \frac{1}{2}t^2.$$

Note that we have to use definite integrals in order to preserve the inequality, and we also used that $y > 1$ along the solution curve. The latter is apparent from the direction field, although a detailed proof might be tricky. Integrating on the left,

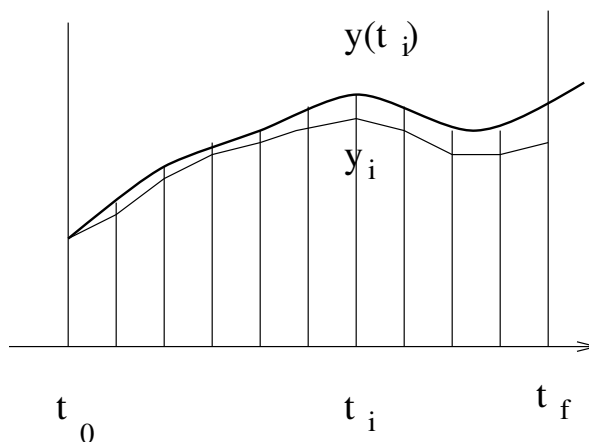
we obtain

$$\begin{aligned} -\frac{2}{u^{1/2}} \Big|_1^y &> \frac{1}{2}t^2 \\ -2\left(\frac{1}{\sqrt{y}} - 1\right) &> \frac{1}{2}t^2 \\ \text{or } \sqrt{y} &> \frac{1}{1 - \frac{1}{4}t^2} \\ \text{or } y &> \frac{1}{(1 - \frac{1}{4}t^2)^2}, \end{aligned}$$

and the expression on the right has an asymptote at $t = 2$.

Euler's Method

There are a variety of *numerical methods* for solving differential equations. Such a method can generate a table of values y_i which are approximations to the true solution $y(t_i)$ for selected values t_i of the independent variable. Usually, the values t_i are obtained by dividing up a specified interval $[t_0, t_f]$ into n equally spaced points.



The simplest method is called *Euler's method*. It is based on the linear approximation

$$y(t+h) = y(t) + y'(t)h + o(h)$$

where as in Chapter III, Section 4,

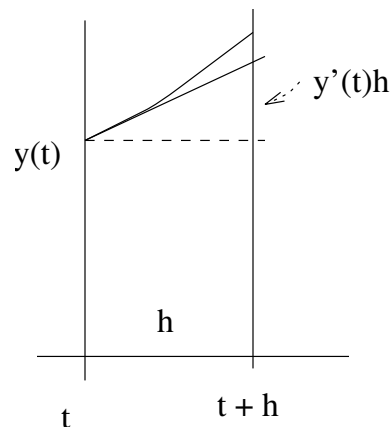
the notation ' $o(h)$ ' indicates a quantity small compared to h . ($\lim_{h \rightarrow 0} o(h)/h = 0$.) Since $y'(t) = f(t, y(t))$, we have approximately

$$y(t+h) \approx y(t) + f(t, y)h.$$

Using this, we may describe Euler's method by the following Pascal-like pseudo-code, where t_0 and y_0 denote the initial values of t and y , t_f is the 'final' value of t , n is the number of steps from t_0 to t_f , and $h = (t_f - t_0)/n$ is the step size.

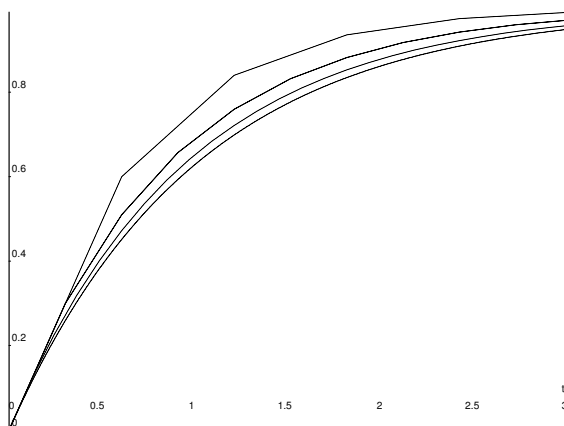
```

read  $t_0, y_0, t_f, n$ ;
 $h := (t_f - t_0)/n$ ;
 $t := t_0$ ;
 $y := y_0$ ;
while ( $t \leq t_f + h/2$ ) do
  begin
     $yy := y + f(t, y) * h$ ;
     $t := t + h$ ;
     $y := yy$ ;
    show( $t, y$ );
  end;
end.
```

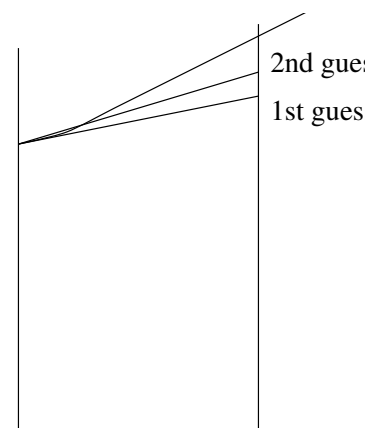


Note the construction ' $t \leq t_f + h/2$ '. This complication is needed because of the way numerical calculations are done in a computer. Namely, the calculation of h will involve either a round-off error or truncation error in almost all cases. As a result, n iterations of the step $t := t + h$ could yield a value for t slightly greater than t_f , but it will certainly yield a value short of $t_f + h/2$. Adding the $h/2$ will insure that the loop stops exactly where we want it to.

Euler's method is very simple to program, but it is not too accurate. It is not hard to see why. At each step, we make an error due to our use of the linear approximation rather than the true solution at that point. In effect that moves us onto a neighboring solution curve. After many steps, these errors will *compound*, i.e., the errors from all previous steps will add additional errors at the current step. Hence, the cumulative error after a large number of steps can get very large. See Figure 1 for the graphs of some solutions of the initial value problem $dy/dt = 1 - y$, $y(0) = 0$ by Euler's method with different values of n . The exact solution in this case is $y = 1 - e^{-t}$. In this case, the approximate solutions all lie above the exact solution. Can you see why?

Figure 1. Euler's method for $n = 5, 10, 25$ and exact solution

One can estimate the error due to the approximation in Euler's method. It turns out that the error in the calculation of the last value $y(t_f)$ is $O(h)$. That means if you double the number of steps, you will generally halve the size of the error. See *Braun* Section 1.13.1 for a discussion of the error analysis. Thus, in principle, you ought to be able to get any desired degree of accuracy simply by making the step size small enough. However, the error estimate $O(h)$ is based on the assumption of exact arithmetic. Numerical considerations such as round-off error become increasingly important as the step size is made smaller. (For example, in the extreme case, if the number of steps n is chosen large enough, the computer may decide that the step size h is zero, and the while loop won't terminate!) As a result, there is a point of diminishing returns where the answer actually starts getting worse as the step size is decreased. To get better answers one needs a more sophisticated method.



2nd guess

1st guess

The Improved Euler Method Euler's method may be improved by incorporating a feedback mechanism which tends to correct errors. We start off as before with the tangent approximation

$$yy := y + f(t, y) * h;$$

but then we use this provisional value yy at $t + h$ to calculate the putative slope $f(t + h, yy)$ at $(t + h, yy)$. This is now averaged with the slope $f(t, y)$ at (t, y) , and the result is used to determine a new value yyy at $t + h$

$$yyy := y + ((f(t, y) + f(t + h, yy))/2) * h;$$

Here is some Pascal like pseudo code implementing this algorithm.

```
read  $t_0, y_0, t_f, n$ ;
```

```

 $h := (t_f - t_0)/n;$ 
 $t := t_0;$ 
 $y := y_0;$ 
while  $(t \leq t_f + h/2)$  do
  begin
     $m := f(t, y);$ 
     $yy := y + m * h;$ 
     $t := t + h;$ 
     $yyy := y + ((m + f(t, yy))/2) * h;$ 
     $y := yyy;$ 
    show( $t, y$ );
  end;
end.

```

See Figure 2 for graphs showing the behavior of the improved Euler method for the initial value problem $y' = 1 - y$, $y(0) = 0$ for various n . Note that the approximations lie below the exact solution in this case. Can you see why? Note also that the improved Euler method does give better accuracy than the Euler method for the same value of n .

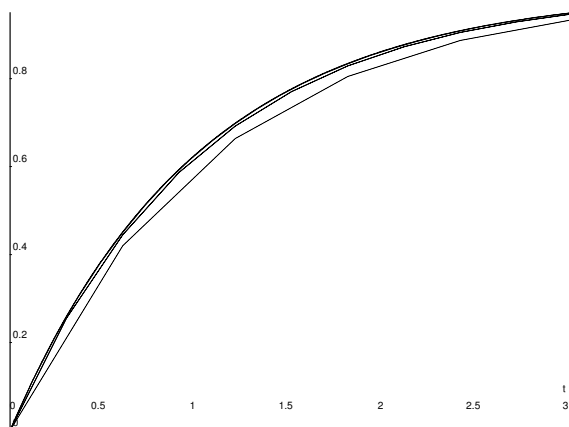


Figure 2. Improved Euler for $n = 5, 10, 25$ versus exact solution.

One can also estimate the error due to the approximation in the improved Euler method. It turns out that the error in the calculation of the last value $y(t_f)$ is $O(h^2)$. That means if you double the number of steps, you will generally reduce the size of the error by a factor of $1/4$. See *Braun* Section 1.13.1 for a discussion of the error analysis.

There are many methods for wringing out the last bit of accuracy when attempting to solve a differential equation numerically. One very popular method is the *Runge-Kutte method* which tries to use quadratic approximations rather than linear approximations. There are people who have devoted a considerable part of their professional careers to the study of numerical solutions of differential equations, and their results are embodied in a variety of software packages.

Exercises for 6.6.

1. Sketch the direction field for the differential equation $y' = t^2 + y^2$ in the rectangle $0 \leq t \leq 1, 0 \leq y \leq 3$. Try to sketch solution curves for the following initial conditions. Use as many points as you feel necessary for an accurate sketch. One shortcut is to plot selected points with slopes m for various m 's. Thus, all points with slope $m = 1$, lie along the circle $t^2 + y^2 = 1$. (You may also use a software package for sketching direction fields and solutions if you can find one and figure out how to use it.)
(a) $y(0) = 0$. (b) $y(0) = 1$.

On the basis of your sketch would you expect any problems in either case in a numerical method (such as Euler's method or the improved Euler's method) to find $y(1.1)$?

2. (optional) Program the improved Euler's method for $y' = t^2 + y^2, y(0) = 1$, and try to determine $y(1.1)$ for (a) $y(0) = 0$ and for (b) $y(0) = 1$.

Chapter 7

Second Order Linear Differential Equations

7.1 Second Order Differential Equations

The general second order differential equation may be put in the form

$$\frac{d^2y}{dt^2} = f(t, y, y'),$$

and a solution is a function $y = y(t)$ defined on some t -interval $t_1 < t < t_2$ and satisfying

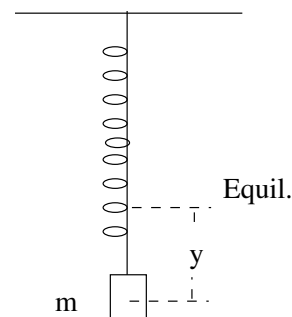
$$y''(t) = f(t, y(t), y'(t)) \quad \text{for } t_1 < t < t_2.$$

Note that the function f on the right is a function of three independent variables, for which t , the function $y(t)$, and its derivative $y'(t)$ are substituted to check a solution.

As we saw in specific cases in Chapter II, a general solution of a second order equation involves two arbitrary constants, so you need two conditions to determine a solution completely. The fundamental existence theorem asserts that if $f(t, y, y')$ is continuous on its domain, then there exists a solution satisfying initial conditions of the form

$$\begin{aligned} y(t_0) &= y_0 \\ y'(t_0) &= y'_0. \end{aligned}$$

(Note that *both* the solution and its derivative are specified at t_0 .) The fundamental uniqueness theorem asserts that if $f_y(t, y, y')$ and $f_{y'}(t, y, y')$ exist and are continuous, then on any given interval containing t_0 there is at most one solution satisfying the given initial conditions.



Example 134 As you saw earlier in this course and in your physics course, the differential equation governing the motion of a weight at the end of spring has the form

$$\frac{d^2 y}{dt^2} = -\frac{k}{m}y$$

where y denotes the displacement of the weight from equilibrium. Similar differential equations govern other examples of *simple harmonic motion*. The general solution can be written

$$y = C_1 \cos(\sqrt{k/m}t) + C_2 \sin(\sqrt{k/m}t)$$

or

$$y = A \cos(\sqrt{k/m}t + \delta)$$

and in either case there are two constants which may be determined by specifying an initial displacement $y(t_0) = y_0$ and an initial velocity $y'(t_0) = y'_0$.

In physics, you may also have encountered the equation for *damped simple harmonic motion* which has the form

$$\frac{d^2 y}{dt^2} = -\frac{k}{m}y - \frac{b}{m} \frac{dy}{dt}.$$

You may also have seen a discussion of solutions of this equation. Again, two arbitrary constants are involved.

We solved the equation for undamped simple harmonic motion in Chapter II by means of a trick which will always work if the function $f(t, y, y') = f(y, y')$ does not depend explicitly on the independent variable t . In that case, you can put $u = y'$ and then by the chain rule

$$\frac{d^2 y}{dt^2} = \frac{du}{dt} = \frac{du}{dy} \frac{dy}{dt} = \frac{du}{dy} u,$$

so the equation can be put in the form

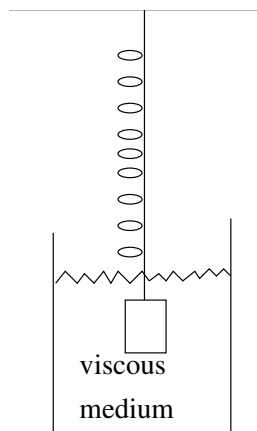
$$u \frac{du}{dy} = f(y, u).$$

This equation can then be solved to express $u = dy/dt$ in terms of y , and then the resulting first order equation may be solved for y . This method would also work for the equation for damped simple harmonic motion. (Try it!).

There is one other circumstance in which a similar trick will work: if $f(t, y, y') = f(t, y')$ does not actually involve the dependent variable y . Then we can put $u = y'$ and we have a first order equation of the form

$$\frac{du}{dt} = f(t, u).$$

Except for these special cases, there are no good methods for solving general second order differential equations. However, for linear equations, which we shall consider



in the rest of this chapter, there are powerful methods. Fortunately, many important second order differential equations arising in applications are linear.

Exercises for 7.1.

1. The exact equation satisfied by a mass at the end of a pendulum of length L is $d^2\theta/dt^2 = -(g/L)\sin\theta$. Put $u = d\theta/dt$ and reduce to a first order equation in θ and t . Do not try to solve this equation.
2. Find a general solution for $y'' = (y')^2$ by putting $u = y'$ and first solving for u .

7.2 Linear Second Order Differential Equations

A second order differential equation is called *linear* if it can be put in the form

$$\frac{d^2y}{dt^2} + p(t)\frac{dy}{dt} + q(t)y = f(t)$$

where $p(t)$, $q(t)$, and $f(t)$ are known functions. This may be put in the form $d^2y/dt^2 = f(t, y, y')$ with $f(t, y, y') = f(t) - q(t)y - p(t)y'$.

Example 135 The equation for *forced, damped harmonic motion* has the form

$$\frac{d^2y}{dt^2} + \frac{b}{m}\frac{dy}{dt} + \frac{k}{m}y = \frac{F_0}{m}\cos\omega t.$$

The term on the right represents a periodic driving force with period $2\pi/\omega$.

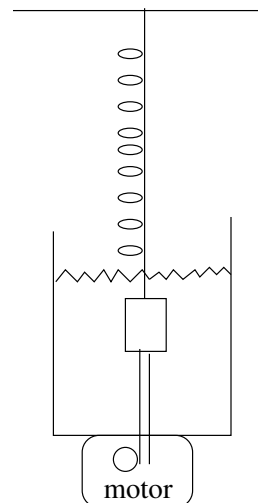
Example 136 The equation

$$\frac{d^2y}{dt^2} - \frac{2t}{1-t^2}\frac{dy}{dt} + \frac{\alpha(\alpha+1)}{1-t^2}y = 0,$$

where α is some constant, is called *Legendre's equation*. Its solutions are used to describe the shapes of the electron 'clouds' in pictures of atoms you see in chemistry books.

Example 137 The equations

$$\begin{aligned}\frac{d^2y}{dt^2} + 2\left(\frac{dy}{dt}\right)^2 + ty &= 0 \\ \frac{d^2y}{dt^2} + \frac{g}{L}\sin y &= 0\end{aligned}$$



are non-linear equations. The second of these is the equation for the motion of a pendulum. This is usually simplified by making the approximation $\sin y \approx y$.

Existence and uniqueness for linear differential equations is a bit easier.

Theorem 7.11 Let $p(t)$, $q(t)$, and $f(t)$ be continuous functions on the interval $t_1 < t < t_2$ and let t_0 be a point in that interval. Then there is a unique solution of

$$y'' + p(t)y' + q(t)y = f(t)$$

on that same interval which satisfies specified initial conditions

$$\begin{aligned} y(t_0) &= y_0 \\ y'(t_0) &= y'_0. \end{aligned}$$

.

Note that the new element in the statement is that the solution is defined and unique on the same interval that the differential equation is defined and continuous on. In the general case, for a non-linear differential equation, we might have to take a smaller interval.

We won't try to prove this theorem in this course.

The basic strategy for solving a linear second order differential equation is the same as the strategy for solving a linear first order differential equation.

1. "(i)" Find a general solution of the homogeneous equation

$$y'' + p(t)y' + q(t)y = 0. \tag{88}$$

Such a solution will involve *two* arbitrary constants.

2. "(ii)" Find one *particular* solution of the inhomogeneous equation

$$y'' + p(t)y' + q(t)y = f(t). \tag{89}$$

Then a general solution of the inhomogeneous equation can be obtained by adding the general solution of the homogeneous equation to the particular solution.

3. "(iii)" Use the initial conditions to determine the constants. Since there are two initial conditions and two constants, we have enough information to do that.

Let's see why this strategy should work. Suppose y_p is one particular solution of the inhomogeneous equation, i.e.,

$$y_p'' + p(t)y_p' + q(t)y_p = f(t).$$

Suppose that y is any *other* solution of the inhomogeneous equation, i.e.,

$$y'' + p(t)y' + q(t)y = f(t).$$

If we subtract the first of these equations from the second and regroup terms, we obtain

$$y'' - y_p'' + p(t)(y' - y_p') + q(t)(y - y_p) = f(t) - f(t) = 0.$$

Thus if we put $u = y - y_p$, we get

$$u'' + p(t)u' + q(t)u = 0,$$

so u is a solution of the homogeneous equation, and

$$y(t) = y_p(t) + u(t).$$

Note how important the linearity was in making this argument work.

Example 138 Consider

$$y'' + 4y = 1 \quad \text{where } y(0) = 0, y'(0) = 1.$$

The homogeneous equation

$$y'' + 4y = 0$$

has the general solution

$$y = C_1 \cos(2t) + C_2 \sin(2t).$$

(This is the equation for simple harmonic motion discussed in the previous section.)

To find a particular solution of the inhomogeneous equation, we note by inspection that a constant solution $y = A$ should work. Then,

$$y'' + 4y = 0 + 4A = 1$$

so $A = 1/4$. Hence, $y_p = 1/4$ is a particular solution. Hence, the general solution of the inhomogeneous equation is

$$y = \underbrace{\frac{1}{4}}_{\text{particular}} + \underbrace{C_1 \cos(2t) + C_2 \sin(2t)}_{\text{general, homogeneous}}.$$

Finally, we determine the constants as follows. First note that

$$y' = -2C_1 \cos 2t + 2C_2 \sin 2t.$$

Then at $t = 0$,

$$\begin{aligned} y = 0 &= \frac{1}{4} + C_1 \cos 0 + C_2 \sin 0 = \frac{1}{4} + C_1 \\ y' = 1 &= -2C_1 \sin 0 + 2C_2 \cos 0 = 2C_2. \end{aligned}$$

Thus, $C_1 = -1/4$ and $C_2 = 1/2$, and the solution matching the given initial conditions is

$$y = \frac{1}{4} - \frac{1}{4} \cos 2t + \frac{1}{2} \sin 2t.$$

Exercises for 7.2.

1. Consider the differential equation $y'' + y = 1$.
 - (a) What is a general solution of the homogeneous equation $y'' + y = 0$?
 - (b) Assume that there is a particular solution of the inhomogeneous equation of the form $y = A$. Find A .
 - (c) Find a general solution of the inhomogeneous equation.
 - (d) Find a solution to the initial value problem $y'' + y = 1$, $y(0) = 1$, $y'(0) = -1$.
2. Consider the differential equation $y'' + 9y = 2 \cos(2t)$.
 - (a) What is a general solution of the homogeneous equation $y'' + 9y = 0$?
 - (b) Assume that there is a particular solution of the inhomogeneous equation of the form $y = A \cos(2t)$. Find A .
 - (c) Find a general solution of the inhomogeneous equation.
 - (d) Find a solution to the initial value problem $y'' + 9y = 2 \cos(2t)$, $y(0) = 1$, $y'(0) = 0$.
3. For which intervals does the existence theorem guarantee a solution for Legendre's equation?

7.3 Homogeneous Second Order Linear Equations

To solve

$$y'' + p(t)y' + q(t)y = 0$$

we need to come up with a solution with two arbitrary constants in it. Suppose that somehow or other, we have found two different solutions $y_1(t)$ and $y_2(t)$ defined on a common t -interval $t_1 < t < t_2$.

Example 139 Two solutions of the equation

$$y'' + 4y = 0$$

are $y_1(t) = \cos(2t)$ and $y_2(t) = \sin(2t)$.

The important insight on which the whole theory is based is that anything of the form

$$y(t) = c_1 y_1(t) + c_2 y_2(t), \quad (90)$$

where c_1 and c_2 are constants, is *again* a solution. To see why, we calculate as follows. We have

$$\begin{aligned} y_1'' + p(t)y_1' + q(t)y_1 &= 0 \\ y_2'' + p(t)y_2' + q(t)y_2 &= 0 \end{aligned}$$

so multiplying the first equation by c_1 , the second by c_2 and adding, we obtain

$$\begin{aligned} c_1 y_1'' + c_2 y_2'' + p(t)(c_1 y_1' + c_2 y_2') + q(t)(c_1 y_1 + c_2 y_2) &= \\ (c_1 y_1 + c_2 y_2)'' + p(t)(c_1 y_1 + c_2 y_2)' + q(t)(c_1 y_1 + c_2 y_2) &= 0, \end{aligned}$$

which says exactly that $y = c_1 y_1 + c_2 y_2$ is a solution. Note that this argument depends very strongly on the fact that the differential equation is linear.

The function $y = c_1 y_1 + c_2 y_2$ is called a *linear combination* of y_1 and y_2 . Thus, we may restate the above statement as follows: *any linear combination of solutions of a linear homogeneous differential equation is again a solution.*

The above analysis provides a method for finding a general solution of the homogeneous equation: find a pair of solutions $y_1(t), y_2(t)$ and use $y = c_1 y_1(t) + c_2 y_2(t)$. However, there is one problem with this; the two solutions might be essentially the same. For example, suppose $y_1 = c y_2$ (i.e., $y_1(t) = c y_2(t)$ for all t in their common domain). Then

$$y = c_1 y_1 + c_2 y_2 = c_1 c y_2 + c_2 y_2 = (c_1 c + c_2) y_2 = c' y_2$$

so the solution does not involve two *arbitrary* constants. Without two such constants, we won't necessarily be able to match arbitrary initial values for $y(t_0)$ and $y'(t_0)$.

With the above discussion in mind we make the following definition. A pair of functions $\{y_1, y_2\}$, defined on a common domain, is called *linearly independent* if neither is a constant multiple of the other. Otherwise, the pair is called *linearly dependent*. There are a couple of subtle points in this definition. First, linear independence (or its negation linear dependence) is a property of the *pair* of functions, not of the functions y_1 and y_2 themselves. Secondly, if $y_1 = c y_2$ then we also have $y_2 = (1/c) y_1$ *except in the case y_1 is identically zero*. In the exceptional case, $c = 0$. Thus, a pair of functions, one of which is identically zero, is always linearly dependent. On the other hand, if $y_1 = c y_2$, and neither y_1 nor y_2 vanishes identically on the domain, then c won't be zero.

Assume now that y_1, y_2 constitute a linearly independent pair of solutions of

$$y'' + p(t)y' + q(t)y = 0,$$

defined on an interval $t_1 < t < t_2$ on which the coefficients $p(t)$ and $q(t)$ are continuous. Suppose t_0 is a point in that interval, and we want to match initial conditions $y(t_0) = y_0$, $y'(t_0) = y'_0$ at t_0 . We shall show that this is always possible with a general solution of the form $y = c_1y_1 + c_2y_2$.

Example 139, continued As above, consider

$$y'' + 4y = 0.$$

with solutions $y_1(t) = \cos 2t$ and $y_2(t) = \sin 2t$. Their quotient is $\tan 2t$, which is not constant, so neither is a constant multiple of the other, and they constitute a linearly independent pair. Let's try to match the initial conditions

$$\begin{aligned} y(\pi/2) &= 1 \\ y'(\pi/2) &= 2. \end{aligned}$$

Let

$$y = c_1y_1(t) + c_2y_2(t) = c_1 \cos 2t + c_2 \sin 2t \quad (91)$$

so

$$y' = c_1y'_1(t) + c_2y'_2(t) = -2c_1 \sin 2t + 2c_2 \cos 2t. \quad (92)$$

Putting $t = \pi/2$, we need to find c_1 and c_2 such that

$$\begin{aligned} y(\pi/2) &= c_1 \cos \pi + c_2 \sin \pi = -c_1 = 1 \\ y'(\pi/2) &= -2c_1 \sin \pi + 2c_2 \cos \pi = -2c_2 = 2. \end{aligned}$$

The solution is $c_1 = -1$, $c_2 = -1$. Thus the solution of the differential equation matching the desired initial conditions is

$$y = -\cos 2t - \sin 2t.$$

(You should check that it works!)

Let's see how this would work in general. Using (91) and (92), matching initial conditions at $t = t_0$ yields

$$\begin{aligned} y(t_0) &= c_1y_1(t_0) + c_2y_2(t_0) = y_0 \\ y'(t_0) &= c_1y'_1(t_0) + c_2y'_2(t_0) = y'_0. \end{aligned}$$

Solve this pair of equations for c_1 and c_2 by the usual method you learned in high school. To find c_1 , multiply the first equation by $y'_2(t_0)$, multiply the second equation by $y_2(t_0)$ and subtract. This yields

$$c_1[y_1(t_0)y'_2(t_0) - y'_1(t_0)y_2(t_0)] = y_0y'_2(t_0) - y'_0y_2(t_0).$$

Hence, provided the coefficient of c_1 is not zero, we obtain

$$c_1 = \frac{y_0y'_2(t_0) - y'_0y_2(t_0)}{y_1(t_0)y'_2(t_0) - y'_1(t_0)y_2(t_0)}.$$

Similarly, multiplying the second equation by $y_1(t_0)$ and the first by $y_1'(t_0)$ and subtracting yields

$$c_2 = \frac{y_0' y_1(t_0) - y_0 y_1'(t_0)}{y_1(t_0) y_2'(t_0) - y_1'(t_0) y_2(t_0)}.$$

Note that the denominators are the same. Also, the above method will work only if this common denominator does not vanish. (In Example 139, the denominator was $(-1)(-2) = 2$.) Define

$$W(t) = y_1(t)y_2'(t) - y_1'(t)y_2(t) = \det \begin{bmatrix} y_1(t) & y_2(t) \\ y_1'(t) & y_2'(t) \end{bmatrix}.$$

This function is called the *Wronskian* of the pair of functions $\{y_1, y_2\}$. Thus, we need to show that the Wronskian $W(t_0) \neq 0$ at the initial point t_0 . To this end, note first that the Wronskian cannot vanish identically for all t in the domain of the differential equation. For,

$$\frac{d}{dt} \frac{y_2(t)}{y_1(t)} = \frac{y_2'(t)y_1(t) - y_2(t)y_1'(t)}{y_1(t)^2} = \frac{W(t)}{y_1(t)^2},$$

so if $W(t)$ vanishes for all t , the quotient y_2/y_1 is constant, and y_2 is a constant multiple of y_1 . That contradicts the linear independence of the pair $\{y_1, y_2\}$. (Actually, the argument is a little more complicated because of the possibility that the denominator y_1 might vanish at some points. See the appendix at the end of this section for the details.)

There is still the possibility that $W(t)$ vanishes at the initial point $t = t_0$, but $W(t)$ does not vanish identically. We shall show that can't ever happen for functions y_1, y_2 which are solutions of the same homogeneous linear differential equation, i.e., the Wronskian is either *never zero* or *always zero* in the domain of the differential equation. To see this, first calculate

$$\begin{aligned} W'(t) &= (y_1(t)y_2'(t) - y_1'(t)y_2(t))' \\ &= y_1'(t)y_2'(t) + y_1(t)y_2''(t) - (y_1''(t)y_2(t) + y_1'(t)y_2'(t)) \\ &= y_1(t)y_2''(t) - y_1''(t)y_2(t). \end{aligned}$$

On the other hand, using the differential equation, we can express the second derivatives of the solutions

$$\begin{aligned} y_1''(t) &= -p(t)y_1'(t) - q(t)y_1(t) \\ y_2''(t) &= -p(t)y_2'(t) - q(t)y_2(t). \end{aligned}$$

Putting these in the above formula yields

$$\begin{aligned} W'(t) &= y_1(t)(-p(t)y_2'(t) - q(t)y_2(t)) - (-p(t)y_1'(t) - q(t)y_1(t))y_2(t) \\ &= -p(t)(y_1(t)y_2'(t) - y_1'(t)y_2(t)) = -p(t)W(t). \end{aligned}$$

This shows that the Wronskian satisfies the first order differential equation

$$\frac{dW}{dt} = -p(t)W,$$

and we know how to solve such equations. The general solution is

$$W(t) = Ce^{-\int p(t)dt}. \quad (93)$$

The important thing about this formula is that the exponential function never vanishes. Hence, the only way $W(t)$ can vanish for any t whatsoever is if $C = 0$, in which case $W(t)$ vanishes identically.

We summarize the above discussion as follows. *A pair $\{y_1, y_2\}$ of solutions of the homogeneous linear equation*

$$y'' + p(t)y' + q(t)y = 0$$

is linearly independent if and only if the Wronskian never vanishes.

Example 139, again The Wronskian of the pair $\{y_1 = \cos 2t, y_2 = \sin 2t\}$ is

$$\det \begin{bmatrix} \cos 2t & \sin 2t \\ -2 \sin 2t & 2 \cos 2t \end{bmatrix} = 2 \cos^2 2t + 2 \sin^2 2t = 2.$$

According to the theory, it is not really necessary to find $W(t)$ for all t . It would have sufficed to find it just at $t_0 = \pi/2$, as in essence we did before, and see that it is not zero.

It sometimes seems a bit silly to calculate the Wronskian to see if a pair of solutions is independent since one feels it should be obvious whether or not one function is a multiple of another. However, for complicated functions, it may not be so obvious. For example, for the functions $y_1(t) = \sin 2t$ and $y_2(t) = \sin t \cos t$ it might not be clear that the first is twice the second if we did not know the trigonometric identity $\sin 2t = 2 \sin t \cos t$. For more complicated functions, there may be all sorts of hidden relations we just don't know.

Example 140 Consider the equation

$$y'' + \frac{2t}{1-t^2}y' + \frac{6}{1-t^2}y = 0 \quad -1 < t < 1.$$

To find the form of the Wronskian, calculate

$$\int p(t) dt = \int \frac{2t}{1-t^2} dt = -\ln(1-t^2).$$

Thus,

$$W(t) = Ce^{\ln(1-t^2)} = C(1-t^2).$$

Warning: We have remarked that the Wronskian never vanishes. That conclusion is valid only on intervals for which the coefficient function $p(t)$ is continuous. If $p(t)$

has a singularity, it is quite possible to have the antiderivative $\int p(t)dt$ approach ∞ as t approaches the singularity. In that case the exponential

$$e^{-\int p(t)dt}$$

would approach 0 as t approaches the singularity. That is the case in the previous example at $t = 1$ and $t = -1$ which are singularities of $p(t) = 2t/(1 - t^2)$. Hence, the fact that $W(t) = C(1 - t^2)$ vanishes at those points does not contradict the validity of the general theory.

An alternate form of the formula for the Wronskian using definite integrals with dummy variables is sometimes useful.

$$W(t) = W(t_0)e^{-\int_{t_0}^t p(s)ds}. \quad (94)$$

Appendix on the Vanishing of the Wronskian Suppose the Wronskian $W(t) = y_1(t)y_2'(t) - y_1'(t)y_2(t)$ vanishes identically, but $y_1(t_1) = 0$ for some specific value t_1 in the interval where y_1 and y_2 are defined. Then the argument showing $y_2(t)/y_1(t)$ is constant fails because there will be a zero in the denominator when applying the quotient rule. Let's see what we can do in that case. $y_1'(t_1) \neq 0$ because otherwise

$$\begin{aligned} y_1(t_1) &= z(t_1) = 0 \\ y_1'(t_1) &= z'(t_1) = 0, \end{aligned}$$

where $z(t)$ is the function which is identically zero for all t . By the uniqueness theorem, that would mean that y_1 is identically zero, which is not the case. Hence, using the fact that $y_1'(t_1) \neq 0$, we can conclude from

$$W(t_1) = y_1(t_1)y_2'(t_1) - y_1'(t_1)y_2(t_1) = -y_1'(t_1)y_2(t_1) = 0$$

that $y_2(t_1) = 0$. Thus, by the same reasoning $y_2'(t_1) \neq 0$. Let $c = y_2'(t_1)/y_1'(t_1)$. Then

$$\begin{aligned} y_2(t_1) &= cy_1(t_1) = 0 \\ y_2'(t_1) &= cy_1'(t_1) \quad \text{by the definition of } c \end{aligned}$$

Hence, by the uniqueness theorem, $y_1(t) = cy_2(t)$ for all t in the common interval on which the solutions are defined.

Note that this is actually quite subtle. In case one of the two functions never vanishes, the quotient rule suffices to show that the identical vanishing of the Wronskian implies the pair of functions is linearly dependent. However, if both functions vanish at some points, we must use the fact that they are both solutions of a homogeneous linear differential equation and apply the basic uniqueness theorem! Solutions of such equations frequently vanish at isolated points so the subtle part of the argument is necessary.

Exercises for 7.3.

1. The equation $y'' + 6y' + 9y = 0$ has solutions $y_1 = e^{-3t}$ and $y_2 = te^{-3t}$. Find a solution $y(t)$ satisfying $y(0) = 1, y'(0) = 0$.
2. Consider the differential equation $t^2y'' + 4ty' + 2y = 0$.
 - (a) Show that $y_1(t) = \frac{1}{t}$ and $y_2(t) = \frac{1}{t^2}$ are solutions for $0 < t < \infty$.
 - (b) Show that the pair of solutions $\{y_1, y_2\}$ is linearly independent.
 - (c) Compute the Wronskian $W(t)$ for this pair of solutions and note that it doesn't vanish for $0 < t < \infty$.
 - (d) What happens to the Wronskian $W(t)$ as $t \rightarrow 0$?
3. For each of the following pairs of functions, calculate the Wronskian $W(t)$ on the interval $-\infty < t < \infty$. In each case, see if you have enough information to decide whether the pair can be a linearly independent pair of solutions of a second order linear homogeneous differential equation on the interval $-\infty < t < \infty$.
 - (a) $\{\sin 3t, \cos 3t\}$.
 - (b) $\{e^{2t}, e^{-t}\}$.
 - (c) $\{t^2 - t, 3t\}$.
 - (d) $\{te^{-t}, e^{-t}\}$.
4. Suppose $\{y_1(t), y_2(t)\}$ is a linearly independent pair of solutions of

$$y'' + \frac{2t}{1+t^2}y' + \frac{1}{1+t^2}y = 0$$

defined on the interval $-1 < t < 1$. Suppose $y_1(0) = 1, y_1'(0) = 2, y_2(0) = 2, y_2'(0) = 1$. Find $W(t)$. Is the pair of solutions linearly independent?

5. Suppose $\{u_1, u_2\}$ is a linearly independent pair of solutions of a second order linear homogeneous differential equation. Let $y_1(t) = u_1(t) + u_2(t)$ and $y_2(t) = u_1(t) - u_2(t)$.
 - (a) Show that the Wronskian $W_y(t)$ of the pair $\{y_1, y_2\}$ is related to the Wronskian $W_u(t)$ of the pair $\{u_1, u_2\}$ by $W_y(t) = -2W_u(t)$. Conclude that $\{y_1, y_2\}$ is also a linearly independent pair of solutions.
 - (b) Show that $\{y_1, y_2\}$ is a linearly independent pair of solutions without using the Wronskian. Hint: Assume $y_1 = cy_2$ and derive a similar equation for u_1 and u_2 .

7.4 Homogeneous Equations with Constant Coefficients

Consider the differential equation

$$y'' + py' + qy = 0$$

where p and q are *constants*. In this case, we can solve the homogeneous equation completely. Since the differential equations arising in some important applications fall in this category, this is quite fortunate. For example, as indicated in Section 1, the equation for damped harmonic motion has constant coefficients.

The essential idea behind the solution method is to search for solutions of the form

$$y = e^{rt}$$

where r is a constant to be determined. This seems a bit arbitrary, but we rely here on the experience of generations of mathematicians who have worked with differential equations. Thus, we take advantage of their discoveries, so we don't have to rediscover everything ourselves.

We have

$$\begin{aligned} y &= e^{rt} \\ y' &= re^{rt} \\ y'' &= r^2e^{rt} \end{aligned}$$

so

$$y'' + py' + qy = r^2e^{rt} + pre^{rt} + qe^{rt} = (r^2 + pr + q)e^{rt}.$$

Since e^{rt} never vanishes, it follows that $y = e^{rt}$ is a solution of the differential equation if and only if

$$r^2 + pr + q = 0. \tag{95}$$

This converts the problem of solving a differential equation into a problem of solving an algebraic equation of the same order. The roots of equation (95) are

$$\begin{aligned} r_1 &= -\frac{p}{2} + \frac{1}{2}\sqrt{p^2 - 4q}, \\ r_2 &= -\frac{p}{2} - \frac{1}{2}\sqrt{p^2 - 4q}. \end{aligned}$$

Note that these will be different as long as $p^2 - 4q \neq 0$. Corresponding to these roots are the two solutions of the differential equation: $y_1 = e^{r_1t}$ and $y_2 = e^{r_2t}$. Moreover, if $r_1 \neq r_2$, then linear independence is no problem since the ratio $e^{r_1t}/e^{r_2t} = e^{(r_1-r_2)t}$ is not constant. Then,

$$y = c_1e^{r_1t} + c_2e^{r_2t}$$

is a general solution.

Example 141 Consider

$$y'' + 3y' - 4y = 0.$$

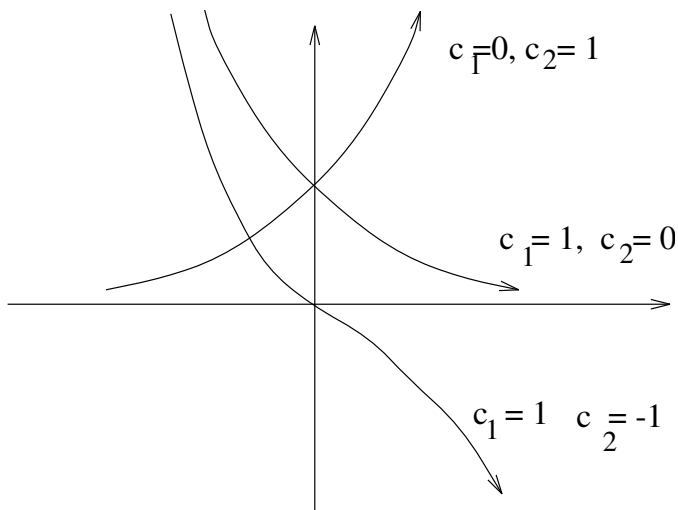
The corresponding algebraic equation is

$$r^2 + 3r - 4 = 0.$$

The roots of this equation are easy to determine by factoring: $r^2 + 3r - 4 = (r + 4)(r - 1)$. They are $r_1 = -4, r_2 = 1$. Hence, the general solution is

$$y = c_1 e^{-4t} + c_2 e^t.$$

The exact shape of the graph of such a solution will depend on the constants c_1 and c_2 .



Suppose the roots of the equation $r^2 + pr + q = 0$ are equal, i.e., $p^2 - 4q = 0$. Then, by the quadratic formula, $r = r_1 = r_2 = -p/2$. The method only gives *one solution* $y_1 = e^{rt}$. Hence, we need to find an additional independent solution. The trick is to know that in this case $y = te^{rt}$ is another solution. For,

$$\begin{aligned} y &= te^{rt} \\ y' &= e^{rt} + tre^{rt} \\ y'' &= re^{rt} + re^{rt} + tr^2e^{rt} = r^2te^{rt} + 2re^{rt}, \end{aligned}$$

so

$$\begin{aligned} y'' + py' + qy &= (r^2t + 2r + pt + p + qt)e^{rt} \\ &= [(r^2 + pr + q)t + 2r + p]e^{rt} = 0. \end{aligned}$$

(Note that these calculations only work in the case $r = -p/2$ is a *double root* of the quadratic equation $r^2 + pr + q = 0$.) The general solution is

$$y = c_1 e^{rt} + c_2 t e^{rt} = (c_1 + c_2 t) e^{rt}.$$

Example 142 Consider

$$y'' + 4y' + 4y = 0.$$

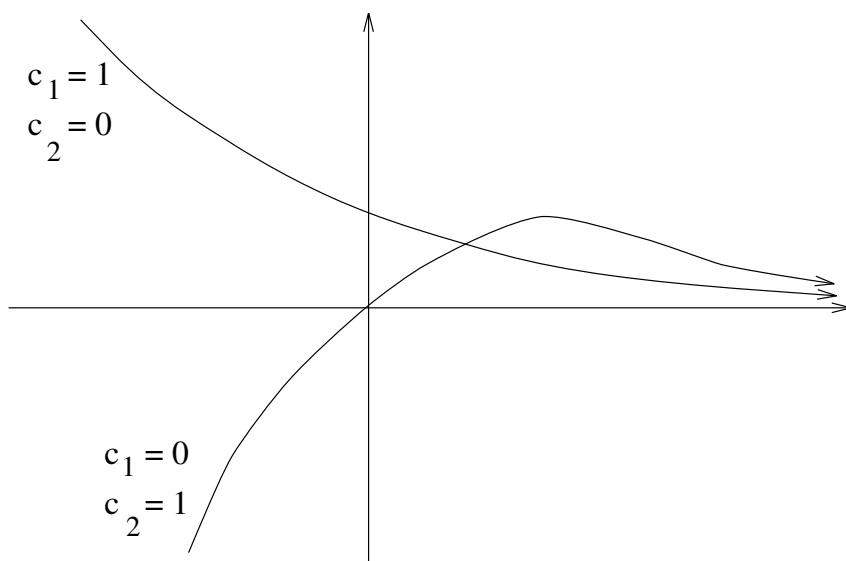
The corresponding algebraic equation is

$$r^2 + 4r + 4 = (r + 2)^2 = 0$$

so $r = -2$ is a double root. Hence, $y_1 = e^{-2t}$ and $y_2 = t e^{-2t}$ constitute an independent pair of solutions, and

$$y = c_1 e^{-2t} + c_2 t e^{-2t} = (c_1 + c_2 t) e^{-2t}$$

is the general solution.



There is one problem with the method in the case of unequal roots r_1, r_2 . Namely, if $p^2 - 4q < 0$, the solutions will be *complex numbers*.

Example 143 Consider

$$y'' + y' + y = 0.$$

The algebraic equation is

$$r^2 + r + 1 = 0$$

and its roots are

$$r_1 = -\frac{1}{2} + \frac{1}{2}\sqrt{1^2 - 3} = -\frac{1}{2} + \frac{1}{2}\sqrt{3}i$$

$$r_2 = -\frac{1}{2} - \frac{1}{2}\sqrt{3}i$$

where $i^2 = -1$. This would suggest that the two basic solutions should be

$$y_1 = e^{(-1/2 + \sqrt{3}i/2)t} \quad \text{and} \quad y_2 = e^{(-1/2 - \sqrt{3}i/2)t}.$$

Unfortunately, you probably have never seen complex exponentials, so we shall make a detour to review complex numbers and talk about their exponentials.

Exercises for 7.4.

- Find a general solution for each of the following differential equations.
 - $y'' - 4y = 0$.
 - $y'' - 5y' + 6y = 0$.
 - $4y'' - 2y' - 2y = 0$.
 - $y'' + 6y' + 9y = 0$.
 - $4y'' - 4y' - y = 0$.
- Solve each of the following initial value problems.
 - $y'' - 2y' - 3y = 0$ given $y(0) = 0, y'(0) = 2$.
 - $y'' - 4y' + 4y = 0$ given $y(0) = 0, y'(0) = -1$.
- Calculate the Wronskian of the pair $\{e^{r_1 t}, e^{r_2 t}\}$. Show that it is never zero as long as $r_1 \neq r_2$. Apply a similar analysis to the pair $\{e^{rt}, te^{rt}\}$.
- Suppose $r_1 \neq r_2$ are two real numbers. Investigate circumstances under which $y = C_1 e^{r_1 t} + C_2 e^{r_2 t}$ can be zero. Is it possible that it never vanishes? What is the maximum number of values of t for which such a function can vanish if the constants C_1 and C_2 are not zero?
- Show that $y = C_1 e^{rt} + C_2 t e^{rt}$ vanishes for at most one value of t if C_1 and C_2 are not both zero.
- The differential equation $t^2 y'' + \alpha t y' + \beta y = 0$ is called *Euler's Equation*.
 - Show that $y = t^r$ is a solution if and only if r is a root of the quadratic equation $r^2 + (\alpha - 1)r + \beta = 0$.
 - Show that if the quadratic equation has two different real roots r_1 and r_2 , then $\{t^{r_1}, t^{r_2}\}$ is a linearly independent pair of solutions defined on the interval $0 < t < \infty$. Hence, in this case the general solution of Euler's equation has the form $y = C_1 t^{r_1} + C_2 t^{r_2}$.
 - Find a general solution of $t^2 y'' - 4t y' + 6y = 0$.

7.5 Complex Numbers

A complex number is an expression of the form

$$\alpha = a + bi$$

where a and b are real numbers. a is called the *real part* of α , and is often denoted $\operatorname{Re}(\alpha)$. b is called the *imaginary part* of α , and it is often denoted $\operatorname{Im}(\alpha)$. The set of all complex numbers is usually denoted \mathbf{C} . Complex numbers are added, subtracted, and multiplied by the usual rules of algebra with the additional rule $i^2 = -1$. Here are some examples of such calculations.

$$\begin{aligned} a + bi + c + di &= a + c + (b + d)i, \\ (a + bi)(c + di) &= ac + adi + bci + bdi^2 = ac - bd + (ad + bd)i. \end{aligned}$$

Complex numbers may be represented geometrically by points (or vectors) in \mathbf{R}^2 : the point (a, b) corresponds to the complex number $\alpha = a + bi$. The horizontal axis is called the *real axis* because all numbers of the form $a = a + 0i$ correspond to points $(a, 0)$ on it. Similarly, the vertical axis is called the *imaginary axis* because all numbers of the form ib correspond to points $(0, b)$ on it. This geometric picture of complex numbers is sometimes called the *Argand diagram*. The length of the vector $\langle a, b \rangle$ is called the *modulus* or *absolute value* of the complex number and is denoted $|\alpha|$. Thus,

$$|\alpha| = \sqrt{a^2 + b^2}.$$

Similarly, the angle θ that the vector makes with the real axis is called the *argument* of α . Of course, the modulus and argument of a complex number are just the polar coordinates of the corresponding point in \mathbf{R}^2 .

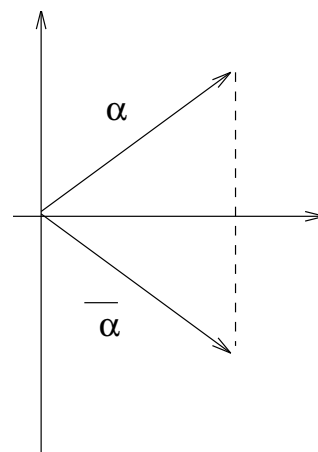
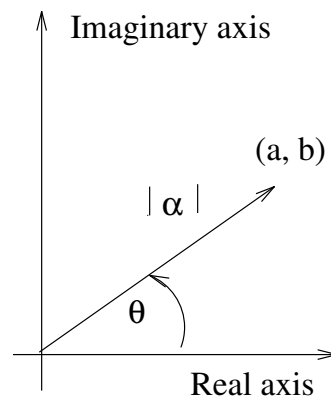
If $\alpha = a + bi$ is a complex number, $\bar{\alpha} = a - bi$ is called its *complex conjugate*. Geometrically, the complex conjugate of α is obtained by reflecting α in the real axis. Note that the two solutions of the quadratic equation

$$r^2 + pr + q = 0$$

are conjugate complex numbers in the case $p^2 - 4q < 0$.

The following rules apply for complex conjugation.

1. $\overline{\alpha + \beta} = \bar{\alpha} + \bar{\beta}$.
2. $\overline{\alpha\beta} = \bar{\alpha}\bar{\beta}$.
3. $|\alpha|^2 = \alpha\bar{\alpha}$.



The proofs of these rules are done by calculating both sides and checking that they give the same result. For example,

$$|\alpha|^2 = a^2 + b^2 \quad \text{and} \\ \alpha\bar{\alpha} = (a + bi)(a - bi) = a^2 - abi + abi - b^2i^2 = a^2 + b^2.$$

Complex numbers may also be divided. Thus, if $\alpha = a + bi$ and $\beta = c + di \neq 0$, then

$$\begin{aligned} \frac{\alpha}{\beta} &= \frac{\alpha\bar{\beta}}{\beta\bar{\beta}} = \frac{\alpha\bar{\beta}}{|\beta|^2} \\ &= \frac{(a + bi)(c - di)}{c^2 + d^2} = \frac{ac + bd}{c^2 + d^2} + \frac{bc - ad}{c^2 + d^2}i. \end{aligned}$$

For example,

$$\frac{1}{i} = \frac{1-i}{i-i} = \frac{-i}{1} = -i.$$

(That is also clear from $i^2 = -1$.)

The next problem is to make an appropriate definition for e^α where α is a complex number. We will use this to consider possible solutions of a differential equation of the form $u(t) = e^{\rho t}$ where ρ is complex and t is a real variable. Such a function u has domain a t -interval on the real line and takes complex values. That is denoted schematically by $u : \mathbf{R} \rightarrow \mathbf{C}$. Since we may identify \mathbf{C} with \mathbf{R}^2 by means of the Argand diagram, this is not really anything new. We suppose that we have available all the usual tools we need for such functions, i.e., differentiation, integration, etc.

Here are some properties we expect the complex exponential to have.

1. "(a)" $e^{\alpha+\beta} = e^\alpha e^\beta$.
2. "(b)" $\frac{d}{dt}e^{\alpha t} = \alpha e^{\alpha t}$.
3. "(c)" It should agree with the ordinary exponential for α real.

Let's apply these rules and see how far they take us in determining a possible definition. First, let $\alpha = a + bi$. Then by (a)

$$e^\alpha = e^{a+bi} = e^a e^{bi}.$$

Since by (c) we already know what e^a is, we need only define e^{bi} . Suppose, as a function of the real variable b

$$e^{ib} = c(b) + is(b)$$

where $c(b)$ and $s(b)$ denote real valued functions. Since $e^{i0} = e^0 = 1$, we know that the functions $c(b)$ and $s(b)$ must satisfy

$$c(0) = 1 \quad s(0) = 0.$$

Also, by (b), we must have

$$\begin{aligned} \frac{d}{db} e^{ib} &= i e^{ib} \\ \text{or} \quad c'(b) + i s'(b) &= i(c(b) + i s(b)) = -s(b) + i c(b). \end{aligned}$$

Comparing real and imaginary parts yields

$$\begin{aligned} c'(b) &= -s(b) \\ s'(b) &= c(b). \end{aligned}$$

It is clear how to choose functions with these properties:

$$\begin{aligned} c(b) &= \cos b \\ s(b) &= \sin b, \end{aligned}$$

so the proper choice for the definition of e^{ib} is

$$e^{ib} = \cos b + i \sin b,$$

and the proper definition of e^α with $\alpha = a + bi$ is

$$e^{a+bi} = e^a \cos b + i e^a \sin b. \tag{96}$$

It is not hard to check from the definition (96) that properties (a), (b), and (c) are true.

Also, this exponential has the property that e^α never vanishes for any complex number α . For,

$$e^{a+bi} = e^a \cos b + i e^a \sin b = 0$$

only if its real and imaginary parts vanish, i.e., $e^a \cos b = e^a \sin b = 0$. Since e^a never vanishes, this can happen only if $\cos b = \sin b = 0$. However, the cosine function and the sine function don't have any common vanishing points, so there is no such b .

Finally, here are some useful formulas we shall use repeatedly.

$$\frac{1}{e^{ib}} = e^{-bi} = \cos(-b) + i \sin(-b) = \cos b - i \sin b$$

which tells us that the inverse of e^{ib} is the same as the complex conjugate. This may also be seen from the fact that the product of e^{bi} and its conjugate is $|e^{ib}|^2$, since

$$|e^{ib}|^2 = \cos^2 b + \sin^2 b = 1.$$

Exercises for 7.5.

- Calculate each of the following quantities
 - $(1 + 3i)(2 - 5i)$.
 - $(1 + 2i)(1 + i) - 3$.
 - $\frac{2 + i}{3 - 2i}$.
 - $(1 + i)^4$.
- Show that $\alpha = e^{i(2\pi/3)}$ satisfies the equation $\alpha^3 = 1$. There are two other complex numbers satisfying that equation. What are they? Draw the Argand diagrams for each of these numbers.
- Note that $e^{i\pi} = \cos \pi + i \sin \pi = -1$. Using this information, find all complex numbers β satisfying $\beta^3 = -1$. Hint. They are all of the form $e^{i\theta}$ for appropriate values of θ .
- Show that $|e^{ib}| = 1$ for any real number b .
 - Show that $\alpha = |\alpha|e^{i\theta}$ for any complex number α . Hint: Consider the polar coordinates of the point in the complex plane corresponding to α .
- Using the definition

$$e^{a+bi} = e^a \cos b + ie^a \sin b$$

prove the formula

$$\frac{d}{dt}e^{(a+bi)t} = (a + bi)e^{(a+bi)t}.$$

- Prove the formula

$$e^{ix+iy} = e^{ix}e^{iy}$$

for x, y real. This is a special case of the law of exponents for complex exponentials. To prove it you will have to use appropriate trigonometric identities.

- Show that there is only one possible choice for real valued functions $c(b), s(b)$ such that $c'(b) = -s(b)$, $s'(b) = c(b)$ and $c(0) = 1, s(0) = 0$. Hint: If $c_1(b), s_1(b)$ is another set of functions satisfying these conditions, then put $u(b) = c(b) - c_1(b), v(b) = s(b) - s_1(b)$. Show that $u'(b) = -v(b), v'(b) = u(b)$ and $u(0) = v(0) = 0$. Then show that $u^2 + v^2$ is constant by taking its derivative. What is that constant, and can you conclude that $u = v = 0$?

7.6 Complex Solutions of a Differential Equation

In our previous discussion of the differential equation

$$y'' + py' + qy = 0 \tag{97}$$

we considered solutions $y = y(t)$ which were functions $\mathbf{R} \rightarrow \mathbf{R}$. However, our analysis of the method of solution suggests that we try to extend this by looking for solutions which are functions $\mathbf{R} \rightarrow \mathbf{C}$. Any such function may be expressed

$$y = y(t) = u(t) + iv(t) \quad (98)$$

where the real and imaginary parts $u(t)$ and $v(t)$ define real valued functions. Putting (98) in (97) yields

$$\begin{aligned} &u'' + iv'' + p(u' + iv') + q(u + iv) = 0 \\ \text{or} \quad &u'' + pu' + qu + i(v'' + pv' + qv) = 0 \\ \text{or} \quad &u'' + pu' + qu = 0 \quad \text{and} \quad v'' + pv' + qv = 0. \end{aligned}$$

Thus, a single complex solution really amounts to a *pair* of real valued solutions. From this perspective, the previous theory of purely real valued solutions appears as the special case where the imaginary part is zero, i.e., $y = u(t) + i0 = u(t)$. Also, all the rules of algebra and calculus still apply for the more general functions, so we may proceed just as before except that we have the advantages of using complex algebra. The only tricky point is to remember that all constants in the extended theory are potentially *complex* numbers whereas previously they were real. In particular, if

$$r^2 + pr + q = 0$$

has two conjugate complex roots r_1, r_2 (which is the case if $p^2 - 4q < 0$), then $y_1 = e^{r_1 t}$ and $y_2 = e^{r_2 t}$ form a linearly independent pair of functions, and a general (complex) solution has the form

$$y = c_1 e^{r_1 t} + c_2 e^{r_2 t},$$

where c_1, c_2 are arbitrary *complex* constants.

Example 144 Consider

$$y'' + 4y = 0 \quad \text{where } y(\pi/2) = 1, y'(\pi/2) = 2.$$

First we solve

$$r^2 + 4 = 0$$

to obtain the two conjugate complex roots $r_1 = 2i, r_2 = -2i$. The general solution is

$$y = c_1 e^{2it} + c_2 e^{-2it}.$$

To match the given initial conditions, note that $y' = 2ic_1 e^{2it} - 2ic_2 e^{-2it}$. At $t = \pi/2$, we have

$$e^{i\pi} = e^{-i\pi} = -1,$$

so

$$\begin{aligned} y(\pi/2) &= c_1 e^{\pi i} + c_2 e^{-\pi i} = -c_1 - c_2 = 1 \\ y'(\pi/2) &= 2ic_1 e^{\pi i} - 2ic_2 e^{-\pi i} = -2ic_1 + 2ic_2 = 2. \end{aligned}$$

Multiply the first equation by $2i$ and add to obtain

$$\begin{aligned} -4ic_1 &= 2i + 2 \\ \text{or} \quad c_1 &= -\frac{1+i}{2i} = -\frac{1-i}{2}. \end{aligned}$$

Here we used

$$\frac{1+i}{i} = \frac{1}{i} + 1 = -i + 1.$$

Similarly, subtraction yields

$$\begin{aligned} -4ic_2 &= 2i - 2 \\ \text{or} \quad c_2 &= -\frac{i-1}{2i} = -\frac{1+i}{2}. \end{aligned}$$

Hence, the solution matching the given initial conditions is

$$y = -\frac{1}{2}[(1-i)e^{2it} + (1+i)e^{-2it}].$$

If you recall, we solved this same problem with real valued functions earlier, and this answer does not look at all the same. However, if we expand the exponentials, we get

$$\begin{aligned} (1-i)e^{2it} &= (1-i)(\cos 2t + i \sin 2t) = \cos 2t + \sin 2t + i(\sin 2t - \cos 2t) \\ (1+i)e^{-2it} &= (1+i)(\cos 2t - i \sin 2t) = \cos 2t + \sin 2t + i(\cos 2t - \sin 2t). \end{aligned}$$

Hence, adding these and multiplying by $-1/2$ yields

$$y = -\cos 2t - \sin 2t$$

which is indeed the same solution obtained before. Notice that the solution is entirely real—its imaginary part is zero—although initially it did not look that way.

The fact that we ended up with a real solution in the above example is not an accident. That will always be the case when p, q and the initial values y_0 and y'_0 are real. The reason is that if we write the solution

$$y = u(t) + iv(t),$$

then the imaginary part $v(t)$ satisfies the initial conditions $v(t_0) = 0, v'(t_0) = 0$, so, by the uniqueness theorem, it must be identically zero.

It is quite common in applications to use complex exponentials to describe oscillatory phenomena. For example, electrical engineers have for generations preferred to represent alternating currents, voltages, and impedances by complex quantities. However, it is sometimes easier to work with real functions. Fortunately, there is a simple way to convert. If $r = a + bi$ is one of a pair of complex conjugate roots of the equation

$$r^2 + pr + q = 0$$

then

$$e^{rt} = e^{at+ibt} = e^{at} \cos bt + ie^{at} \sin bt$$

is a solution, and

$$y_1 = e^{at} \cos bt \quad \text{and} \quad y_2 = e^{at} \sin bt$$

are two real solutions. They form a linearly independent pair since their quotient is not constant. Hence,

$$y = c_1 e^{at} \cos bt + c_2 e^{at} \sin bt$$

(where the constants c_1 and c_2 are real) is a general real solution. Note that in this analysis, we only used one of the two roots (and the corresponding e^{rt}). However, the other root is the complex conjugate $\bar{r} = a - bi$, so it yields real solutions

$$e^{at} \cos(-bt) = e^{at} \cos(bt) \quad \text{and} \quad e^{at} \sin(-bt) = -e^{at} \sin(bt)$$

which are the same functions except for sign.

Example 144, again The roots are $2i$ and $-2i$. So taking $a = 0, b = 2$ yields the solutions

$$e^{0t} \cos 2t = \cos 2t \quad \text{and} \quad e^{0t} \sin 2t$$

which confirms what we already know.

Exercises for 7.6.

- Find a general complex solution for each of the following differential equations.
 - $y'' + y' + y = 0$.
 - $y'' + 2y' + 2y = 0$.
 - $2y'' - 3y' + 2y = 0$.
- Find a general real solution for each of the differential equations in the previous problem.
- Solve the initial value problem

$$y'' + 9y = 0 \quad \text{given} \quad y(0) = 1, y'(0) = -1.$$

Solve the problem two ways. (a) Find a general complex solution and determine complex constants to match the initial conditions. (b) Find a general real solution and determine real constants to match the initial conditions. (c) Verify that the two different solutions are the same.

4. Solve the initial value problem

$$y'' - 2y' + 5y = 0 \quad \text{given} \quad y(0) = 1, y'(0) = -1.$$

Solve the problem two ways. (a) Find a general complex solution and determine complex constants to match the initial conditions. (b) Find a general real solution and determine real constants to match the initial conditions. (c) Verify that the two different solutions are the same.

5. Calculate the Wronskian of the pair of functions $\{e^{at} \cos(bt), e^{at} \sin(bt)\}$. Show that it never vanishes.

7.7 Oscillations

The differential equation governing the motion of a mass at the end of spring may be written

$$m \frac{d^2 y}{dt^2} + b \frac{dy}{dt} + ky = 0, \quad (99)$$

where $m > 0$ is the mass, $b \geq 0$ represents a coefficient of friction, and $k > 0$ is the spring constant. The equation governing the charge Q on a capacitor in a simple oscillatory circuit is

$$L \frac{d^2 Q}{dt^2} + R \frac{dQ}{dt} + \frac{1}{C} Q = 0, \quad (100)$$

where L is the inductance, R is the resistance, and C is the capacitance in the circuit. Both these equations are of the form $y'' + py' + qy = 0$ with $p \geq 0$ and $q > 0$, and we now have the tools in hand to completely solve that equation. We shall do that in the context of (99) for mechanical oscillations, but the theory applies equally well to electrical oscillations. Hence, we may take $p = b/m$ and $q = k/m$, but many of the formulas are a bit neater if we multiply everything through by m as in (99).

As we saw in the previous sections, the first step is to solve the quadratic equation

$$mr^2 + br + k = 0.$$

There are three cases:

1. "(a)" $b^2 - 4km > 0$. The roots are real and unequal.
2. "(b)" $b^2 - 4km = 0$. The roots are real and equal.
3. "(c)" $b^2 - 4km < 0$. The roots are complex conjugates and unequal.

We treat each in turn.

Case (a). Assume $b^2 > 4km$, and put $\Delta = \sqrt{b^2 - 4km}$. Then according to the quadratic formula, the roots are

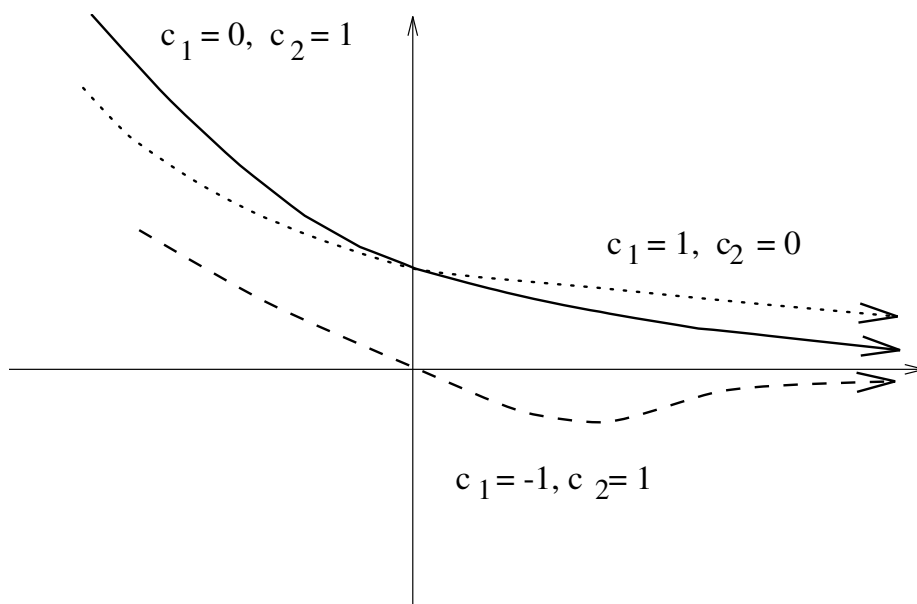
$$r_1 = \frac{-b + \Delta}{2m}$$

$$r_2 = \frac{-b - \Delta}{2m},$$

so the general solution is

$$y = c_1 e^{\frac{1}{2m}(-b+\Delta)t} + c_2 e^{\frac{1}{2m}(-b-\Delta)t}.$$

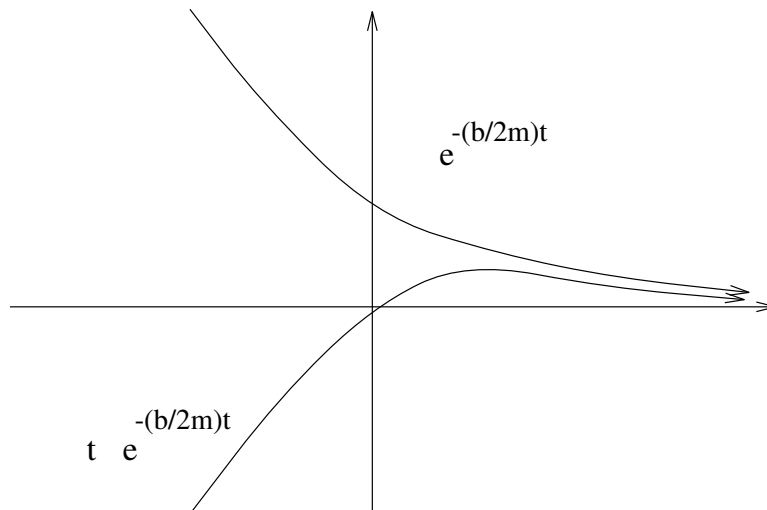
Some typical solutions for different values of the constants are indicated in the diagram.



Note that since $\Delta = \sqrt{b^2 - 4km} < \sqrt{b^2} = b$, both roots are negative. Hence, both exponentials approach zero as $t \rightarrow \infty$. If the signs of the constants differ, then the solution will vanish for precisely one value of t , but otherwise it will never vanish. Thus, the solution dies out without actually oscillating. This is called the *overdamped case*. *Case (b).* Suppose $b^2 = 4km$. Then the quadratic equation has two equal roots $r = r_1 = r_2 = -b/2m$. The general solution is

$$y = c_1 e^{-\frac{b}{2m}t} + c_2 t e^{-\frac{b}{2m}t} = (c_1 + c_2 t) e^{-\frac{b}{2m}t}.$$

This solution also vanishes for exactly one value of t and approaches zero as $t \rightarrow \infty$. This is called the *critically damped case*.



Case (c). Suppose $b^2 < 4km$. The roots are

$$r = \frac{-b + \sqrt{b^2 - 4km}}{2m} = -\frac{b}{2m} + i \frac{\sqrt{4km - b^2}}{2m}$$

and its complex conjugate \bar{r} . Put

$$\omega_1 = \frac{\sqrt{4km - b^2}}{2m} = \sqrt{\frac{k}{m} - \left(\frac{b}{2m}\right)^2}.$$

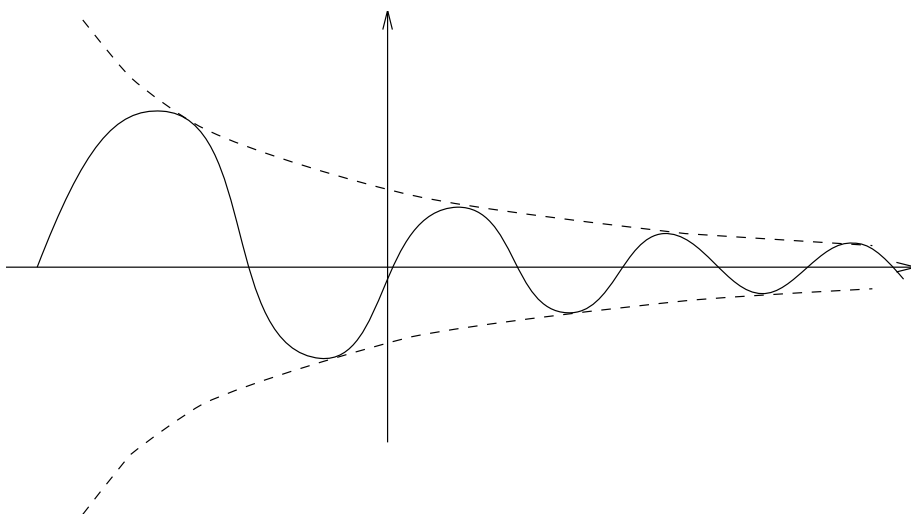
We may obtain two linearly independent real solutions by taking the real and imaginary parts of

$$e^{rt} = e^{(-\frac{b}{2m} + i\omega_1)t} = e^{-\frac{b}{2m}t}(\cos \omega_1 t + i \sin \omega_1 t).$$

These are $y_1 = e^{-\frac{b}{2m}t} \cos \omega_1 t$ and $y_2 = e^{-\frac{b}{2m}t} \sin \omega_1 t$ so the general real solution is

$$\begin{aligned} y &= c_1 e^{-\frac{b}{2m}t} \cos \omega_1 t + c_2 e^{-\frac{b}{2m}t} \sin \omega_1 t \\ &= e^{-\frac{b}{2m}t} (c_1 \cos \omega_1 t + c_2 \sin \omega_1 t). \end{aligned}$$

The expression in parentheses oscillates with angular frequency ω_1 while the exponential approaches zero as $t \rightarrow \infty$. This is called the *underdamped case*.



If $b = 0$, then the system will oscillate indefinitely with angular frequency $\omega_0 = \sqrt{k/m}$. $\omega_0/2\pi$ is called the resonant frequency of the system, even in the case $b \neq 0$. If k/m is much larger than $(b/2m)^2$, then $\omega_0 \approx \omega_1$ and the system will oscillate for quite a long time before noticeably dying out. See the Exercises for more discussion of these points.

Exercises for 7.7.

1. If m is measured in grams, y in centimeters, and t in seconds, then the spring constant k should be measured in gm/sec^2 and the coefficient of friction b in gm/sec . Assuming we use these units, determine in each of the following cases if the system oscillates. If so determine the frequencies $\frac{\omega_0}{2\pi}$ and $\frac{\omega_1}{2\pi}$.

(a) $m = 1, k = 0.5, b = 0.005$.

(b) $m = 4, k = 1, b = 4$.

2. Show that the three cases for the differential equation governing the charge on a capacitor break up as follows. (a) Overdamped: $R > 2\sqrt{\frac{L}{C}}$. (b) Critically damped: $R = 2\sqrt{\frac{L}{C}}$. (c) Underdamped: $R < 2\sqrt{\frac{L}{C}}$.

Show that in the underdamped case, $\omega_0 = \frac{1}{\sqrt{LC}}$ and $\omega_1 = \sqrt{\frac{1}{LC} - \left(\frac{R}{2L}\right)^2}$.

3. Inductance L is measured in *henries*, resistance R is measured in *ohms*, and capacitance is measured in *farads*. Find the frequencies $\frac{\omega_0}{2\pi}$ and $\frac{\omega_1}{2\pi}$ if $L = 0.5$ henries, $R = 50$ ohms, and $C = 2 \times 10^{-4}$ farads.

What are the full names of the famous physicists after whom these units are named?

4. Approximately how many oscillatory cycles are required in the previous problem for the amplitude of the charge to drop to half its initial value?

7.8 The Method of Reduction of Order

In the case of equal roots, the equation with constant coefficients

$$y'' + py' + qy = 0$$

has the solution e^{rt} and also a second solution te^{rt} not dependent on the first. This is a fairly common situation. We have a method to find one solution $y_1(t)$ of

$$y'' + p(t)y' + q(t)y = 0,$$

but we need to find another solution which is not a constant multiple of $y_1(t)$. To this end, we look for solutions of the form $y = v(t)y_1(t)$ with $v(t)$ not constant. We have

$$\begin{aligned} y &= vy_1 \\ y' &= v'y_1 + vy_1' \\ y'' &= v''y_1 + v'y_1' + v'y_1' + vy_1'' = v''y_1 + 2v'y_1' + vy_1''. \end{aligned}$$

Thus

$$\begin{aligned} y'' + p(t)y' + q(t)y &= v''y_1 + 2v'y_1' + vy_1'' + pv'y_1 + pv'y_1' + qvy_1 \\ &= v''y_1 + (2y_1' + py_1)v' + v(y_1'' + py_1' + qy_1). \end{aligned}$$

Since y_1 is a solution, we have $y_1'' + py_1' + qy_1 = 0$. Hence, $y'' + py' + qy = 0$ amounts to the requirement

$$v'' + a(t)v' = 0 \quad \text{where } a(t) = \frac{2y_1'(t) + p(t)y_1(t)}{y_1(t)} = 2\frac{y_1'(t)}{y_1(t)} + p(t).$$

This may be treated as a first order equation in v' , and the general solution is

$$v' = Ce^{-\int a(t)dt}.$$

Since we need only one solution, we may take $C = 1$. Moreover,

$$\begin{aligned} \int a(t)dt &= 2 \int \frac{y_1'(t)}{y_1(t)} dt + \int p(t)dt \\ &= 2 \ln |y_1(t)| + \int p(t)dt = \ln y_1(t)^2 + \int p(t)dt. \end{aligned}$$

Hence,

$$\begin{aligned} v' &= e^{-\int a(t)dt} = e^{-\ln y_1(t)^2} e^{-\int p(t)dt} \quad \text{or} \\ v' &= \frac{1}{y_1(t)^2} e^{-\int p(t)dt}. \end{aligned} \quad (101)$$

Equation (101) may be integrated once more to determine v and ultimately another solution $y_2 = vy_1$.

Example 145 Consider

$$y'' + py' + qy = 0$$

where p and q are constant and $r = -p/2$ is a double root. Take $y_1 = e^{rt}$. Then

$$a(t) = 2 \frac{re^{rt}}{e^{rt}} + p = 2r + p = 0.$$

Hence, v' satisfies

$$v'' = 0$$

from which we derive $v' = c_1, v = c_1t + c_2$. Again, since we need only one solution, we may take $c_1 = 1, c_2 = 0$ to obtain $v = t$. This yields finally a second solution $y_2 = ty_1 = te^{rt}$, which is what we decided to try before.

Example 146 Consider Legendre's equation for $\alpha = 1$

$$y'' - \frac{2t}{1-t^2}y' + \frac{2}{1-t^2}y = 0.$$

You can check quite easily that $y_1(t) = t$ defines a solution. We look for a linearly independent solution of the form $y = v(t)t$. In this case $p(t) = -2t/(1-t^2)$ so

$$\int p(t)dt = \ln(1-t^2)$$

and by (101)

$$v' = \frac{1}{t^2} e^{-\ln(1-t^2)} = \frac{1}{t^2(1-t^2)}.$$

The right hand side may be integrated by partial fractions to obtain

$$v = -\frac{1}{t} + \frac{1}{2} \ln \left(\frac{1+t}{1-t} \right)$$

so we end up ultimately with

$$y_2 = vt = -1 + \frac{t}{2} \ln \left(\frac{1+t}{1-t} \right).$$

You would not be likely to come up with that by trial and error!

Exercises for 7.8.

1. One solution of $(1+t^2)y'' - 2ty' + 2y = 0$ is $y_1(t) = t$. Find a second solution by the method of reduction of order.
2. One solution of $y'' - 2ty' + 2y = 0$ is $y_1 = t$. Use the method of reduction of order to find another solution. Warning. You won't be able to determine v explicitly in this case, but push the integration as far as you can.
3. (a) Find one solution of Euler's equation $t^2y'' - 4ty' + 6y = 0$.
 (b) Find a second solution by the method of reduction of order.
 (c) Solve $t^2y'' - 4ty' + 6y = 0$ given $y(1) = 0, y'(1) = 1$.

7.9 The Inhomogeneous Equation. Variation of Parameters

We now investigate methods for finding a *particular* solution of the *inhomogeneous* equation

$$y'' + p(t)y' + q(t)y = f(t).$$

The first method is called *variation of parameters*. Let $\{y_1, y_2\}$ be a linearly independent pair of solutions of the homogeneous equation. We look for particular solutions of the inhomogeneous equation of the form

$$y = u_1y_1 + u_2y_2 \tag{102}$$

where u_1 and u_2 are functions to be determined. (The idea is that $c_1y_1 + c_2y_2$ would be a general solution of the homogeneous equation, and maybe we can get a solution of the inhomogeneous equation by replacing the constants by functions.)

We have

$$y' = u_1'y_1 + u_1y_1' + u_2'y_2 + u_2y_2'.$$

In order to simplify the calculation, look for u_1, u_2 satisfying

$$u_1'y_1 + u_2'y_2 = 0 \tag{103}$$

so

$$y' = u_1y_1' + u_2y_2'. \tag{104}$$

Then

$$y'' = u_1'y_1' + u_1y_1'' + u_2'y_2' + u_2y_2''. \tag{105}$$

Hence, using (102), (104), and (105), we have

$$\begin{aligned} y'' + py' + qy &= u_1'y_1' + u_1y_1'' + u_2'y_2' + u_2y_2'' + p(u_1y_1' + u_2y_2') + q(u_1y_1 + u_2y_2) \\ &= u_1'y_1' + u_2'y_2' + u_1(y_1'' + py_1' + qy_1) + u_2(y_2'' + py_2' + qy_2) \\ &= u_1'y_1' + u_2'y_2' \end{aligned}$$

because $y_1'' + py_1' + qy_1 = y_2'' + py_2' + qy_2 = 0$. (Both y_1 and y_2 are solutions of the homogeneous equation.) It follows that $y'' + p(t)y' + q(t)y = f(t)$ if and only if

$$u_1'y_1' + u_2'y_2' = f(t).$$

Putting this together with (103) yields the pair of equations

$$\begin{aligned} u_1'y_1 + u_2'y_2 &= 0 \\ u_1'y_1' + u_2'y_2' &= f(t) \end{aligned} \quad (106)$$

These can be solved for u_1' and u_2' by the usual methods. The solutions are

$$u_1' = \frac{-y_2 f}{y_1 y_2' - y_1' y_2} \quad u_2' = \frac{y_1 f}{y_1 y_2' - y_1' y_2}.$$

The denominator in each case is of course just the Wronskian $W(t)$ of the pair $\{y_1, y_2\}$, so we know it never vanishes. We may now integrate to obtain

$$u_1 = - \int \frac{y_2(t)f(t)}{W(t)} dt \quad u_2 = \int \frac{y_1(t)f(t)}{W(t)} dt.$$

Finally, we obtain the particular solution

$$\begin{aligned} y_p(t) &= y_1(t)u_1(t) + y_2(t)u_2(t) \\ &= -y_1(t) \int \frac{y_2(t)f(t)}{W(t)} dt + y_2(t) \int \frac{y_1(t)f(t)}{W(t)} dt. \end{aligned} \quad (107)$$

Example 147 Consider the equation

$$y'' - \frac{4}{t}y' + \frac{6}{t^2}y = t \quad \text{for } t > 0.$$

The homogeneous equation is a special case of Euler's equation as discussed in the exercises.

It has the solutions $y_1 = t^2$ and $y_2 = t^3$ and it is clear that these form a linearly independent pair. Let's apply the above method to determine a particular solution.

$$W(t) = \det \begin{bmatrix} y_1 & y_2 \\ y_1' & y_2' \end{bmatrix} = \det \begin{bmatrix} t^2 & t^3 \\ 2t & 3t^2 \end{bmatrix} = 3t^4 - 2t^4 = t^4.$$

Hence, taking $f(t) = t$, the variation of parameters formula gives

$$\begin{aligned} y_p &= -t^2 \int \frac{t^3 t}{t^4} dt + t^3 \int \frac{t^2 t}{t^4} dt \\ &= -t^2 t + t^3 \ln t = t^3(\ln t - 1). \end{aligned}$$

Thus the general solution is

$$y = y_p + c_1 y_1 + c_2 y_2 = t^3(\ln t - 1) + c_1 t^2 + c_2 t^3.$$

The variation of parameters formula may also be expressed using definite integrals with a dummy variable.

$$\begin{aligned} y_p &= -y_1(t) \int_{t_0}^t \frac{y_2(s)f(s)}{W(s)} ds + y_2(t) \int_{t_0}^t \frac{y_1(s)f(s)}{W(s)} ds \\ &= \int_{t_0}^t \frac{y_1(t)y_2(s) - y_2(t)y_1(s)}{W(s)} f(s) ds. \end{aligned} \quad (108)$$

Exercises for 7.9.

- Find a general solution of each of the following inhomogeneous equations by variation of parameters.
 - $y'' - 4y = e^{-t}$.
 - $y'' + 4y = \cos t$.
 - $y'' - 4y' + 4y = t$.
 - $y'' - 4y' + 4y = e^{-2t}$.
 - $y'' - 2y' + 2y = e^t \cos t$.
- Solve the initial value problem $y'' + 5y' + 6y = e^{4t}$ given $y(0) = 0, y'(0) = 1$.
- Find a general solution of $t^2 y'' - 2y = t$. Hint: The homogeneous equation is a special case of Euler's Equation.

7.10 Finding a Particular Solution by Guessing

The variation of parameters method is quite general, but it often involves a lot of unnecessary calculation. For example, the equation

$$y'' + 5y' + 6y = e^{4t}$$

has the linearly independent pair of solutions $y_1 = e^{-2t}, y_2 = e^{-3t}$. Because of all the exponentials appearing in the variation of parameters formula, there is quite a lot of cancellation, but it is apparent only after considerable calculation. You should try it out to convince yourself of that. On the other hand, a bit of experimentation suggests that something of the form $y = Xe^{4t}$ might work, and if we put this in the equation, we get

$$\begin{aligned} 16Xe^{4t} + 5(4Xe^{4t}) + 6Xe^{4t} &= e^{4t} \\ \text{or} \quad 42Xe^{4t} &= e^{4t} \end{aligned}$$

from which we conclude that $X = 1/42$. Thus, $y_p = (1/42)e^{4t}$ is a particular solution, and

$$y = \frac{1}{42}e^{4t} + c_1e^{2t} + c_2e^{3t}$$

is the general solution.

Of course, guessing won't work in complicated cases—see the previous section for an example where guessing would be difficult—but fortunately it often does work in the cases important in applications. For this reason, it is given a name: *the method of undetermined coefficients*. The idea is that we know by experience that guesses of a certain form will work for certain equations, so we try something of that form and all that is necessary is to determine some coefficient(s), as in the example above. In this section, we shall consider appropriate guesses for the equation

$$y'' + py' + qy = Ae^{\alpha t}, \quad (109)$$

where p, q and A are real constants, and α is a (possibly) complex constant. This covers applications to oscillatory phenomena. There is a lot more known about appropriate guesses in other cases, and if you ever need to use it, you should refer to a good book on differential equations. (See Section 2.5 of *Braun* for a start.)

The appropriate guess for a particular solution of (109) is $y = Xe^{\alpha t}$ where X is a (complex) constant to be determined. We have

$$\begin{aligned} y &= Xe^{\alpha t} \\ y' &= X\alpha e^{\alpha t} \\ y'' &= X\alpha^2 e^{\alpha t} \end{aligned}$$

so we need to solve

$$y'' + py' + qy = (\alpha^2 + p\alpha + q)Xe^{\alpha t} = Ae^{\alpha t}$$

for X . It is obvious how to do that *as long as the expression in parentheses does not vanish*. That will be the case when α is not a root of $r^2 + pr + q = 0$, i.e., the exponential on the right is *not a solution of the homogeneous equation*.

Hence, if α is not a root of the equation $r^2 + pr + q = 0$, then $X = A/(\alpha^2 + p\alpha + q)$, and a particular solution of the inhomogeneous equation (109) is given by

$$y_p = \frac{A}{\alpha^2 + p\alpha + q}e^{\alpha t}. \quad (110)$$

We still have to deal with the case that α is a root of the equation $r^2 + pr + q = 0$. Exactly what to do in this case depends on the nature of the roots of that equation. Assume first that *the roots r_1, r_2 are unequal*, and $\alpha = r_1$. In that case, the appropriate guess is $y = Xte^{\alpha t}$. We have

$$\begin{aligned} y &= Xte^{\alpha t} \\ y' &= Xe^{\alpha t} + Xt\alpha e^{\alpha t} \\ y'' &= 2X\alpha e^{\alpha t} + Xt\alpha^2 e^{\alpha t} \end{aligned}$$

so we need to solve

$$y'' + py' + qy = X(2\alpha + p)e^{\alpha t} + Xt(\alpha^2 + p\alpha + q)e^{\alpha t} = Ae^{\alpha t}.$$

Since α is a root, the second term in parentheses vanishes, so we need to solve

$$X(2\alpha + p) = A.$$

Since α , by assumption is not a double root, $\alpha \neq -p/2$, so the coefficient does not vanish, and we have $X = A/(2\alpha + p)$. Hence, the particular solution is

$$y_p = \frac{At}{2\alpha + p}e^{\alpha t}. \quad (111)$$

The final case to consider is that in which $\alpha = -p/2$ is a double root of $r^2 + pr + q = 0$. Then the appropriate guess is $y = Xt^2e^{\alpha t}$. We shall omit the details here, but you should work them out to see that you understand the process. The answer is $X = A/2$, and

$$y_p = \frac{At^2}{2}e^{\alpha t}$$

is a particular solutions.

Do you see a rule for what the denominators should be ?

The above analysis has wide ramifications because of the fact that α can be complex. In particular, the equation

$$y'' + py' + qy = Ae^{i\omega t} = A \cos \omega t + iA \sin \omega t, \quad (112)$$

where ω is a positive real constant, may be thought of as a pair of real equations

$$u'' + pu' + qu = A \cos \omega t \quad (113)$$

$$v'' + pv' + qv = A \sin \omega t \quad (114)$$

where u and v are respectively the real and imaginary parts of y . That suggests the following strategy for solving an equation of the form (112). Solve (112) instead, and then take the real part. We shall now do that under the various assumptions on whether or not $\alpha = i\omega$ is a root of $r^2 + pr + q = 0$.

Assume first that $i\omega$ is not a root of $r^2 + pr + q = 0$. That will always be the case, for example, if $p \neq 0$ since the roots are of the form $-p/2 \pm \sqrt{p^2 - 4q}/2$. Then the particular solution of (112) is

$$y_p = \frac{A}{-\omega^2 + ip\omega + q}e^{i\omega t}.$$

Let $Z = q - \omega^2 + ip\omega$. Then we can write the denominator in the form

$$Z = |Z|e^{i\delta}$$

where

$$|Z| = \sqrt{(q - \omega^2)^2 + p^2\omega^2}$$

and δ is the argument of Z , so

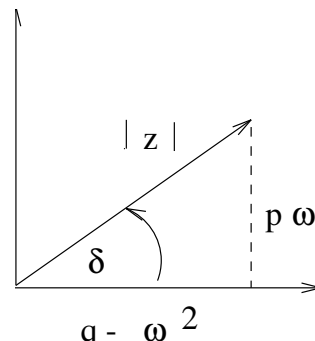
$$\tan \delta = \frac{p\omega}{q - \omega^2}.$$

(If $q = \omega^2$, interpret the last equation as asserting that $\delta = \pm\pi/2$ with the sign depending on the sign of p .) Then the particular solution may be rewritten

$$y_p = \frac{A}{|Z|e^{i\delta}} e^{i\omega t} = \frac{A}{|Z|} e^{i(\omega t - \delta)}.$$

If we take the real part of this, we obtain the following particular solution of the equation $u'' + pu' + qu = A \cos \omega t$:

$$u_p = \frac{A}{|Z|} \cos(\omega t - \delta). \quad (115)$$



We still have to worry about the case in which $i\omega$ is a root of $r^2 + pr + q = 0$. As mentioned above, we must have $p = 0$ and $\omega^2 = q$. The roots $\pm i\omega$ are unequal, so the particular solution of (112) in this case is

$$\begin{aligned} y_p &= \frac{At}{p + 2i\omega} e^{i\omega t} \\ &= \frac{At}{2\omega} (-i)(\cos \omega t + i \sin \omega t) \\ &= \frac{At}{2\omega} (\sin \omega t - i \cos \omega t). \end{aligned}$$

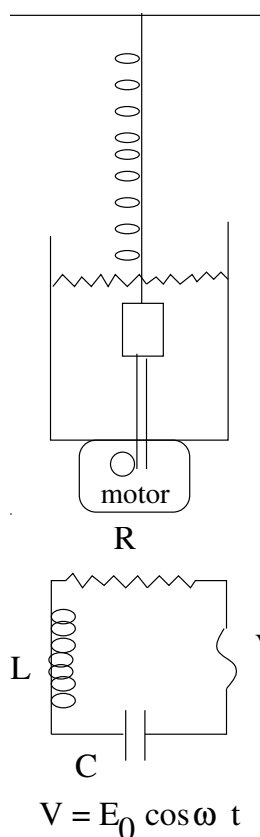
Taking the real part of this yields the following particular solution of $u'' + pu' + qu = A \cos \omega t$ in the case $\omega^2 = q$:

$$u_p = \frac{At}{2\omega} \sin \omega t. \quad (116)$$

Exercises for 7.10.

- Find a particular solution of $y'' + 2y' - 3y = t$ by trying a solution of the form $y = At + B$.
- Find a particular solution of $y'' + 2y' - 3y = e^t$ by trying a solution of the form $y = (At + B)e^t$. Would a solution of the form $y = Ae^t$ work?
- Find a particular complex solution of each of the following.
 - $y'' + y' + y = e^{it}$.
 - $y'' + 4y' + 5y = 6e^{2it}$.
 - $y'' + 3y' + 2y = 4e^{3it}$.
 - $y'' - 2y' + 2y = 2e^{(2+i)t}$.

4. Find the general solution of $y'' + 9y = e^{3it}$.
5. Find a particular real solution of each of the following. Hint. Consider the appropriate complex equation.
 - (a) $y'' + 4y = 3 \cos 3t$.
 - (b) $y'' + 4y' + 5y = 6 \cos 2t$.
 - (c) $y'' + 3y' + 2y = 4 \cos 3t$.
 - (d) $y'' - 2y' + 2y = 2e^{2t} \cos t$.
6. Find a particular solution of $y'' + 4y = 3 \cos 3t$ by trying something of the form $y = A \cos 3t + B \sin 3t$.
7. Find the general solution of each of the following.
 - (a) $y'' + 9y = 3 \cos 3t$.
 - (b) $y'' + 4y' + 5y = 6 \cos 2t$.
 - (c) $y'' + y' = \sin 2t$.



7.11 Forced Oscillations

Consider the equation for a mass at the end of a spring being driven by a periodic force

$$m \frac{d^2 y}{dt^2} + b \frac{dy}{dt} + ky = F_0 \cos \omega t \quad (117)$$

where m, b, k , and F_0 are positive constants.

The electrical analogue of this is the equation

$$L \frac{dI}{dt} + RI + \frac{Q}{C} = E_0 \cos \omega t \quad (118)$$

holding for the circuit indicated in the diagram, where Q is the charge on the capacitor C , $I = \frac{dQ}{dt}$ is the current, R is the resistance, and L is the inductance. The term on the right represents a periodic driving voltage.

As previously, we shall analyze the mechanical case, but the conclusions are also valid for the electric circuit. The fact that the same differential equation may govern different physical phenomena was the basis of the idea of an *analogue computer*.

Dividing through by m , we may write equation (118)

$$\frac{d^2 y}{dt^2} + \frac{b}{m} \frac{dy}{dt} + \frac{k}{m} y = \frac{F_0}{m} \cos \omega t \quad (119)$$

Its general solution will have the form

$$y = y_p(t) + h(t)$$

where $y_p(t)$ is one of the particular solutions considered in the previous section and $h(t)$ is a general solution of the homogeneous equation considered in Section 7. If you refer back to Section 7, you will recall that *provided* $b > 0$, $h(t) \rightarrow 0$ as $t \rightarrow \infty$. Hence, if we wait long enough, the particular solution will predominate. For this reason, the solution of the homogeneous equation is called the *transient part* of the solution, and the particular solution is called the *steady state* part of the solution. In most cases, all you can observe is the steady state solution. Since we have assumed that $b > 0$, (i.e., $p \neq 0$), we are in the case where $i\omega$ is not a root of $r^2 + pr + q = 0$. Hence, the desired steady state solution is, from (15) in the previous section,

$$y = \frac{A}{|Z|} \cos(\omega t - \delta)$$

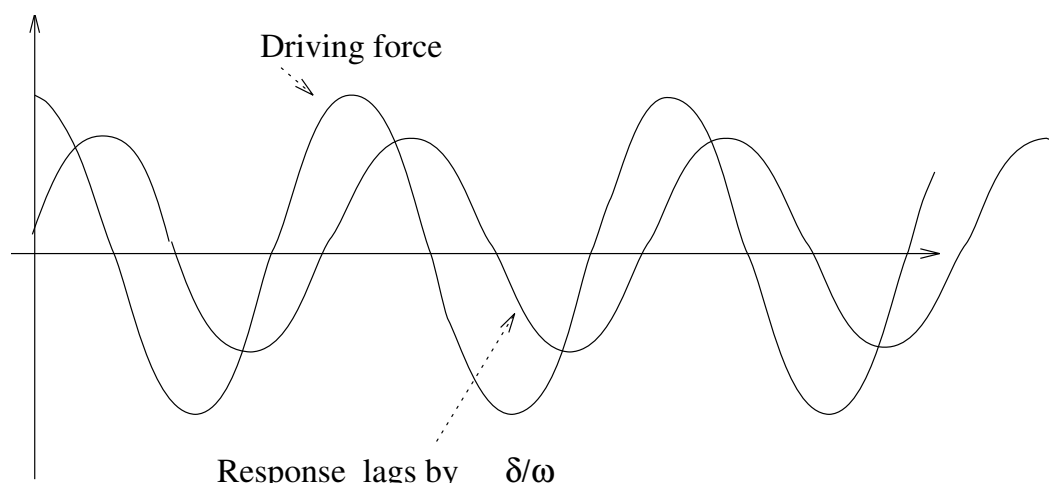
where

$$\begin{aligned} A &= \frac{F_0}{m} \\ |Z| &= \sqrt{(q - \omega^2)^2 + p^2\omega^2} = \sqrt{\left(\frac{k}{m} - \omega^2\right)^2 + \frac{b^2}{m^2}\omega^2} \\ &= \frac{1}{m} \sqrt{(k - m\omega^2)^2 + b^2\omega^2} \\ \tan \delta &= \frac{p\omega}{q - \omega^2} = \frac{(b/m)\omega}{k/m - \omega^2} \\ &= \frac{b\omega}{k - m\omega^2}. \end{aligned}$$

Hence, the steady state solution is

$$y = \frac{F_0}{\sqrt{(k - m\omega^2)^2 + (b\omega)^2}} \cos(\omega t - \delta).$$

where $\tan \delta = \frac{b\omega}{k - m\omega^2}$.

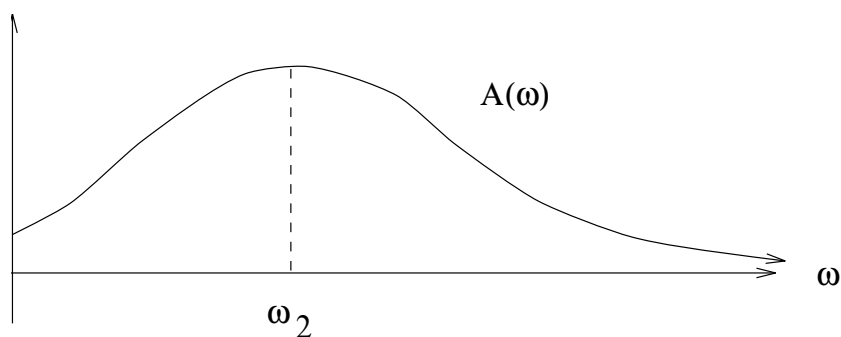


Note that the response to the driving force lags by a phase angle δ . This lag is a fundamental aspect of forced oscillators, and it is difficult to understand without a thorough understanding of the solution of the differential equation.

The amplitude of the steady state solution

$$A(\omega) = \frac{F_0}{\sqrt{(k - m\omega^2)^2 + (b\omega)^2}}$$

may be plotted as a function of the angular frequency ω .



Its maximum occurs for the value

$$\omega_2 = \sqrt{\frac{k}{m} - \frac{b^2}{2m^2}}.$$

(This is derived by the usual method from calculus: set the derivative with respect to ω equal to zero and calculate. See the Exercises.) If you recall, the quantity

$\omega_0 = \sqrt{k/m}$ is called the resonant frequency of the system. Using that, the above expression may be rewritten

$$\omega_2 = \sqrt{\omega_0^2 - \frac{b^2}{2m^2}}.$$

In Section 7, we introduced one other quantity:

$$\omega_1 = \sqrt{\omega_0^2 - \frac{b^2}{4m^2}},$$

which is the angular frequency of the damped, unforced oscillator. We have the inequalities

$$\omega_0 > \omega_1 > \omega_2.$$

If b is small, all three of these will be quite close together.

For many purposes, it is more appropriate to consider the square of the velocity $v^2 = (dy/dt)^2$ instead of the displacement y . For example, the kinetic energy of the system is $(1/2)mv^2$. (Similarly, in the electrical case the quantity RI^2 gives a measure of the power requirements of the system.) If as above we ignore the transient solution, we may take

$$\frac{dy}{dt} = -\frac{F_0\omega}{\sqrt{(k - m\omega^2)^2 + (b\omega)^2}} \sin(\omega t - \delta).$$

and the square of its amplitude is given by

$$B(\omega) = \frac{F_0^2\omega^2}{(k - m\omega^2)^2 + (b\omega)^2}.$$

If we divide through by ω^2 , this may be rewritten

$$B(\omega) = \frac{F_0^2}{(k/\omega - m\omega)^2 + b^2}.$$

It is easy to see where $B(\omega)$ attains its maximum, even without calculus. The maximum occurs when the denominator is at a minimum, but since the denominator is a sum of squares, its minimum occurs when its first term vanishes, i.e., when

$$\begin{aligned} \frac{k}{\omega} - m\omega &= 0 \\ \text{i.e., } \omega^2 &= \frac{k}{m}. \end{aligned}$$

Thus, the maximum of $B(\omega)$ occurs for $\omega = \omega_0 = \sqrt{k/m}$. This is one justification for calling ω_0 the resonant frequency.

Exercises for 7.11.

1. A spring is subject to the periodic force $F(t) = 2\cos(10\pi t)$. For each of the following values of m, b and k , determine the amplitude and phase of the response.
 - (a) $m = 1, k = 0.5, b = 0.005$.
 - (b) $m = 4, k = 1, b = 4$.
2. In each of the cases in the previous problem, assume the spring is displaced one unit of distance at $t = 0$ and is released with no initial velocity. Estimate in each case the number of cycles of the forced oscillation required before the transient response drops to one percent of the response due to the forced oscillation.
3. Show that the maximum of

$$A(\omega) = \frac{F_0}{\sqrt{(k - m\omega^2)^2 + (b\omega)^2}}$$

occurs for the value $\omega_2 = \sqrt{\frac{k}{m} - \frac{b^2}{2m^2}}$.

4. Show that the phase angle $\delta = \pi/2$ for the resonant frequency $\omega = \omega_0 = \sqrt{k/m}$. Thus, at resonance, the force and the response are one quarter cycle out of phase.
5. Show that the current $I = dQ/dt$ in the solution of equation (118) is of the form

$$I = \frac{E_0}{|Z|} \cos(\omega t - \delta)$$

where $Z = R + i(\omega L - \frac{1}{\omega C})$ and δ is the argument of Z . Hint: Replace $E_0 \cos(\omega t)$ by $E_0 e^{i\omega t}$ in equation (118), differentiate once to obtain a second order equation for I , and then apply the results of Section 10.

In the theory of alternating current circuits, the quantity Z is called the (complex) *impedance*. By analogy with Ohm's Law, the complex quantity E_0/Z describes the current in the sense that its absolute value is its magnitude and its argument describes its phase.

Chapter 8

Series

8.1 Series Solutions of a Differential Equation

There are several second order linear differential equations which arise in mathematical physics which cannot be solved by the methods introduced in the previous chapter. Here are some examples

$$\begin{aligned}y'' - \frac{2t}{1-t^2}y' + \frac{\alpha(\alpha+1)}{1-t^2}y &= 0 && \text{Legendre's equation} \\t^2y'' + ty' + (t^2 - \nu^2)y &= 0 && \text{Bessel's equation} \\ty'' + (1-t)y' + \lambda y &= 0 && \text{Laguerre's equation}\end{aligned}$$

A related equation we studied (in the Exercises to Chapter VII, Sections 4 and 8) is

$$y'' + \frac{\alpha}{t}y' + \frac{\beta}{t^2}y = 0 \quad \text{Euler's equation.}$$

These equations and others like them arise as steps in solving important partial differential equations such as Laplace's equation, the wave equation, the heat equation, or Schroedinger's equation. Their solutions are given appropriate names: *Legendre functions*, *Bessel Functions*, etc. Sometimes these *special functions* can be expressed in terms of other known functions, but usually they are entirely new functions. One way to express these solutions is as *infinite series*, and we investigate that method here.

Example Consider

$$(1-t^2)y'' - 2ty' + 6y = 0 \tag{120}$$

which is Legendre's equation for $\alpha = 2$. The simplest possible functions are the *polynomial functions*, those which can be expressed as combinations of powers of t

$$y = a_0 + a_1t + a_2t^2 + a_3t^3 + a_4t^4 + \cdots + a_dt^d.$$

The highest power of t occurring in such an expression is called its *degree*. We shall try to find a solution of (120) which is a polynomial function *without* committing ourselves about its degree in advance. We have

$$\begin{aligned} y &= a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + \cdots + \\ y' &= \quad a_1 + 2a_2 t + 3a_3 t^2 + 4a_4 t^3 + \cdots \\ y'' &= \quad \quad 2a_2 + 6a_3 t + 12a_4 t^2 + \cdots \end{aligned}$$

Note that these equations tell us that $y(0) = a_0$ and $y'(0) = a_1$, so the first two coefficients just give us the initial conditions at $t = 0$.

We have

$$\begin{aligned} 1 y'' &= 2a_2 + 6a_3 t + 12a_4 t^2 + 20a_5 t^3 + \cdots \\ -t^2 y'' &= \quad \quad -2a_2 t^2 - 6a_3 t^3 - \cdots \\ -2ty' &= \quad -2a_1 t - 4a_2 t^2 - 6a_3 t^3 - \cdots \\ 6y &= 6a_0 + 6a_1 t + 6a_2 t^2 + 6a_3 t^3 + \cdots \end{aligned}$$

The sum on the left is $(1 - t^2)y'' - 2ty' + 6y$ which is assumed to be zero. If we add up the coefficients of corresponding powers of t on the right, we obtain

$$0 = 2a_2 + 6a_0 + (6a_3 + 4a_1)t + 12a_4 t^2 + (20a_5 - 6a_3)t^3 + \cdots$$

Setting the coefficients of each power of t equal to zero, we obtain

$$2a_2 + 6a_0 = 0 \quad 6a_3 + 4a_1 = 0 \quad 12a_4 = 0 \quad 20a_5 - 6a_3 = 0 \quad \cdots$$

Thus,

$$\begin{aligned} a_2 &= -3a_0 \\ a_3 &= -\frac{2}{3}a_1 \\ a_4 &= 0 \\ a_5 &= \frac{3}{10}a_3 = -\frac{1}{5}a_1 \\ &\vdots \end{aligned}$$

The continuing pattern is more or less clear. Each succeeding coefficient a_n is expressible as a multiple of the coefficient a_{n-2} two steps lower. Since $a_4 = 0$, it follows that all coefficients a_n with n even must vanish for $n \geq 4$. On the other hand, the best that can be said about the coefficients a_n with n odd is that ultimately each of them may be expressed in terms of a_1 . It follows that if we separate out even and odd powers of t , the solution looks something like

$$\begin{aligned} y &= a_0 - 3a_0 t^2 + a_1 t - \frac{2}{3}a_1 t^3 - \frac{1}{5}a_1 t^5 + \cdots \\ &= a_0(1 - 3t^2) + a_1(t - \frac{2}{3}t^3 - \frac{1}{5}t^5 + \cdots). \end{aligned}$$

Consider first the solution satisfying the initial conditions $a_0 = y(0) = 1, a_1 = y'(0) = 0$. The corresponding solution is

$$y_1 = 1 - 3t^2$$

which is a polynomial of degree 2, just as expected. (However, notice that we had no way to predict in advance that it would be of degree 2.) The meaning of the ‘solution’ obtained by putting $a_0 = 0$ and $a_1 = 1$ is not so clear. It appears to be a ‘polynomial’

$$y_1 = t - \frac{2}{3}t^3 - \frac{1}{5}t^5 + \dots$$

which goes on forever. In fact, it is what is called an *power series*. This is just one example of many where one must represent a function in terms of an infinite series. We shall concern ourselves in the rest of this chapter with the theory of such series with the ultimate aim of understanding how they may be used to solve differential equations. For many of you this may be review, but you should pay close attention since some points you are not familiar with may come up.

Exercises for 8.1.

1. Find a polynomial solution of $y'' - \frac{2t}{1-t^2}y' + \frac{2}{1-t^2}y = 0$ as follows. First multiply through by $1 - t^2$ to avoid denominators. Then apply the method discussed in this section. Which of the two sets of initial conditions, $a_0 = 1, a_1 = 0$ or $a_0 = 0, a_1 = 1$, results in a polynomial solution?
2. Find a polynomial solution of $(1 - t^2)y'' - 2ty' + 12y = 0$.
3. Apply the method of this section to the first order differential equation $y' = y$. Show that with the initial condition $y(0) = a_0 = 1$ you obtain the ‘polynomial’ solution

$$y(t) = 1 + t + \frac{1}{2}t^2 + \frac{1}{3!}t^3 + \frac{1}{4!}t^4 + \dots$$

This is actually an infinite series which, we shall see in the ensuing sections, represents the function e^t .

8.2 Definitions and Examples

A series is a sequence of terms u_n connected by ‘+’ signs

$$u_1 + u_2 + u_3 + \dots + u_n + \dots$$

The idea is that if there are infinitely many terms, the summation process is supposed to go on forever. Of course, in reality that is impossible, but perhaps after we have added up sufficiently many terms, the contributions of additional terms will be so negligible that they really won't matter. We shall make this idea a bit more precise below.

The terms of the series can be positive, negative or zero. They can depend on variables or they can be constant. The summation can start at any index. In the general theory we assume the first term is u_1 , but the summation could start just as well with u_0 , u_6 , u_{-3} , or any convenient index.

Here are some examples of important series

$$\begin{array}{ll} 1 + t + t^2 + \cdots + t^n + \cdots & \text{geometric series} \\ 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} + \cdots & \text{harmonic series} \\ t - \frac{t^3}{3!} + \frac{t^5}{5!} - \cdots + (-1)^n \frac{t^{2n+1}}{(2n+1)!} + \cdots & \text{Taylor series for } \sin t \end{array}$$

In the first series, the general term is t^n and the numbering starts with $n = 0$. In the second series, the general term is $1/n$ and the numbering starts with $n = 1$.

Series are often represented more compactly using 'Σ'-notation

$$\sum_{n=1}^{\infty} u_n.$$

For example,

$$\begin{array}{ll} \sum_{n=0}^{\infty} t^n & \text{is the geometric series} \\ \sum_{n=1}^{\infty} \frac{1}{n} & \text{is the harmonic series} \end{array}$$

The notion of the sum of an infinite series is made precise as follows. Let

$$\begin{array}{l} s_1 = u_1 \\ s_2 = u_1 + u_2 \\ s_3 = u_1 + u_2 + u_3 \\ \vdots \\ s_n = u_1 + u_2 + \cdots + u_n = \sum_{j=1}^n u_j \\ \vdots \end{array}$$

s_n is called the n th *partial sum*. It is obtained by adding the n th term u_n to the previous partial sum, i.e.,

$$s_n = s_{n-1} + u_n.$$

Consider the behavior of the *sequence* of partial sums $s_1, s_2, \dots, s_n, \dots$ as $n \rightarrow \infty$. If this sequence approaches a finite limit

$$\lim_{n \rightarrow \infty} s_n = s$$

then we say the series *converges* and that s is its *sum*. Otherwise, we say that the series *diverges*.

Example Consider the geometric series

$$1 + t + t^2 + \dots + t^n + \dots = \sum_{n=0}^{\infty} t^n.$$

The partial sums are

$$\begin{aligned} s_0 &= 1 \\ s_1 &= 1 + t \\ s_2 &= 1 + t + t^2 \\ &\vdots \\ s_n &= 1 + t + t^2 + \dots + t^n = \frac{1 - t^{n+1}}{1 - t} \\ &\vdots \end{aligned}$$

The formula for s_n should be familiar to you from high school algebra, it is the sum of the first $n + 1$ terms of a geometric progression with starting term 1 and ratio t . It applies as long as $t \neq 1$. Let $n \rightarrow \infty$. There are several cases to consider.

Case (a). $|t| < 1$. Then, $\lim_{n \rightarrow \infty} t^n = 0$ so $\lim_{n \rightarrow \infty} s_n = \frac{1}{1 - t}$. Hence, in this case the series converges and the sum is $1/(1 - t)$.

Case (b) $t = 1$. Then $s_n = n \rightarrow \infty$, so the series diverges. We might also say in a case like this that the sum is ∞ .

Case (c) $t = -1$. Then the series is

$$1 - 1 + 1 - 1 + 1 - \dots$$

and the partial sums s_n alternate between 1 (n odd) and 0 (n even). As a result they don't approach any definite limit as $n \rightarrow \infty$.

Note that in cases (b) and (c) it is *not* true that $1/(1 - t)$ is the sum of the series, since that sum is not well defined. However, in case (c) ($t = -1$), we have $1/(1 - t) = 1/2$ which is the average of the two possible partial sums 1 and 0.

Case (d). Assume $t > 1$. In this case, it might be more appropriate to write

$$s_n = \frac{t^{n+1} - 1}{t - 1}. \quad (121)$$

As $n \rightarrow \infty$, $t^{n+1} \rightarrow \infty$, so the series diverges. However, since t^n increases without limit, it would also be appropriate in this case to say the sum is ∞ .

Case (e). Assume $t < -1$. Then in (121), the term t^{n+1} oscillates wildly between positive and negative values, and the same is true of s_n . Hence, s_n approaches no definite limit as $n \rightarrow \infty$, and the series diverges.

To summarize, *the geometric series $\sum_{n=0}^{\infty} t^n$ converges to the sum $\frac{1}{1-t}$ for $-1 < t < 1$ and diverges otherwise.*

Properties of Convergence We list some important properties of convergent and divergent series. 1. *Importance of the tail.* Whether or not a series converges does not depend on any *finite* number of terms. Thus, one should not be misled by what happens to the first few partial sums. (In this context ‘few’ might mean the first 10,000,000 terms!) Convergence of a series depends on the limiting behavior of the sequence of partial sums, and no finite number of terms in that sequence will tell you for sure what is happening in the limit.

Of course, the actual *sum* of a convergent series depends on all the terms.

2. *The general term u_n of a convergent series always approaches zero.* For, as noted earlier

$$s_n = s_{n-1} + u_n.$$

If s_n approaches a finite limit s , then it is not hard to see that s_{n-1} approaches this same limit s . Hence,

$$\lim_{n \rightarrow \infty} u_n = \lim_{n \rightarrow \infty} s_n - \lim_{n \rightarrow \infty} s_{n-1} = s - s = 0.$$

This rule provides a quick check for divergence in special cases. For example, the series

$$\frac{1}{2} + \frac{2}{3} + \dots + \frac{n}{n+1} + \dots$$

cannot possibly converge because

$$\lim_{n \rightarrow \infty} \frac{n}{n+1} = 1 \neq 0.$$

(That limit can be calculated using L'Hôpital's Rule or rewriting $\frac{n}{n+1} = \frac{1}{1+1/n}$.)

3. *The harmonic series $\sum_{n=1}^{\infty} 1/n$ diverges.* This is important because it shows us that the converse of the previous assertion is not true. That is, it is possible that

$u_n \rightarrow 0$ as $n \rightarrow \infty$ and the series still diverges. Many people find this counter-intuitive. Apparently, it is hard to believe that the sum of infinitely many things can be a finite number, and having convinced oneself that it can happen, one seeks an explanation. The explanation that comes to mind is that it can happen because the terms are getting smaller and smaller. However, as the example of the harmonic series shows, *that is not enough*. There are of course many other examples. Here is a *proof* that the harmonic series diverges. We use the principle that if s_n approaches a finite limit s , then any subsequence obtained by considering infinitely many selected values of n would also approach this same finite limit. Consider, in particular,

$$\begin{aligned} s_2 &= 1 + \frac{1}{2} = \frac{3}{2} \\ s_4 &= s_2 + \frac{1}{3} + \frac{1}{4} > \frac{3}{2} + 2\left(\frac{1}{4}\right) = \frac{4}{2} \\ s_8 &= s_4 + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} > \frac{4}{2} + 4\left(\frac{1}{8}\right) = \frac{5}{2} \\ s_{16} &= s_8 + \underbrace{\frac{1}{9} + \cdots + \frac{1}{16}}_{8 \text{ terms}} > \frac{5}{2} + 8\left(\frac{1}{16}\right) = \frac{6}{2} \\ &\vdots \end{aligned}$$

Continuing in this way, we can show in general that

$$s_{2^k} > \frac{k+2}{2}.$$

Since the right hand side is unbounded as $k \rightarrow \infty$, the sequence s_{2^k} cannot approach a finite limit. It follows that s_n cannot approach a finite limit. The harmonic series

also provides a warning to those who would compute without thinking. If you naively add up the terms of the harmonic series using a computer, the series will appear to approach a finite sum. The reason is that after the partial sums attain a large enough value, each succeeding term will get lost in round-off error and not contribute to the accumulated sum. The exact ‘sum’ you will get will depend on the word size in the computer, so it will be different for different computers. We shall see other examples of this phenomenon below. The moral is to be very careful about how you add up a large number of small terms in a computer if you want an accurate answer. 4. *Algebraic manipulation of series*. One may perform many of the operations with series that one is familiar with for finite sums, but some things don’t work.

For example, the sum or difference of two convergent series is convergent. Similarly, any multiple of a convergent series is convergent.

Example 148 Each of the series $\sum_{n=0}^{\infty} \frac{1}{2^n}$ and $\sum_{n=0}^{\infty} \frac{1}{3^n}$ is a geometric series with $|t| < 1$. ($t = 1/2$ for the first and $t = 1/3$ for the second.) Hence, both series

converge and so does the series

$$\sum_{n=0}^{\infty} \left(\frac{1}{2^n} + \frac{1}{3^n} \right).$$

In fact, its sum is obtained by adding the sums of the two constituent series

$$\frac{1}{1-1/2} + \frac{1}{1-1/3} = 2 + \frac{3}{2} = \frac{7}{2}.$$

Example 149

$$\sum_{n=0}^{\infty} t^{n+1} = t \sum_{n=0}^{\infty} t^n = t \frac{1}{1-t} = \frac{t}{1-t}$$

is valid for $-1 < t < 1$.

On occasion, you can combine divergent series to get a convergent series.

Example 150 Consider the two series

$$\sum_{n=1}^{\infty} \frac{1}{n} \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{1}{n+1}.$$

The first is the harmonic series, which we saw is not convergent. The second series is

$$\frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n+1} + \cdots$$

which is the harmonic series with the first term omitted. Hence, it is also divergent. However, the difference of these two divergent series is

$$\sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{n+1} \right) = \sum_{n=1}^{\infty} \frac{1}{n(n+1)},$$

and this series is convergent. (This follows by one of the tests for convergence we shall investigate in the next section, but it may also be checked directly. See the Exercises.) However, *not all manipulations are valid*.

Example 151 We saw that the series

$$1 - 1 + 1 - 1 + 1 - 1 + \cdots$$

is not convergent. However, we may produce a convergent series of regrouping terms

$$(1 - 1) + (1 - 1) + (1 - 1) + \cdots = 0 + 0 + 0 + \cdots$$

or

$$1 + (-1 + 1) + (-1 + 1) + \cdots = 1 + 0 + 0 + 0 + \cdots$$

Generally speaking, regrouping of terms so as to produce a new series won't produce correct results, although in certain important cases it does work. We shall return to this point in Sections 3 and 5.

Exercises for 8.2.

1. In each of the following cases, determine if the the indicated *sequence* s_n converges, and if so find its limit. The sequence could have arisen as the sequence of partial sums of a series, but it could also have arisen some other way. Usually the simplest way to determine the limit is to use L'Hôpital's Rule, but in some cases other methods may be necessary. (If you don't know L'Hôpital's Rule, ask your instructor.)

(a) $s_n = \frac{n^2 + 2n + 3}{3n^3 - 5n^2}.$

(b) $s_n = 1 - \frac{99^n}{100^n}.$ (Don't use L'Hôpital's Rule.)

(c) $s_n = 1 - (-1)^n.$

(d) $s_n = \frac{\sin(nt)}{n}$ where t is a variable.

(e) $s_n = \frac{\ln n}{n}.$

(f) $s_n = \left(\frac{1}{n}\right)^{1/n}.$ (Consider $\ln s_n$ and apply L'Hôpital's Rule to that.)

2. In each of the following cases, determine if the indicated *series* converges.

(a) $1 + 1/5 + (1/5)^2 + \cdots = \sum_{n=0}^{\infty} (1/5)^n.$

(b) $1 + 2 + 3 + \cdots = \sum_{n=1}^{\infty} n.$

(c) $1 - 4 + 9 - \cdots = \sum_{n=1}^{\infty} (-1)^{n+1} n^2.$

(d) $2 - 2/3 + 2/9 - \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{2}{3^n}.$

(e) $(1+t) - (1+t)^2 + (1+t)^3 - \cdots = \sum_{n=1}^{\infty} (-1)^{n+1} (1+t)^n.$ where $t > 0$. Does it matter how small t is? Suppose $t = 10^{-100}$. What if $-2 < t < 0$?

(f) $1/2 - 2/3 + 3/4 - \cdots = \sum_{n=1}^{\infty} (-1)^n \frac{n}{n+1}.$

(g) $\sum_{n=1}^{\infty} \left(\frac{1}{3^n} - \frac{1}{4^n} \right).$

(h) $\sum_{n=1}^{\infty} \cos(n\pi)$

(i) $\sum_{n=0}^{\infty} \frac{2^n + 99^n}{100^n}.$

3. Consider the repeating decimal

$$x = .2315231523152315\dots$$

Treating this as the sum of the infinite series $\sum_{n=0}^{\infty} at^n$ where $a = .2315$ and $t = .0001$, determine x .

Every infinite decimal which repeats in this way can be evaluated by this method. The method can be adapted also to infinite decimals which *eventually* start repeating.

4. Suppose that when a ball is dropped from a given height, it bounces back to r times that height where $0 < r < 1$. If it is originally dropped from height h , find the total distance it moves as it bounces back and forth infinitely many times. Hint: Except for the initial drop, each bounce involves moving through the same distance twice.
5. Two trains initially 100 miles apart approach, each moving at 20 mph. A bird, flying at 40 mph, starts from the first train, flies to the second train, returns to the first train, ad infinitum until the two trains meet. Use a geometric series to find the total distance the bird flies.

There is an easier way to do this. Do you see it? A story is told of the famous mathematician John von Neumann who when given the problem immediately responded with the answer. In response to the remark, "I see you saw the easy method", he is supposed to have answered "What easy method?"

6. For

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \sum_{n=1}^{\infty} \frac{1}{n} - \frac{1}{n+1},$$

show that $s_n = 1 - \frac{1}{n+1}$. Hint: The partial sum s_n is what is called a 'collapsing sum', i.e., mutual cancellation for adjacent terms eliminates everything except the '1' in the first term $1 - \frac{1}{2}$, and the $\frac{1}{n+1}$ in the last term $\frac{1}{n} - \frac{1}{n+1}$.

Conclude that the series converges and the sum is 1.

7. What is wrong with the following argument?

$$1 - 1 + 1 - 1 + \dots = 1 - (1 - 1 + 1 - \dots)$$

or $x = 1 - x$

where x is the sum. It follows that $x = 1/2$.

8.3 Series of Non-negative Terms

Some of the strange things that happen with series result from trying to balance one ‘infinity’ against another. Thus, if infinitely many terms are positive, and infinitely many terms are negative, the sum of the former might be ‘ $+\infty$ ’ while the sum of the latter might be ‘ $-\infty$ ’, and the result of combining the two is anyone’s guess. For series $\sum_n u_n$ in which all the terms have the same sign, that is not a problem, so such series are much easier to deal with. In this section we shall consider series in which all terms $u_n \geq 0$, but analogous conclusions are valid for series in which all $u_n \leq 0$.

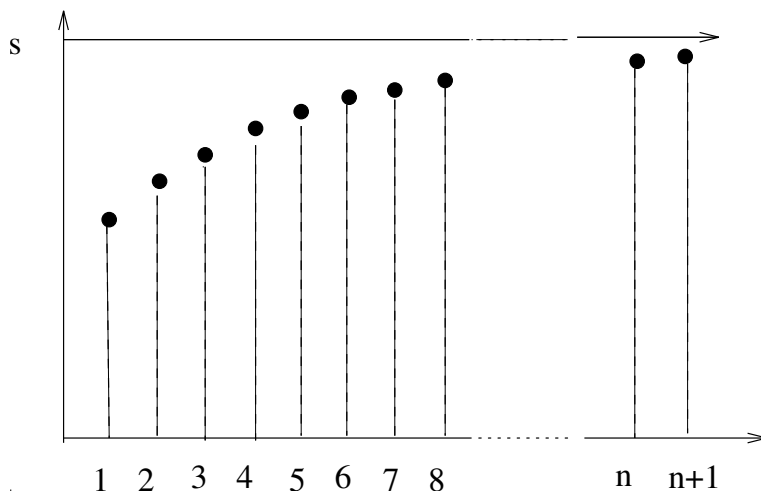
Let $\sum_{n=1}^{\infty} u_n$ be such a series with non-negative terms u_n . In the sequence of partial sums we have

$$s_n = s_{n-1} + u_n \geq s_{n-1},$$

i.e., the next term in the sequence is never less than the previous term. There are only two possible ways such a sequence can behave.

1. "(a)" The sequence s_n can grow without bound. In that case we say $s_n \rightarrow \infty$ as $n \rightarrow \infty$.
2. "(b)" The sequence s_n can remain bounded so no s_n ever exceeds some pre-assigned upper limit. In that case, the sequence must approach a finite limit s , i.e., $s_n \rightarrow s$ as $n \rightarrow \infty$.

The fact that non-decreasing sequences behave this way is a fundamental property of the real number system called *completeness*.



Example Consider the decimal expansion of π . It may be considered the limit of the sequence

$$\begin{aligned}s_1 &= 3 \\s_2 &= 3.1 \\s_3 &= 3.14 \\s_4 &= 3.141 \\s_5 &= 3.1415 \\s_6 &= 3.14159 \\&\vdots\end{aligned}$$

This is a sequence of numbers which is bounded above (for example, by 4), so (b) tells us that it approaches a finite limit, and that limit is the real number π .

Note that for series, with terms alternately positive and negative, the above principle does not apply.

Example For the series $1 - 1 + 1 - 1 + 1 - \dots$, the partial sums are

$$\begin{aligned}s_1 &= 1 \\s_2 &= 0 \\s_3 &= 1 \\s_4 &= 0 \\&\vdots\end{aligned}$$

so although the sequence is bounded above (and also below), it never approaches a finite limit.

For series consisting of terms all of the same sign (or zero), rearranging the terms does not affect convergence or divergence or the sum when convergent. We won't try to prove this in this course because we would first have to define just precisely what we mean by 'rearranging' the terms of a series. You can find a proof in any good calculus book. (See for example *Calculus* by Tom M. Apostol).

The Integral Test Suppose we have a series $\sum_{n=1}^{\infty} u_n$ where each term is positive ($u_n > 0$) and moreover the terms decrease

$$u_1 > u_2 > \dots > u_n > \dots$$

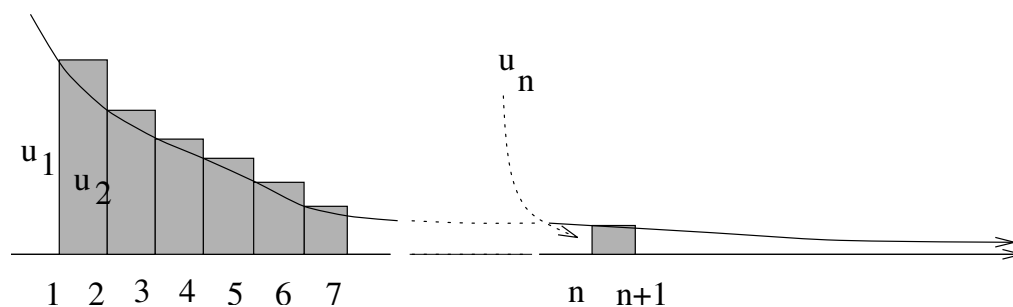
Suppose in addition that $\lim_{n \rightarrow \infty} u_n = 0$. Often it is possible to come up with a function $f(x)$ of a real variable such that $u_n = f(n)$ for each n . This function should be continuous and it should also be decreasing and approach 0 as $x \rightarrow \infty$. Then, there is a very simple test to check whether or not the series converges. $\sum_n u_n$ converges if and only if the improper integral $\int_1^{\infty} f(x) dx$ is finite.

Example 152 The harmonic series $\sum_n 1/n$ satisfies the above conditions and we may take $f(x) = 1/x$. Then

$$\int_1^\infty \frac{1}{x} dx = \lim_{X \rightarrow \infty} \int_1^X \frac{dx}{x} = \lim_{X \rightarrow \infty} \ln x \Big|_1^X = \lim_{X \rightarrow \infty} \ln X = \infty.$$

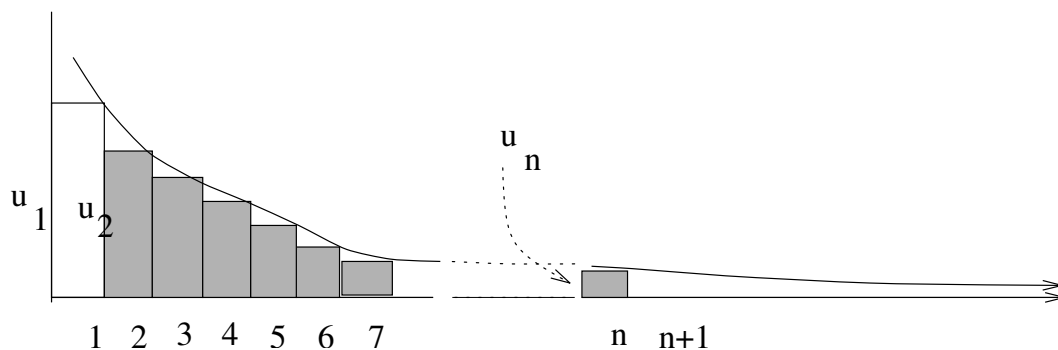
It follows that the series diverges, a fact we already knew.

The reason why the integral test works is fairly clear from some pictures. From the diagram below



$$s_n = u_1 + u_2 + \cdots + u_n \geq \int_1^{n+1} f(x) dx$$

Hence, if $\int_1^\infty f(x) dx = \lim_{n \rightarrow \infty} \int_1^{n+1} f(x) dx = \infty$, it follows that the sequence s_n is not bounded and (a) applies, i.e., $s_n \rightarrow \infty$ and the series diverges. On the other hand from the diagram below,



$$s_n - u_1 = u_2 + u_3 + \cdots + u_n \leq \int_1^n f(x)dx.$$

Hence, by similar reasoning if the integral is finite, the sequence s_n is bounded, so by (b) it approaches a limit, and the series converges.

Example 153 The series $\sum_1^\infty \frac{1}{n^p}$ with $p > 0$ is called the ‘ p -series’. The conditions of the integral test apply, and we may take $f(x) = 1/x^p$. Then, with the exception of the case $p = 1$, which we have already dealt with, we have

$$\int_1^\infty \frac{dx}{x^p} = \lim_{X \rightarrow \infty} \int_1^X \frac{dx}{x^p} = \lim_{X \rightarrow \infty} \left. \frac{x^{-p+1}}{-p+1} \right|_1^X = \lim_{X \rightarrow \infty} \frac{1}{1-p} [X^{1-p} - 1].$$

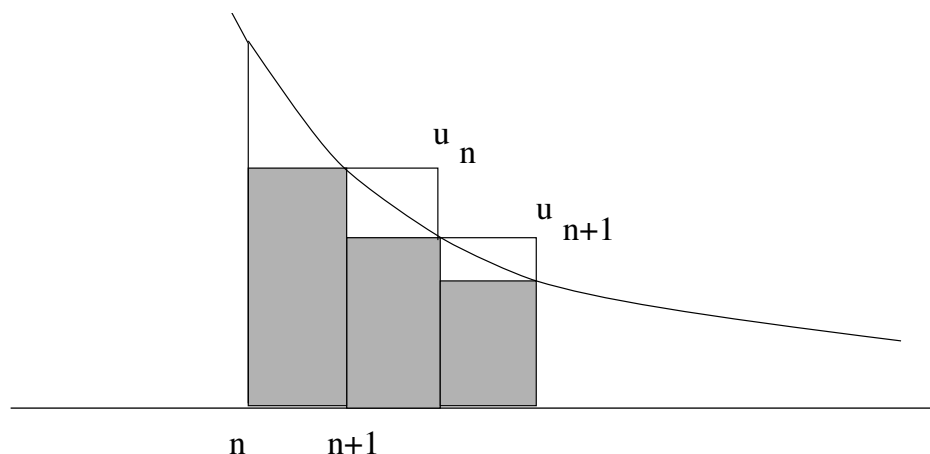
If $p > 1$, then $X^{1-p} = \frac{1}{X^{p-1}} \rightarrow 0$ so the above limit is $\frac{1}{p-1}$ (finite), and the series converges. On the other hand, if $0 < p < 1$, then $1-p > 0$ and $X^{1-p} \rightarrow \infty$, so the above limit is not finite and the series diverges. For $p = 1$, the series is the harmonic series which we already discussed. Hence, we conclude that the ‘ p -series’ diverges for $0 < p \leq 1$ and converges for $p > 1$.

Error estimates with the integral test. The integral test may be refined to estimate the speed at which a series converges. Suppose $\sum_n u_n$ is convergent and its sum is s . Decompose s as follows

$$s = \underbrace{u_1 + u_2 + \cdots + u_n}_{s_n} + \underbrace{u_{n+1} + u_{n+2} + \cdots}_{R_n}$$

Here, $R_n = s - s_n = \sum_{j=n+1}^\infty u_j$ is the error you make if you stop adding terms after the n th term. It is sometimes called the *remainder*. From the diagram below, you can see that

$$\int_{n+1}^\infty f(x)dx < R_n = \sum_{j=n+1}^\infty u_j < \int_n^\infty f(x)dx. \quad (122)$$



Example 154 Consider the series $\sum_{n=1}^{\infty} 1/n^2$. We have

$$\int_{n+1}^{\infty} \frac{dx}{x^2} < R_n < \int_n^{\infty} \frac{dx}{x^2}.$$

Evaluating the two integrals, we obtain

$$\frac{1}{n+1} < R_n < \frac{1}{n}.$$

This gives us a pretty good estimate of the size of the error. For n large, the upper and lower estimates are about the same, so it makes sense to say $R_n \approx 1/n$. For example, I used Mathematica to add up 100 terms of this series and got (to 5 decimal places) 1.63498, but it is known that the true sum of the series is $\pi^2/6$ which (to 5 decimal places) is 1.64493. As you see, they differ in the second decimal place, and the error is about $1/100 = .01$ as predicted.

A common blunder. It is often suggested that when adding up terms of a series on a computer, it suffices to stop when the next term you would add is less than the error you are willing to tolerate. The above example shows that this suggestion is nonsense. For, if you were willing to accept an error of $.0001 = 10^{-4}$, you would then stop when $1/n^2 < .10^{-4}$ or $n^2 > 10^4$ or $n > 100$. However, the above analysis shows that the actual error at that stage would be about $1/100 = .01$.

This rule makes even less sense for a divergent series like the harmonic series because it tells you that its sum is finite.

The Comparison Test By a variation of the reasoning used to derive the integral test, one may derive the following criteria to determine if a series $\sum_n u_n$ of non-negative terms converges or diverges. (a) If you can find a convergent series $\sum_{n=1}^{\infty} a_n$ such that $u_n \leq a_n$ for every n , then $\sum_{n=1}^{\infty} u_n$ also converges. (b) If you can find a

divergent series $\sum_{n=1}^{\infty} b_n$ of non-negative terms such that $b_n \leq u_n$, then $\sum_{n=1}^{\infty} u_n$ also diverges.

Example 155 The series $\sum_{n=1}^{\infty} \frac{\ln n}{n}$ diverges because it can be compared to the harmonic series, i.e.,

$$\frac{\ln n}{n} \geq \frac{1}{n} \quad \text{for } n > 1,$$

so $\sum \ln n/n$ diverges since $\sum 1/n$ diverges. (Why doesn't it matter that the comparison fails for $n = 1$?)

Example 156 Consider the series

$$\sum_{n=1}^{\infty} \frac{n}{n^3 - n + 2}.$$

We have

$$\frac{n}{n^3 - n + 2} = \frac{1}{n^2 - 1 + 2/n} < \frac{1}{n^2 - 1} < \frac{1}{n^2 - n^2/2} = \frac{1}{n^2/2} = \frac{2}{n^2}.$$

The first equality is obtained by dividing numerator and denominator by n . Each inequality is obtained by making the denominator smaller. Since $\sum_n 2/n^2 = 2 \sum_n 1/n^2$ and $\sum_n 1/n^2$ converges, so does $\sum_n 2/n^2$. Hence, the comparison test tells us that $\sum_n n/(n^3 - n + 2)$ converges.

The last example illustrates a common strategy when applying the comparison test. We try to estimate the approximate behavior of u_n for large n and thereby come up with an appropriate comparison series. We then try (usually by a tricky argument) to show that the given series is less than the comparison series (if we expect convergence) or greater than the comparison series (if we expect divergence.) Unfortunately, it is not always easy to dream up the relevant comparisons. Hence, it is often easier to use the following version of the comparison test.

The Limit Comparison Test Suppose $\sum_n u_n$ and $\sum_n c_n$ are series of non-negative terms such that $\lim_{n \rightarrow \infty} \frac{u_n}{c_n}$ exists and is not 0 or ∞ . Then $\sum_n u_n$ converges if and only if $\sum_n c_n$ converges. If the limiting ratio is 0, and $\sum_n c_n$ converges, then $\sum_n u_n$ converges. If the limiting ratio is ∞ , and $\sum_n c_n$ diverges, then $\sum_n u_n$ diverges.

Example 156, again $u_n = n/(n^3 - n + 2)$ looks about like $c_n = 1/n^2$. Consider the ratio

$$\frac{u_n}{c_n} = \frac{n}{n^3 - n + 2} \frac{n^2}{1} = \frac{n^3}{n^3 - n + 1} = \frac{1}{1 - 1/n^2 + 2/n^3} \rightarrow 1$$

as $n \rightarrow \infty$. Since 1 is finite and non-zero, and since the comparison series $\sum_n 1/n^2$ converges, the series $\sum_n n/(n^3 - n + 2)$ also converges.

It is important to emphasize that this test does not work for series for which some terms are positive and some are negative.

The Proof The proof of the limit comparison test is fairly straightforward, but if you are not planning to be a Math major and you are willing to accept such things on faith, you should skip it. Suppose

$$\lim_{n \rightarrow \infty} \frac{u_n}{c_n} = r \quad \text{with } 0 < r < \infty.$$

Then, for all n sufficiently large

$$\frac{r}{2} < \frac{u_n}{c_n} < 2r.$$

(All the values of u_n/c_n have to be very close to r when n is large enough.) Hence, for all n sufficiently large,

$$\frac{r}{2}c_n < u_n < 2rc_n$$

If $\sum_n c_n$ converges, then so does $\sum_n 2rc_n$, so by the comparison test, so does $\sum_n u_n$. By similar reasoning, if $\sum_n c_n$ diverges, so does $\sum_n u_n$. Note that we have used the remark about the importance of the tail—see Section 2—which asserts that when deciding on matters of convergence or divergence, it is enough to look only at all sufficiently large terms.

Exercises for 8.3.

1. Apply the integral test in each of the following cases to see if the indicated series converges or diverges.

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

(b) $\sum_{n=1}^{\infty} \frac{1}{n^2+1}$.

(c) $\sum_{n=1}^{\infty} \frac{1}{\sqrt{2n+1}}$.

(d) $\sum_{n=2}^{\infty} \frac{1}{n \ln n}$.

(e) $\sum_{n=1}^{\infty} \frac{\ln n}{n^3}$.

2. Why doesn't the integral test apply to each of the following series?

(a) $\sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n}$.

(b) $\sum_{n=1}^{\infty} \frac{n}{n+1}$.

3. Apply one of the two comparison tests to determine if each of the following series converges or diverges.

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

(b) $\sum_{n=1}^{\infty} \frac{1}{n^2+1}$.

(c) $\sum_{n=1}^{\infty} \frac{1}{\sqrt{2n+1}}$.

(d) $\sum_{n=2}^{\infty} \frac{\ln n}{n}$.

(e) $\sum_{n=1}^{\infty} \frac{\ln n}{n^3}$.

4. Determine if each of the following series converges or diverges by applying one of the tests discussed in this section.

(a) $\sum_{n=1}^{\infty} \frac{n+1}{n^3+3n+1}$.

(b) $\sum_{n=1}^{\infty} \frac{n^2+2n+2}{n^3+n+2}$.

(c) $\sum_{n=1}^{\infty} \frac{1}{2n-1}$.

(d) $\sum_{n=1}^{\infty} \frac{n}{e^n}$.

(e) $\sum_{n=1}^{\infty} \frac{1+\sin nt}{n^2}$ where t is a real number.

5. (a) Estimate the error we make in calculating $\sum_{n=1}^{\infty} \frac{1}{n^3}$ if we stop after 1000 terms.

(b) Estimate the least number of terms we need to use to calculate $\sum_{n=1}^{\infty} \frac{1}{n^3}$ accurately to within 5×10^{-16} . (This is close to but not the same as asking that it be accurate to 15 decimal places.)

6. (a) Estimate how many terms of the series $\sum_{n=1}^{\infty} \frac{1}{n(n+1)}$ are necessary to calculate the sum accurately to within 5×10^{-6} . (This will require a little work because you have to find $\int \frac{dx}{x(x+1)}$.

(b) Write a computer program to do the calculation. See what you get and compare it to the true answer which you know to be 1. (See Section 2.)

7. (a) Estimate how many terms of the series $\sum_{n=2}^{\infty} \frac{1}{n \ln n}$ are necessary to compute the sum accurately to within 5×10^{-4} .

(b) Why is this problem silly?

8. (Optional) Suppose $\sum_n u_n$ and $\sum_n c_n$ are series of non-negative terms such that $\frac{u_n}{c_n} \rightarrow 0$. Show that if $\sum_n c_n$ converges, then so also does $\sum_n u_n$. Hint: Show that for all n sufficiently large, $u_n < c_n$ and use the previous version of the comparison test.

8.4 Alternating Series

A series with alternating positive and negative terms is called an *alternating series*. Such a series is usually represented

$$v_1 - v_2 + v_3 - v_4 + \dots$$

where v_1, v_2, v_3, \dots are all positive. Of course, variations of this scheme are possible. For example, the first term might have an index other than 1, and also there is no reason not to start with a negative term. However, to simplify the discussion of the general theory, we shall assume the series is as presented above. For alternating series, we start to encounter the problem of ‘balancing infinities’, but they are still relatively well behaved. Also, many series important in applications are alternating series.

Examples of Alternating Series

$$\sum_{n=0}^{\infty} t^n \quad \text{if } t \text{ is negative} \quad (123)$$

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \dots \quad (124)$$

$$\sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots \quad (125)$$

(123) is a special case of the geometric series, and we shall see later that (125) is a series for $\sin x$. **Theorem 8.12** Let $v_1 - v_2 + v_3 - \dots$ be an alternating series with

$$v_1 > v_2 > \dots > v_n > \dots \quad (\text{all } > 0).$$

If $\lim_{n \rightarrow \infty} v_n = 0$, then the series converges. If s is the sum of the series, then the error $R_n = s - s_n$ after n terms satisfies

$$|R_n| < v_{n+1}.$$

That is, the absolute value of the error after n terms is bounded by the next term of the series.

Example The series

$$1 - \frac{1}{2} + \frac{1}{3} - \cdots + (-1)^{n+1} \frac{1}{n} + \cdots$$

converges since $\frac{1}{n} \rightarrow 0$. Also,

$$s = 1 - \underbrace{\frac{1}{2} + \frac{1}{3} - \cdots + (-1)^{n+1} \frac{1}{n}}_{s_n} + R_n$$

where $|R_n| < \frac{1}{n+1}$. According to Mathematica, for $n = 100$, $s_n = 0.688172$ (to 6 decimal places), and the theorem tells us the error is less than $1/100 = .01$. Hence, the correct value for the sum is somewhere between .67 and .69.

Suppose on the other hand, we want to use enough terms to get an answer such that the error will be less than $.0005 = 5 \times 10^{-4}$. For that, we need

$$\begin{aligned} \frac{1}{n+1} &\leq 5 \times 10^{-4} \quad \text{or} \\ n+1 &\geq \frac{1}{5 \times 10^{-4}} = \frac{1}{5} 10^4 = 2 \times 10^3. \end{aligned}$$

Thus, $n = 2000$ would work.

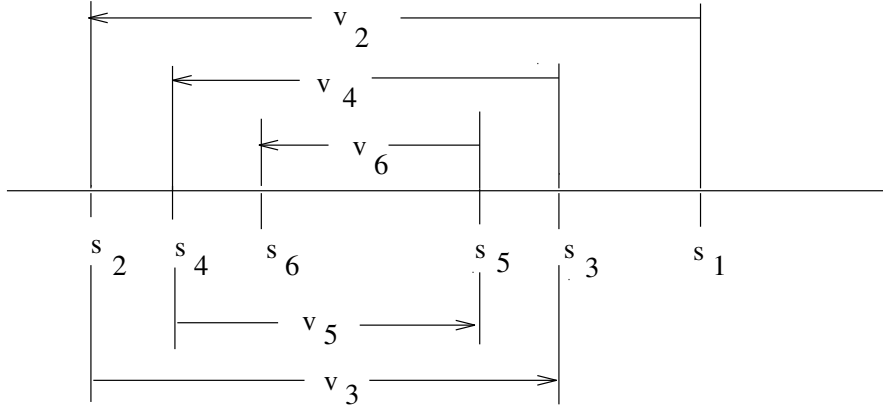
An aside on rounding.

In the above example, we asked for enough terms to be sure that the error is less than 5×10^{-4} . This is *different from* asking that it be accurate to three decimal places, although the two questions are closely related. For example, suppose the true answer to a problem is 1.32436... An error of $+0.0004$ would change this to 1.32476.. which would be rounded *up* to the incorrect answer 1.325 if we used only three decimal places. An alternative to rounding is *truncating*, i.e. cutting off all digits past the desired one. That would solve our problem in the case of a positive error, but an error of -0.0004 would in this case result in the incorrect answer 1.323 if we truncated.

By using extra decimal places, it is possible to reduce the likelihood of a false rounding error, but it is not possible to eliminate it entirely. For example, suppose the true answer is 2.426499967 and we insist on enough terms so that the error is less than 5×10^{-8} . Then, an error of $+0.00000004$ would cause us to round up incorrectly to 2.427.

Since the issue of rounding can be so confusing, we shall instead generally ask that the error be smaller than five in the next decimal position. You should generally read a request for a certain number of decimal places in this way but you should also remember that rounding may change the answer a bit in such cases.

The Proof. The proof of the theorem is a little tricky, but it is illuminating enough that you should try to follow it. You will get a better idea of how the error behaves, and that is important. The diagram below makes clear what happens.



Here is the same thing in words. If n is odd,

$$s_{n-1} < s_{n-1} + \underbrace{u_n - u_{n-1}}_{>0} = s_{n+1}. \quad (126)$$

Hence, the even numbered partial sums s_n form an increasing sequence. Similarly,

$$s_{n+1} = s_{n-1} - \underbrace{(u_n - u_{n-1})}_{>0} < s_{n-1}, \quad (127)$$

so the odd numbered partial sums s_n form a decreasing sequence. Also, if n is odd,

$$s_{n+1} = s_n - v_{n+1} < s_n,$$

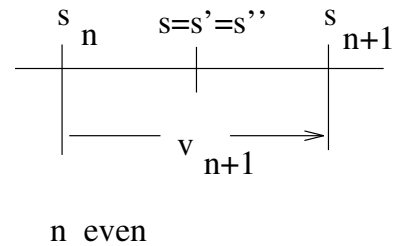
so it follows that every odd partial sum is greater than every even partial sum:

$$s_2 < s_4 < s_6 < \cdots < s_7 < s_5 < s_3 < s_1.$$

The even partial sums form an increasing sequence, bounded above, so by the completeness property (b), they must approach a finite limit s' . By similar reasoning, since the odd partial sums form a *decreasing* sequence, bounded below, they also approach a finite limit s'' . Moreover, it is not too hard to see that for n even

$$s_n \leq s' \leq s'' \leq s_{n+1}.$$

However, $s_{n+1} - s_n = v_{n+1} \rightarrow 0$ as $n \rightarrow \infty$. It follows that $s'' - s' \rightarrow 0$ as $n \rightarrow \infty$, but since $s'' - s'$ does not depend on n , it must be equal to zero. In other words the even partial sums *and* the odd partial sums must approach the *same* limit $s = s' = s''$. It follows that $\lim_{n \rightarrow \infty} s_n = s$, and the series converges.



Note also that we get the following refined error estimate from the above analysis.

$$\begin{aligned} s_n &< s < s_n + v_{n+1} && \text{if } n \text{ is even} \\ s_n - v_{n+1} &< s < s_n && \text{if } n \text{ is odd.} \end{aligned}$$

Exercises for 8.4.

1. Determine if each of the following series converges.

(a) $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2}$.

(b) $\sum_{n=1}^{\infty} \frac{(-1)^n n}{n+1}$.

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

(d) $\sum_{n=2}^{\infty} \frac{(-1)^n}{\ln n}$.

(e) $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^{1/n}}$.

2. How many terms of the series $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}$ are needed to be sure that the answer is accurate to within 5×10^{-11} . (The series actually adds up to $\ln 2$.)

3. We shall see later that

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

- (a) How many terms of this series are necessary to calculate $1/e$ to within 5×10^{-5} .
 - (b) Do that calculation. (You may want to program it.)
 - (c) Compare with $1/e$ as evaluated on your calculator or by computer.
4. Calculate $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2}$ to within 5×10^{-3} . (The sum of the series is actually $\pi^2/12$.)

8.5 Absolute and Conditional Convergence

A series $\sum_n u_n$ is called *absolutely convergent* if the series $\sum_n |u_n|$ converges. If $\sum_n u_n$ converges, but $\sum_n |u_n|$ diverges, the original series is called *conditionally convergent*. Absolutely convergent series are quite well behaved, and one can manipulate them almost as though they were finite sums. Conditionally convergent series, on the other hand, are often hard to deal with.

Examples The series

$$1 - 1/4 + 1/9 - \cdots + (-1)^{n+1}(1/n^2) + \cdots = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2}$$

is absolutely convergent because $\sum_n 1/n^2$ converges. The series

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

is conditionally convergent. It does converge by the alternating series test, but the series of absolute values $\sum_n 1/n$ diverges.

Of course, for series with non-negative terms, absolute convergence means the same thing as convergence. Also, for series with all terms $u_n \leq 0$, absolute convergence means the same thing as convergence. (Just consider the series $-\sum_n u_n = \sum_n (-u_n)$ for which all the terms are non-negative.)

Theorem 8.13 If $\sum_n |u_n|$ converges, then $\sum_n u_n$ converges. That is, an absolutely convergent series is convergent.

This theorem may look ‘obvious’, but that is the result of the choice of terminology. The relation between the series $\sum_n u_n$ and the corresponding series of absolute values $\sum_n |u_n|$ is rather subtle. For example, even if they both converge, they certainly won’t have the same sum.

The proof. Assume $\sum_n |u_n|$ converges. Define two new series as follows. Let

$$p_n = \begin{cases} u_n & \text{if } u_n > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$q_n = \begin{cases} |u_n| = -u_n & \text{if } u_n < 0 \\ 0 & \text{otherwise} \end{cases}.$$

Then $\sum_n p_n$ is the series obtained when all negative terms of the original series are replaced by 0’s, and $\sum_n q_n$ is the *negative* of the series obtained if all positive terms of the original series are replaced by 0’s. Both are series of non-negative terms, and

$$u_n = p_n + (-q_n) = p_n - q_n \quad \text{for each } n.$$

(Either p_n or $-q_n$ is zero, and the other is u_n .) Hence, to show that $\sum_n u_n$ converges, it would suffice to show that both $\sum_n p_n$ and $\sum_n q_n$ converge. Consider the former. For each n , we have

$$0 \leq p_n \leq |u_n|.$$

Indeed, p_n equals one or the other of the bounds depending on the sign of u_n . Hence, by the comparison test, since $\sum_n |u_n|$ converges, so does $\sum_n p_n$. A similar argument shows that $\sum_n q_n$ converges. Hence, we conclude that $\sum_n u_n$ converges. That completes the proof.

One consequence of the above argument is that for absolutely convergent series, rearranging the terms of the series does not affect convergence or divergence or the sum when convergent. The reason is that for an absolutely convergent series,

$$\sum_n u_n = \sum_n p_n - \sum_n q_n, \quad (128)$$

and each of the series on the right is a convergent series of non-negative terms. For such series, rearranging terms is innocuous. Suppose on the other hand that the two series $\sum_n p_n$ and $\sum_n q_n$ are both divergent. Then (128) would assert that the $\sum_n u_n$ is the difference of two ‘infinities’, and any attempt to make sense of that is fraught with peril. However, in that case

$$\sum_n |u_n| = \sum_n p_n + \sum_n q_n$$

is a sum of two divergent series of non-negative terms and certainly doesn’t converge, so the series $\sum_n u_n$ is not absolutely convergent.

The Ratio Test There is a fairly simple test which will establish absolute convergence for series to which it applies. Consider the ratio

$$\frac{|u_{n+1}|}{|u_n|}$$

of succeeding terms of the series of absolute values. Suppose this approaches a finite limit r . (Note that we must have $r \geq 0$ since it is a limit of non-negative terms.) That means that, for large n , the series $\sum_n |u_n|$ looks very much like a geometric series of the form

$$\sum_n ar^n = a \sum_n r^n.$$

However, the geometric series converges exactly when $0 \leq r < 1$. We should be able to conclude from this that the original series also converges exactly in those circumstances. Unfortunately, because the comparison involves only a limiting ratio, the analysis is not precise enough to tell us what happens when $r = 1$. The precise statement of the test is the following.

The Ratio Test Suppose

$$r = \lim_{n \rightarrow \infty} \frac{|u_{n+1}|}{|u_n|}$$

exists. If $r < 1$, the series $\sum_n u_n$ is absolutely convergent. If $r > 1$, the series $\sum_n u_n$ diverges. If $r = 1$, the test provides no information.

Example 157 Consider the series

$$\sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

(As you may already know, this is the series for e^x .) To apply the ratio test, we need to consider

$$\frac{|x^{n+1}/(n+1)!|}{|x^n/n!|} = \frac{|x|^{n+1}}{(n+1)!} \frac{n!}{|x|^n} = \frac{|x|}{n+1}.$$

The limit of this as $n \rightarrow \infty$ is $r = 0 < 1$. Hence, by the ratio test, the series $\sum_n x^n/n!$ is absolutely convergent for every possible x .

Example 158 Consider the series

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

(As we shall see this is the series for $\ln(1+x)$ at least when it converges.) The ratio test tells us to consider

$$\frac{|(-1)^{n+1}x^{n+1}/(n+1)|}{|(-1)^n x^n/n|} = \frac{|x|^{n+1}}{n+1} \frac{n}{|x|^n} = \frac{n}{n+1} |x|.$$

Since, $\lim_{n \rightarrow \infty} \frac{n}{n+1} = 1$, the limiting ratio in this case is $|x|$. Thus, the series $\sum_{n=1}^{\infty} (-1)^n x^n/n$ converges absolutely if $|x| < 1$ and diverges if $|x| > 1$. Unfortunately, the ratio test does not tell us what happens when $|x| = 1$. However, *we can settle those case by using other criteria*. Thus, for $x = 1$, the series is

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{n}$$

and that converges by the alternating series test. On the other hand, for $x = -1$, the series is

$$\sum_{n=1}^{\infty} (-1)^n \frac{(-1)^n}{n} = \sum_{n=1}^{\infty} \frac{1}{n} = \sum_{n=1}^{\infty} \frac{1}{n}$$

which is the harmonic series, so it diverges.

The detailed proof that the ratio test works is a bit subtle. We include it here for completeness, but you will be excused if you skip it.

The proof. Suppose $\lim_{n \rightarrow \infty} \frac{|u_{n+1}|}{|u_n|} = r$. That means that if r_1 is slightly less than r , and r_2 is slightly greater than r , then for all sufficiently large n

$$r_1 < \frac{|u_{n+1}|}{|u_n|} < r_2.$$

Moreover, by making n large enough, we can arrange for the numbers r_1 and r_2 to be as close to r as we might like.

By the general principle that the tail of a series dominates in discussions of convergence, we can ignore the finite number of terms for which the above inequality does not hold. Renumbering, we may assume

$$r_1|u_n| < |u_{n+1}| < r_2|u_n|$$

holds for all n starting with $n = 1$. Then,

$$|u_2| < r_2|u_1|, |u_3| < r_2|u_2| < r_2^2|u_1|, \dots, |u_n| < |u_1|r_2^{n-1}$$

and similarly for the lower bound, so putting $a = |u_1|$, we may assume

$$ar_1^{n-1} < |u_n| < ar_2^{n-1} \quad (129)$$

for all n . Suppose $r < 1$. Then, by taking r_2 sufficiently close to r , we may assume that $r_2 < 1$. So we may use (129) to compare $\sum_n |u_n|$ to $\sum_n ar_2^{n-1}$ which is a convergent geometric series. Hence, $\sum_n |u_n|$ converges. Suppose on the other hand that $r > 1$. Then by choosing r_2 sufficiently close to r , we may assume $r_2 > 1$. Putting this in (129) yields

$$1 < |u_n|$$

for all n . Hence, $u_n \rightarrow 0$ as $n \rightarrow \infty$ is not possible, and $\sum_n u_n$ must diverge.

Note that the above argument breaks down if $r = 1$. In that case, we cannot assume that $r_2 < 1$ or $r_1 > 1$, so neither part of the argument works.

The Root Test There is a test which is related to the ratio test which is a bit simpler to use. Instead of considering the ratio of successive terms, one considers the n th root $|u_n|^{1/n}$. If this approaches a finite limit r , then roughly speaking, for large n , $|u_n|$ behaves like r^n . Thus, we are led to compare $\sum_n |u_n|$ to the geometric series $\sum_n r^n$. We get convergence or divergence, as in the ratio test, which depends on the value of the limit r . As in the ratio test, the analysis is not precise enough to tell us what happens if that limit is 1.

The Root Test Suppose $\lim_{n \rightarrow \infty} |u_n|^{1/n} = r$. If $0 \leq r < 1$, then $\sum_n u_n$ converges absolutely. If $1 < r$, then $\sum_n u_n$ diverges. If $r = 1$, it may converge or diverge.

Example 159 Consider the series

$$x + \frac{x^2}{4} + \frac{x^3}{9} + \cdots = \sum_{n=1}^{\infty} \frac{x^n}{n^2}.$$

We have

$$\left| \frac{x^n}{n^2} \right|^{1/n} = \frac{|x|}{n^{2/n}}$$

so in this case

$$\lim_{n \rightarrow \infty} |u_n|^{1/n} = |x| \lim_{n \rightarrow \infty} n^{-2/n}.$$

The limit on the right is evaluated by a tricky application of L'Hôpital's rule. The idea is that if $\lim_{n \rightarrow \infty} \ln a_n = L$, then $\lim_{n \rightarrow \infty} a_n = e^L$. In this case,

$$\lim_{n \rightarrow \infty} \ln(n^{-2/n}) = \lim_{n \rightarrow \infty} \left(-\frac{2}{n} \ln n\right) = \lim_{n \rightarrow \infty} \frac{-2 \ln n}{n} = \lim_{n \rightarrow \infty} \frac{-2/n}{1} = 0.$$

(The next to last step was done using L'Hôpital's rule.) Hence,

$$\lim_{n \rightarrow \infty} n^{-2/n} = e^0 = 1.$$

It follows that the limiting ratio for the series is $r = |x|$. Thus, if $|x| < 1$ the series $\sum_n \frac{x^n}{n^2}$ converges absolutely, and if $|x| > 1$, it diverges. For $|x| = 1$, we *must settle the question by other means*. For $x = 1$, the series is $\sum_n 1/n^2$ which we know converges. (It is a ' p -series' with $p = 2 > 1$.) For $x = -1$, the series is $\sum_n (-1)^n/n^2$, and we just decided its series of absolute values converges. Hence, $\sum_n (-1)^n/n^2$ converges absolutely. (You could also see it converges by the alternating series test, but absolute convergence is stronger.)

The proof of the root test is similar to that of the ratio test. Again you will be excused if you skip it, but here it is.

The proof. Suppose $\lim_{n \rightarrow \infty} |u_n|^{1/n} = r$. If r_1 is slightly less than r and r_2 is slightly larger, then, for n sufficiently large,

$$r_1 < |u_n|^{1/n} < r_2.$$

Again, by the principle that we may ignore finitely many terms of a series when investigating convergence, we may assume the inequalities holds for all n . Raising to the n power, we get

$$r_1^n < |u_n| < r_2^n.$$

If $r < 1$, we may assume $r_2 < 1$, and compare $\sum_n |u_n|$ with a convergent geometric series. On the other hand, if $1 < r$, we may assume $1 < r_2$, and then $|u_n| \rightarrow \infty$ which can't happen for a convergent series.

Exercises for 8.5.

1. In each case, tell if the indicated series is absolutely convergent, conditionally convergent, or divergent. If the ratio or root test does not work, you may want to try some other method.

(a) $\sum_{n=1}^{\infty} t^n$ where $|t| < 1$.

(b) $\sum_{n=1}^{\infty} \frac{(-1)^n(n+4)}{5n+6}$.

(c) $\sum_{n=1}^{\infty} \frac{(-1)^{5n+2}t^n}{n^2}$ where $|t| < 1$.

$$(d) \sum_{n=2}^{\infty} \frac{(-1)^n}{\ln n}.$$

$$(e) \sum_{n=1}^{\infty} \frac{(-2)^n}{n^n}.$$

$$(f) \sum_{n=0}^{\infty} \frac{(-1)^n n!}{n^n}. \text{ Hint: You should know what } \lim_{n \rightarrow \infty} (1 + 1/n)^n \text{ is from your previous courses.}$$

$$(g) \sum_{n=1}^{\infty} \frac{n+1}{2n^3 + 4n + 5}.$$

2. Determine all values of t such that $\sum_{n=0}^{\infty} \frac{t^n}{\sqrt{n+1}}$ converges. For which of those values does it converge absolutely?

8.6 Power Series

At the beginning of this chapter, we saw that we might want to consider series of the form $\sum_{n=0}^{\infty} a_n t^n$ as solutions of differential equations. A bit more generally, if we were given initial conditions at $t = t_0$, we might consider series of the form

$$\sum_{n=0}^{\infty} a_n (t - t_0)^n.$$

Such a series is called a *power series* centered at t_0 .

A power series will generally converge for some, but perhaps not all, values of the variable. Usually, you can determine those values for which it converges by an application of the ratio or root test.

Example 160 Consider the power series

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

To apply the ratio test, consider

$$\frac{\frac{n+1}{2^{n+1}} |t-3|^{n+1}}{\frac{n}{2^n} |t-3|^n} = \frac{|t-3|}{2} \frac{n+1}{n}.$$

However,

$$\lim_{n \rightarrow \infty} \frac{n+1}{n} = 1$$

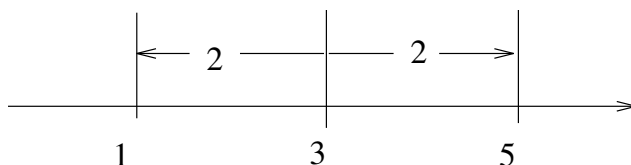
(by L'Hôpital's rule, or by dividing numerator and denominator by n). Hence, the limiting ratio is $|t - 3|/2$. Hence, the series converges absolutely if

$$\begin{aligned}\frac{|t - 3|}{2} &< 1 \\ \text{i.e., } |t - 3| &< 2 \\ \text{i.e., } -2 &< t - 3 < 2 \\ \text{i.e., } 1 &< t < 5.\end{aligned}$$

Similarly, if $|t - 3|/2 > 1$, the series diverges. That inequality amounts to $t < 1$ or $5 < t$. The case $|t - 3|/2 = 1$, i.e., $t = 1$ or $t = 5$, must be handled separately. I leave the details to you. It turns out that the series diverges both for $t = 1$ and $t = 5$. Hence, the exact range of convergence is

$$1 < t < 5$$

and the series converges absolutely on this interval.

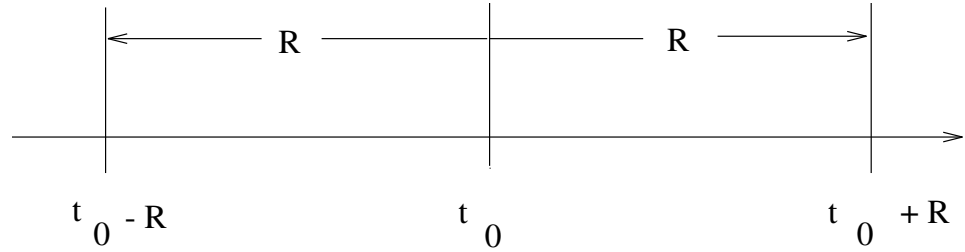


The above analysis could have been done equally well using the root test.

The behavior exhibited in Example 160, or in similar examples in the previous section, is quite general. For any power series $\sum_n a_n(t - t_0)^n$ there is a number R such that the series converges absolutely in the interval

$$t_0 - R < t < t_0 + R,$$

diverges outside that interval, and may converge or diverge at either endpoint. R is called the *radius of convergence* of the series. (The terminology comes from the corresponding concept in complex analysis where the condition $|z - z_0| < R$ characterizes all points in the complex plane inside a circular disk of radius R centered at z_0 .) R may be zero, in which case the series converges only for $t = t_0$ (where all terms but the first are 0). R may be infinite, in which case the series converges for all t . In many cases, it is finite, and it may be important to know its value.



Understanding how the ratio or root test is used to determine the radius of convergence gives you a good idea of why a power series converges in a connected interval rather than at a random collection of points, so you should work enough examples to be sure you can do such calculations. Unfortunately, in the general case, neither the ratio nor the root test may apply, so the argument justifying the existence of an interval of convergence is a bit tricky.

We shall outline the argument here, but you need not study it at this time if the examples satisfy you that it all works.

The argument why the range of convergence is an interval. If the series converges only for $t = t_0$, then we take $R = 0$, and we are done. Suppose instead that there is a $t_1 \neq t_0$ at which the series $\sum_n a_n(t_1 - t_0)^n$ converges. Then the general term

$$a_n(t_1 - t_0)^n \rightarrow 0$$

as $n \rightarrow \infty$. One consequence is that the absolute value of the general term must be bounded, i.e., there is a bound M such that

$$|a_n||t_1 - t_0|^n < M$$

for all n . Put $R_1 = |t_1 - t_0|$. Then, if $|t - t_0| < R_1$,

$$|a_n||t - t_0|^n = |a_n||t_1 - t_0|^n \left(\frac{|t - t_0|}{R_1} \right)^n < Mr^n$$

where $r = |t - t_0|/R_1 < 1$. Comparing with a geometric series, we see that $\sum_n |a_n||t - t_0|^n$ converges, so $\sum_n a_n(t - t_0)^n$ is absolutely convergent as long as $|t - t_0| < R_1$.

Consider next all possible $R_1 > 0$ such that the series converges absolutely in the range $t_0 - R_1 < t < t_0 + R_1$. If there is no upper bound to this set of numbers, then the series converges absolutely for all t . In that case, we set the radius of convergence $R = \infty$. Suppose instead that there is some upper bound to the set of all such R_1 . By a variation of the completeness arguments we have used before (concerning bounded increasing sequences), it is possible to show in this case that there is a single value R such that $\sum_n a_n(t - t_0)^n$ converges absolutely for $|t - t_0| < R$, but that the series diverges for $|t - t_0| > R$. This R is the desired radius of convergence.

Differentiation and Integration of power Series The rationale behind the use of series to solve differential equations is that they are generalizations of polynomial functions. By the usual rules of calculus, polynomial functions are easy to differentiate or integrate. For example, if

$$f(t) = a_0 + a_1(t - t_0) + a_2(t - t_0)^2 + \cdots + a_n(t - t_0)^n + \cdots$$

then

$$f'(t) = 0 + a_1 \cdot 1 + 2a_2(t - t_0) + \cdots + na_n(t - t_0)^{n-1} + \cdots$$

That is, to differentiate a polynomial (or any finite sum), you differentiate each term, and then add up the results. A similar rule works for integration. Unfortunately, these rules do not always work for infinite sums. It is possible for the derivatives of the terms of a series to add up to something other than the derivative of the sum of the series.

Example Consider the series

$$\sum_{n=1}^{\infty} \left(\frac{\sin nt}{n} - \frac{\sin(n+1)t}{n+1} \right).$$

The partial sums are

$$\begin{aligned} s_1(t) &= \sin t - \frac{1}{2} \sin 2t \\ s_2(t) &= \sin t - \frac{1}{2} \sin 2t + \frac{1}{2} \sin 2t - \frac{1}{3} \sin 3t = \sin t - \frac{1}{3} \sin 3t \\ &\vdots \\ s_n(t) &= \cdots = \sin t - \frac{1}{n+1} \sin(n+1)t \\ &\vdots \end{aligned}$$

However, $\frac{|\sin(n+1)t|}{n} \leq \frac{1}{n}$ for any t , so its limit is 0 as $n \rightarrow \infty$. Hence,

$$\lim_{n \rightarrow \infty} s_n(t) = \sin t$$

so the series converges for every t and its sum is $\sin t$. On the other hand, the series of derivatives is

$$\sum_{n=1}^{\infty} (\cos nt - \cos(n+1)t).$$

The partial sums of this series are calculated essentially the same way, and

$$s'_n(t) = \cos t - \cos(n+1)t.$$

For most values of t , $\cos(n+1)t$ does not approach a definite limit as $n \rightarrow \infty$. (For example, try $t = \pi/2$. You alternately get 0, -1 , or 1 for different values of n .) Hence, the series of derivatives is generally not even a convergent series.

Even more bizarre examples exist in which the series of derivatives converges but to something other than the derivative of the sum of the original series.

Similar problems arise when you try to integrate a series term by term. Some of this will be discussed when you study Fourier Series and Boundary Value Problems, next year. Fortunately, for a power series, *within its interval of absolute convergence* the derivative of the sum is the sum of the derivatives and similarly for integrals. This fact is fundamental for any attempt to use power series to solve a differential equation. It may also be used to make a variety of other calculations with series.

Theorem 8.14 Suppose $\sum_{n=0}^{\infty} a_n(t-t_0)^n$ converges absolutely to $f(t)$ for $|t-t_0| < R$. Then

$$(a) \quad f'(t) = \sum_{n=1}^{\infty} n a_n(t-t_0)^{n-1} \text{ for } |t-t_0| < R.$$

$$(b) \quad \int_{t_0}^t f(s) ds = \sum_{n=0}^{\infty} \frac{a_n}{n+1} (t-t_0)^{n+1} \text{ for } |t-t_0| < R. \quad \text{Moreover, the series } \sum_{n=1}^{\infty} n a_n(t-t_0)^{n-1} \text{ and } \sum_{n=0}^{\infty} \frac{a_n}{n+1} (t-t_0)^{n+1} \text{ have the same radius of convergence as } \sum_{n=0}^{\infty} a_n(t-t_0)^n.$$

The proof of this theorem is a bit involved. An outline is given in the appendix to this section, which you may want to skip.

Example 161 We know

$$\frac{1}{1-t} = \sum_{n=0}^{\infty} t^n \quad \text{for } |t| < 1.$$

(1 is the radius of convergence.) Hence,

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

The last expression was obtained by substituting $n+1$ for n . This has the effect of changing the lower limit ' $n=1$ ' to ' $n+1=1$ ' or ' $n=0$ '. Such substitutions are often useful when manipulating power series, particularly in the solution of differential equations.

The differentiated series also converges absolutely for $|t| < 1$.

Example 162 If replace t by $-t$ in the formula for the sum of a geometric series, we get

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

and this converges absolutely for $|-t| = |t| < 1$. Hence, by (b), we get

$$\int_0^t \frac{ds}{1+s} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n+1} t^{n+1} = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} t^n$$

and this is valid for $|t| < 1$. Note the shift obtained by replacing n by $n-1$. Doing the integral on the left yields

$$\ln(1+t) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} t^n = t - \frac{t^2}{2} + \frac{t^3}{3} + \dots \quad \text{for } |t| < 1.$$

Series of this kind may be used to make numerical calculations. For example, suppose we want to calculate $\ln 1.01$ to within 5×10^{-6} , i.e., *loosely speaking*, we want the answer to be accurate to five decimal places. We can use the series

$$\ln(1+.01) = .01 - \frac{(.01)^2}{2} + \frac{(.01)^3}{3} + \dots + (-1)^{n+1} \frac{(.01)^n}{n} + \dots$$

and include enough terms so the error R_n satisfies $|R_n| < 5 \times 10^{-6}$. Since the series is an alternating series with decreasing terms, we know that $|R_n|$ is bounded by the absolute value of the next term in the series, i.e., by $\frac{(.01)^{n+1}}{n+1}$. Hence, to get the desired accuracy, it would suffice to choose n large enough so that

$$\frac{(.01)^{n+1}}{n+1} = \frac{10^{-2(n+1)}}{n+1} < 5 \times 10^{-6}.$$

There is no good way to determine such an n by a deductive process, but trial and error works reasonably well. Certainly, $n = 2$ would be good enough. Let's see if $n = 1$ would also work.

$$\frac{10^{-4}}{2} = .5 \times 10^{-4} = 5 \times 10^{-5},$$

and that is not small enough. Hence, $n = 2$ is the best that this particular estimate of the error will give us. Hence,

$$\ln(1.01) = .01 - \frac{(.01)^2}{2} \approx 0.00995.$$

You should check this with your calculator to see if it is accurate. (Remember however that the calculator is also using some approximation, and it doesn't know the true value any better than you do.)

The estimate R_n of the error after using n terms of a series is tricky to determine. For alternating series, we know that the next term rule applies, but we shall see examples later of where much more sophisticated methods need to be used.

Appendix. Proof of the Theorem

We first note that it is true that the sum of a power series within its radius of convergence is always a *continuous function*. The proof is not extremely hard, but we shall omit it here. You may see some further discussion of this point in your later mathematics courses. We shall use this fact implicitly several places in what follows. In particular, knowing the sum of a series is continuous allows us to conclude it is integrable and also to apply the fundamental theorem of calculus.

Statement (a) in the theorem may be proved once statement (b) has been established. To see this, argue as follows. Let $\sum_{n=0}^{\infty} a_n(t-t_0)^n = f(t)$ have radius of convergence $R > 0$. Consider the differentiated series $\sum_{n=1}^{\infty} na_n(t-t_0)^{n-1}$. If the ratio (or root) test applies, it is not hard to see that this series has the same radius of convergence R as the original series. (See the Exercises.) If the ratio test does not apply, there is a tricky argument to establish this fact in any case. Suppose that point is settled. Define

$$g(t) = \sum_{n=1}^{\infty} na_n(t-t_0)^{n-1} \quad \text{for } |t-t_0| < R.$$

(We hope that $g(t) = f'(t)$, but we don't know that yet.) Assume (b) is true and apply it to the series for $g(t)$. We get

$$\int_{t_0}^t g(s)ds = \sum_{n=1}^{\infty} \frac{n}{n} a_n(t-t_0)^n = \sum_{n=1}^{\infty} a_n(t-t_0)^n.$$

However,

$$\begin{aligned} f(t) &= \sum_{n=0}^{\infty} a_n(t-t_0)^n = a_0 + \sum_{n=1}^{\infty} a_n(t-t_0)^n \\ &= a_0 + \int_{t_0}^t g(s)ds. \end{aligned}$$

This formula, together with the fundamental theorem of calculus assures us that $f(t)$ is a differentiable function and

$$f'(t) = 0 + \frac{d}{dt} \int_{t_0}^t g(s)ds = g(t)$$

as needed. The proof of (b) is harder. Let

$$f(t) = \sum_{n=0}^{\infty} a_n(t-t_0)^n \quad \text{for } |t-t_0| < R.$$

Fix $t > t_0$. (The argument for $t < t_0$ is similar.) We want to integrate $f(s)$ over the range $t_0 \leq s \leq t$. For any given N , we may decompose the series for $f(s)$ into two parts

$$f(s) = \sum_{n=0}^N a_n(s-t_0)^n + \underbrace{\sum_{n=N+1}^{\infty} a_n(s-t_0)^n}_{R_N(s)}.$$

Integrate both sides to obtain

$$\begin{aligned}\int_{t_0}^t f(s)ds &= \sum_{n=0}^N \int_{t_0}^t a_n(s-t_0)^n ds + \int_{t_0}^t R_N(s)ds \\ &= \sum_{n=0}^N \frac{a_n}{n+1} (s-t_0)^{n+1} \Big|_{t_0}^t + \int_{t_0}^t R_N(s)ds \\ &= \sum_{n=0}^N \frac{a_n}{n+1} (t-t_0)^{n+1} + \int_{t_0}^t R_N(s)ds.\end{aligned}$$

To complete the proof, we need to see what happens in the last equality as $N \rightarrow \infty$. The first term is the N th partial sum of the desired series, so it suffices to show that the second error term approaches zero. However,

$$\left| \int_{t_0}^t R_N(s)ds \right| \leq \int_{t_0}^t |R_N(s)|ds.$$

On the other hand, we have

$$\begin{aligned}|R_N(s)| &= \left| \sum_{n=N+1}^{\infty} a_n(s-t_0)^n \right| \leq \sum_{n=N+1}^{\infty} |a_n||s-t_0|^n \\ &\leq \sum_{n=N+1}^{\infty} |a_n||t-t_0|^n.\end{aligned}$$

(This follows by the same reasoning as in the comparison test.) Note that since $|t-t_0| < R$, the series $\sum_n a_n(t-t_0)^n$ converges *absolutely*, so the series on the right of the above inequality is the tail of a convergent series; call it T_N . Then, we have

$$|R_N(s)| \leq T_N$$

for $|s-t_0| < |t-t_0|$, where T_N is independent of s . Hence,

$$\left| \int_{t_0}^t R_N(s)ds \right| \leq \int_{t_0}^t |R_N(s)|ds \leq \int_{t_0}^t T_N ds = T_N(t-t_0).$$

Since T_N is the tail of a convergent series, it goes to zero as $N \rightarrow \infty$. That completes the proof.

Exercises for 8.6.

1. Show that the power series $\sum_{n=0}^{\infty} n!t^n$ has radius of convergence $R = 0$.

2. Find the interval of convergence and the radius of convergence for each of the following power series.

(a) $\sum_{n=0}^{\infty} nt^n$.

(b) $\sum_{n=0}^{\infty} \frac{3^n}{n^2} (t-5)^n$.

(c) $\sum_{n=0}^{\infty} \frac{n!}{(2n)!} (t-1)^n$.

(d) $\sum_{n=0}^{\infty} \frac{n^2}{10^n} (t+1)^n$.

(e) $\sum_{n=0}^{\infty} \frac{2^n}{(2n)!} t^{2n}$.

3. Using the geometric series

$$f(t) = \frac{1}{1+t} = \sum_{n=0}^{\infty} (-1)^n t^n,$$

find series expansions for $-f'(t) = \frac{1}{(1+t)^2}$ and $\frac{f''(t)}{2} = \frac{1}{(1+t)^3}$. For which values of t does the theorem in the section assure you that these expansions are valid?

4. Find series expansion for $f(t) = t \ln(1+t)$ and $g(t) = \frac{\ln(1+t)}{t}$. What are the intervals of convergence of these series?

5. Assume the expansion

$$f(t) = t - \frac{t^3}{3} + \frac{t^5}{5} - \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n t^{2n+1}}{2n+1}$$

is valid for $-1 < t < 1$. Show that $f'(t) = \frac{1}{1+t^2}$. Given that $f(0) = 0$, conclude $f(t) = \tan^{-1} t$.

6. Assume that the expansion

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

is valid for all t . Show that $f'(t) = f(t)$. Given that $f(0) = 1$, what can you conclude about $f(t)$?

7. Suppose the ratio test applies to the power series $\sum_n a_n t^n$ and it has radius of convergence R . Show that the series $\sum_n n a_n t^{n-1}$ and $\sum_n \frac{a_n}{n+1} t^{n+1}$ also have radius of convergence R .

8. Let $s_n(t) = ne^{-nt}$. (You may assume that $s_n(t)$ is the n th partial sum of an appropriate series.) Show that $\lim_{n \rightarrow \infty} s_n(t) = 0$ for $t \neq 0$. Show on the other hand that $\int_0^1 s_n(t) dt = 1 - e^{-n}$. Conclude that

$$\int_0^1 \lim_{n \rightarrow \infty} s_n(t) dt \neq \lim_{n \rightarrow \infty} \int_0^1 s_n(t) dt$$

in this case.

8.7 Analytic Functions and Taylor Series

Our aim, as enunciated at the beginning of this chapter, is to use power series to solve differential equations. So suppose that

$$f(t) = \sum_{n=0}^{\infty} a_n(t-t_0)^n \quad \text{for } |t-t_0| < R,$$

where R is the radius of convergence of the series on the right. (Note that this assumes that $R > 0$. There are power series with $R = 0$. Such a series converges only at $t = t_0$ and won't be of much use. See the Exercises for the previous section for an example.) A function $f : \mathbf{R} \rightarrow \mathbf{R}$ is said to be *analytic* if at each point t_0 in its domain, it may be represented by such a power series in an interval about t_0 . Analytic functions are the best possible functions to use in applications. For example, we know by the previous section, that such a function is differentiable. Even better, its derivative is also analytic because it has a power series expansion with the same radius of convergence. Hence, the function also has a second derivative, and by extension of this argument, *must have derivatives of every order*. Moreover, it is easy to relate the coefficients of the power series to these derivatives. We have

$$\begin{aligned} f(t) &= \sum_{n=0}^{\infty} a_n(t-t_0)^n = a_0 + a_1(t-t_0) + a_2(t-t_0)^2 + \dots \\ f'(t) &= \sum_{n=1}^{\infty} n a_n(t-t_0)^{n-1} = a_1 + 2a_2(t-t_0) + \dots \\ f''(t) &= \sum_{n=2}^{\infty} n(n-1)a_n(t-t_0)^{n-2} = 2a_2 + \dots \\ &\vdots \\ f^{(k)}(t) &= \sum_{n=k}^{\infty} n(n-1)(n-2)\dots(n-k+1)a_n(t-t_0)^{n-k} = k(k-1)\dots 2 \cdot 1 a_k + \dots \\ &\vdots \end{aligned}$$

Here $f^{(k)}$ denotes the k th derivative of f . Note that the series for $f^{(k)}(t)$ starts off with $k!a_k$. Make sure you understand why! Now put $t = t_0$ in the above formulas. All terms involving $t - t_0$ to a positive power vanish, and we get

$$\begin{aligned} f(t_0) &= a_0 \\ f'(t_0) &= a_1 \\ f''(t_0) &= 2a_2 \\ &\vdots \\ f^{(k)}(t_0) &= k!a_k \\ &\vdots \end{aligned}$$

We can thus solve for the coefficients

$$\begin{aligned} a_0 &= f(t_0) \\ a_1 &= f'(t_0) \\ a_2 &= \frac{f''(t_0)}{2} \\ &\vdots \\ a_k &= \frac{f^{(k)}(t_0)}{k!} \\ &\vdots \end{aligned}$$

Hence, within the radius of convergence, the power series expansion of $f(t)$ is

$$f(t) = \sum_{n=0}^{\infty} \frac{f^{(n)}(t_0)}{n!} (t - t_0)^n.$$

This series is called the *Taylor series* of the function f .

The above series was derived under the assumption that f is analytic, but we can form that series for any function whatsoever, provided derivatives of all orders exist at t_0 . If the function is analytic, then it will equal its Taylor series within its radius of convergence.

Example 163 Let $f(t) = e^t$ and let $t_0 = 0$. Then,

$$f'(t) = e^t, f''(t) = e^t, \dots, f^{(n)}(t) = e^t, \dots,$$

so

$$a_n = \frac{e^0}{n!} = \frac{1}{n!}.$$

Hence, the Taylor series for e^t at 0 is

$$\sum_{n=0}^{\infty} \frac{1}{n!} t^n.$$

It is easy to determine the radius of convergence of this series by means of the ratio test.

$$\frac{|t|^{n+1}/(n+1)!}{|t|^n/n!} = \frac{|t|}{n+1} \rightarrow 0 < 1.$$

Hence, the series converges for all t and the radius of convergence is infinite. We shall see that e^t equals the series for every t , so $f(t) = e^t$ defines an analytic function on \mathbf{R} . However, this need not always be the case. For example, we might have $f(t)$ equal to the Taylor series for some but not all values in the interval of convergence of that series.

Example 164 Let $f(t) = \cos t$ and let $t_0 = 0$. Then, $f'(t) = -\sin t$, $f''(t) = -\cos t$, $f'''(t) = \sin t$, and $f^{(4)}(t) = \cos t$. This pattern then repeats indefinitely with a period of 4. In particular,

$$f^{(n)}(0) = \begin{cases} 1 & \text{if } n \text{ is even and a multiple of 4} \\ -1 & \text{if } n \text{ is even and not a multiple of 4} \\ 0 & \text{otherwise} \end{cases}.$$

Hence, the Taylor series is

$$1 - \frac{t^2}{2} + \frac{t^4}{4!} - \frac{t^6}{6!} + \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{t^{2n}}{(2n)!}.$$

As above, a simple application of the ratio test shows that this series has infinite radius of convergence. We shall see below that $\cos t$ is in fact analytic and equals its Taylor series for all t .

Example 165 We showed in the previous section that the expansion

$$\ln(1+t) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{t^n}{n}$$

is valid for t in the interval of convergence of the series, ($|t| < 1$). This tells us that the function $\ln(1+t)$ is certainly analytic for $-1 < t < 1$, and the series on the right is its Taylor series.

If we substitute $t-1$ for t in that expansion we get

$$\ln t = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{(t-1)^n}{n}$$

and this gives us the Taylor series of $\ln t$ for $t_0 = 1$.

Taylor's Formula with Remainder Because analytic functions are so important, it is useful to have a mechanism for determining when a function is equal to its Taylor series. To this end, let

$$R_N(t) = f(t) - \sum_{n=0}^N \frac{f^{(n)}(t_0)}{n!} (t-t_0)^n.$$

The *remainder*, $R_N(t)$, is the difference between the value of the function and the value of the appropriate partial sum of its Taylor series. To say the series converges to the function is to say that the limit of the sequence of partial sums is $f(t)$, i.e., that $R_N(t) \rightarrow 0$. To determine if this is the case, it is worthwhile having a method for estimating $R_N(t)$. This is provided by the following formula which is called the *Cauchy form of the remainder*.

$$R_N(t) = \frac{1}{N!} \int_{t_0}^t (t-s)^N f^{(N+1)}(s) ds. \quad (130)$$

This formula is in fact valid as long as the $f^{(N+1)}(s)$ exists and is continuous in the interval from t_0 to t . We shall derive this formula later in this section, but it is instructive to look at the first case $N = 0$. Here

$$f(t) = f(t_0) + R_0(t)$$

where $R_0(t) = \int_{t_0}^t f'(s) ds$

so the formula tells us that the change $f(t) - f(t_0)$ in the function is the integral of its rate of change $f'(s)$. The general case may be considered an extension of this for higher derivatives, but just why the factor $(t-s)^n$ comes up in the integral won't be clear until you see the proof.

In most cases, formula (130) isn't much use as written. The point is that we want to use the partial sum

$$\sum_{n=0}^N \frac{f^{(n)}(t_0)}{n!} (t-t_0)^n$$

as an approximation to the function value $f(t)$. If we could calculate the remainder term $R_N(t)$ exactly, we could also calculate $f(t)$ exactly, and there would be no need to use an approximation. Hence, in most cases, we want to get an *upper bound* on the size of $R_N(t)$ rather than an exact value. *Suppose it is known that*

$$|f^{(N+1)}(s)| \leq M$$

for s in the interval between t_0 and t . Then replacing $|f^{(N+1)}(s)|$ by M in (130) yields (for $t > t_0$)

$$|R_N(t)| \leq \frac{1}{N!} \int_{t_0}^t (t-s)^N M ds = -\frac{M}{N!} \frac{(t-s)^{N+1}}{N+1} \Big|_{t_0}^t = \frac{M}{(N+1)!} (t-t_0)^{N+1}.$$

A similar argument works for $t < t_0$ except for some fiddling with signs. The result which holds in both cases is

$$|R_N(t)| \leq \frac{M}{(N+1)!} |t-t_0|^{N+1}. \quad (131)$$

The expression on the right is what we get from the *next term* of the Taylor series if we replace $f^{(N+1)}(t_0)$ by the maximum value of that derivative in the interval from t_0 and t .

Example 166 Consider the Taylor series for $f(t) = \cos t$ centered at $t_0 = 0$

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

Only even numbered terms appear in this series. We want to estimate the error $R_N(t)$ using the inequality (131). We might as well take $N = 2K + 1$ odd. (Why?) We know the even derivatives are all $\pm \cos t$, so we can take $M = 1$ as an upper bound for $|f^{(N+1)}(s)| = |f^{(2K+2)}(s)|$. Hence, the estimate for the remainder is

$$|R_{2K+1}(t)| \leq \frac{1}{(2K+2)!} |t|^{2K+2}.$$

(Note that in this case the estimate on the right is just the absolute value of the next term in the series.) Thus for $K = 5, t = 1$ radian, we get

$$|R_{11}| \leq \frac{1}{12!} 1^{12} = 2.08768 \times 10^{-9}.$$

Because the odd numbered terms of the series are zero, the number of non-zero terms in the approximating sum is 6. According to Mathematica

$$\sum_{n=0}^5 (-1)^n \frac{1}{(2n)!} = 0.540302304$$

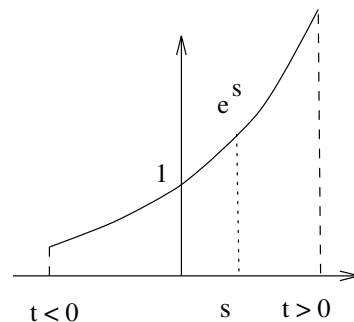
to 9 decimal places. Use your calculator to see if it agrees that this result is as accurate as the above error estimate suggests.

Example 167 Take $f(t) = \sin t$ and $t_0 = 0$. Calculations as above lead to the Taylor series

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

All the even numbered terms are zero. We get the estimate for the remainder for $N = 2K + 2$ —(why don't we use $N = 2K + 1$)—

$$|R_{2K+2}(t)| \leq \frac{|t|^{2K+3}}{(2K+3)!}.$$



Example 168 Take $f(t) = e^t$ and $t_0 = 0$. The Taylor series is

$$\sum_{n=0}^{\infty} \frac{t^n}{n!}.$$

The estimate of the remainder, however, is a bit more involved. The crucial parameter M is the upper bound of $|f^{(N+1)}(s)| = e^s$ for s in the interval from 0 to t , but this depends on the sign of t . Suppose first that $t < 0$. Then, since e^s is an

increasing function of s , its maximum value is attained at the right endpoint, i.e., at $s = 0$. Hence, we should take $M = e^0 = 1$. Thus,

$$|R_N(t)| \leq \frac{1}{(N+1)!} |t|^{N+1}$$

is a plausible estimate for the remainder. On the other hand, if $t > 0$, the same reasoning shows that we should take $M = e^t$. This seems a bit circular, since it is e^t that we are trying to calculate. However, there is nothing preventing us from using an M *larger than* e^t if it is easier to calculate. For example, suppose we want to compute $e = e^1$ using the series

$$\sum_{n=0}^{\infty} \frac{1}{n!} 1^n = \sum_{n=0}^{\infty} \frac{1}{n!}.$$

The maximum value of e^s in this case would be $e^1 = e$, but that is certainly less than 3. Hence, we take $M = 3$ and use the error estimate

$$|R_N(1)| \leq \frac{3}{(N+1)!} |1|^{N+1}.$$

Thus, for $N = 15$ (actually 16 terms), we could conclude

$$|R_{15}(1)| \leq \frac{3}{16!} = 1.43384 \times 10^{-13}.$$

According to Mathematica, the sum for $N = 15$ is 2.71828182846. You should try this on your calculator to see if it agrees.

We may use the estimate of the error to determine values of t for which the Taylor series converges to the function. The following lemma is useful in that context.

Lemma 8.15 For any number t ,

$$\lim_{n \rightarrow \infty} \frac{t^n}{n!} = 0.$$

Proof. We saw earlier (by the ratio test) that the Taylor series for e^t

$$\sum_{n=0}^{\infty} \frac{t^n}{n!}$$

has infinite radius of convergence. Hence, it converges for all t , and its general term $t^n/n!$ must approach 0. \square

For each of the functions we considered above, the estimate of the remainder involved $|t|^n/n!$ or something related to it. Hence, the Lemma tells us $R_N(t) \rightarrow 0$ as

$N \rightarrow \infty$ in each of these cases, for any t . Thus, we have the following Taylor series expansions, valid for all t ,

$$\begin{aligned}\cos t &= \sum_{n=0}^{\infty} (-1)^n \frac{t^{2n}}{(2n)!} \\ \sin t &= \sum_{n=0}^{\infty} (-1)^n \frac{t^{2n+1}}{(2n+1)!} \\ e^t &= \sum_{n=0}^{\infty} \frac{t^n}{n!}.\end{aligned}$$

In principle, each of these series may be used to calculate the function to any desired degree of accuracy by using sufficiently many terms. However, in practice this is only useful for relatively small values of t . For example, we saw that

$$|R_{2K+1}(t)| \leq \frac{|t|^{2K+2}}{(2K+2)!}$$

is a plausible estimate of the error when using the Taylor series for $\cos t$. However, here are some values for $t = 10$

n	$t^n/(2n)!$
1	50
2	416.667
3	1388.89
4	2480.16
5	2755.73
6	2087.68
	\vdots
10	41.1032
	\vdots
20	1.2256210^{-8}
	\vdots

As you see, the terms can get quite large before the factorial term in the denominator begins to dominate. Using the series to calculate $\cos t$ for large t might lead to some nasty surprises. (Can you think of a better way to do it?)

Derivation of the Cauchy Remainder We assume that all the needed derivatives exist and are continuous functions. Consider the expression

$$R_N(t, s) = f(t) - f(s) - f'(s)(t-s) - \frac{f''(s)}{2}(t-s)^2 - \cdots - \frac{f^{(N)}(s)}{N!}(t-s)^N$$

as a function of both t and s . Take the derivative of both sides with respect to s . Note that by the product rule, the derivative of a typical term is

$$\frac{\partial}{\partial s} \left(-\frac{f^{(n)}(s)}{n!} (t-s)^n \right) = -\frac{f^{(n+1)}(s)}{n!} (t-s)^n + \frac{f^{(n)}(s)}{(n-1)!} (t-s)^{n-1}$$

If we write these terms in the reverse order, we get

$$\begin{aligned} \frac{\partial R_N}{\partial s} &= 0 - f'(s) + f'(s) - f''(s)(t-s) + f''(s)(t-s) - \cdots - \frac{f^{(N+1)}(s)}{N!} (t-s)^N \\ &= -\frac{f^{(N+1)}(s)}{N!} (t-s)^N. \end{aligned}$$

since all terms except the last cancel. Integrating with respect to s , we obtain

$$\int_{t_0}^t \frac{\partial R_N}{\partial s} ds = -\frac{1}{N!} \int_{t_0}^t f^{(N+1)}(s) (t-s)^N ds.$$

However, the integral on the left is

$$R_N(t, s)|_{s=t_0}^{s=t} = R_N(t, t) - R_N(t, t_0).$$

Since $R_N(t, t) = f(t) - f(t) - 0 - \cdots - 0 = 0$, it follows that

$$-R_N(t, t_0) = -\frac{1}{N!} \int_{t_0}^t f^{(N+1)}(s) (t-s)^N ds,$$

which is the desired formula except for the minus signs.

Determining the Radius of Convergence of a Taylor Series We have seen that generally speaking the Taylor series expansion

$$f(t) = \sum_{n=0}^{\infty} \frac{f^{(n)}(t_0)}{n!} (t-t_0)^n$$

is valid in some interval $t_0 - R < t < t_0 + R$. There is a simple rule for determining R from f , which applies in many cases.

Consider the example

$$f(t) = \frac{1}{1-t} = 1 + t + t^2 + \cdots + t^n + \cdots$$

We know the series on the right converges absolutely for $-1 < t < 1$ and that the expansion of the function is valid in this range. ($t_0 = 0, R = 1$.) You will note that the function in this case has a singularity at $t = 1$. That suggests the following rule:

Radius of Convergence of a Taylor Series For each point t_0 , let R be the distance to the nearest singularity of the function f . Then the radius of convergence

of the Taylor series for f centered at t_0 is R , and the series converges absolutely to the function in the interval $t_0 - R < t < t_0 + R$.

Unfortunately, it is easy to come up with an example for which this rule appears to fail. Put $u = -t^2$ in the formula

$$\frac{1}{1-u} = 1 + u + u^2 + \cdots + u^n + \cdots$$

to get

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

This is valid for $|u| = |-t^2| < 1$, i.e., $|t| < 1$. Cleaning up the minus signs yields

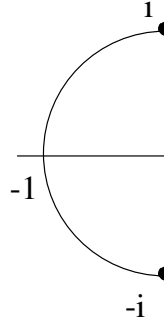
$$\frac{1}{1+t^2} = 1 - t^2 + t^4 - t^6 + \cdots + (-1)^n t^{2n} + \cdots$$

for $-1 < t < 1$. The function $f(t) = 1/(1+t^2)$ does not appear to have a singularity either at $t = 1$ or $t = -1$, which contradicts the rule. However, if you consider the function of a *complex* variable z defined by

$$f(z) = \frac{1}{1+z^2}$$

this function does have singularities at $z = \pm i$ where the denominator vanishes. Also, very neatly, the distance in the complex plane from $z_0 = 0$ to each of these singularities is $R = 1$, and that is the proper radius of convergence of the series. *Thus the rule does apply if we extend the function $f(t)$ to a function $f(z)$ of a complex variable and look for the singularities of that complex function.* In many simple cases, it is obvious how to do this, but in general, the notion of singularity for functions $f(z)$ of a complex variable and how to go about extending functions $f(t)$ of a real variable requires extensive theoretical discussion. This will be done in your complex analysis course. We felt it necessary, however, to enunciate the rule, even though it could not be fully explained, because it is so important in studying the behavior of series solutions of differential equations. It is one of many circumstances in which behavior in the complex plane, which would be invisible just by looking at real points, can affect the behavior of a mathematical system and of physical processes modeled by such a system.

A Subtle Point We have been using implicitly the fact that the sum $f(t) = \sum_{n=0}^{\infty} a_n(t-t_0)^n$ of a power series is an analytic function within the interval of convergence $t_0 - R < t < t_0 + R$ of that series. If so, then the Taylor series of $f(t)$ at any *other* point t_1 in $(t_0 - R, t_0 + R)$ should also converge to $f(t)$. (Look carefully at the definition of ‘analytic’ at the beginning of this section!) In fact, such is the case, and the interval of convergence of the Taylor series at t_1 extends at least to the closer of the two endpoints $t_0 - R$, $t_0 + R$ of the interval of convergence of the original power series. One must prove these assertions to justify the conclusion that the sum of a power series is analytic, but we won’t do that in this course. It would



be rather difficult to do without getting into complex variable theory. Hence, you will have to wait for a course in that subject for a proof.

Exercises for 8.7.

- Find the Taylor Series $\sum_{n=0}^{\infty} \frac{f^{(n)}(t_0)}{n!} (t - t_0)^n$ for each of the following functions for the indicated value of t_0 . Do it by finding all the derivatives and evaluating them at t_0 . Having done that, see if you can also find the Taylor series by some simpler method.
 - $f(t) = e^{-t}$, $t_0 = 0$.
 - $f(t) = e^t$, $t_0 = 1$.
 - $f(t) = \frac{1}{t}$, $t_0 = 1$.
 - $f(t) = \cos t$, $t_0 = \frac{\pi}{2}$.
 - $f(t) = \ln t$, $t_0 = 1$.

- By differentiating the Taylor series expansion for $\sin t$ at $t_0 = 0$, check that you get the Taylor series expansion for $\cos t$.
- In complex variable courses, one studies power series where the independent variable, usually called z , is allowed to assume complex values. It is very much like the theory outlined in this section. With this in mind, verify the identity

$$e^{it} = \cos t + i \sin t$$

by calculating the series on both sides of the equation.

- How many terms of the Taylor series expansion of $\cos t$ at $t_0 = 0$ are needed to calculate $\cos(1)$ to within 5×10^{-16} ? Hint: The series is an alternating series
- How many terms of the Taylor series expansion of e^t at $t_0 = 0$ are necessary to calculate e to within 5×10^{-16} ? How about e^{10} to within 5×10^{-16} ? Hint: The series is *not* an alternating series.
- Use Taylor series to calculate $\sin(100)$ accurately to within 5×10^{-4} . Hint: Don't use the series at $t_0 = 0$.
- For each of the following functions, determine the radius of convergence of its Taylor series for $t_0 = 0$.
 - $f(t) = \frac{1+t}{t-2}$.
 - $f(t) = \frac{2t}{1+2t^2}$.

$$(c) f(t) = \frac{1}{1-t^3}.$$

$$(d) f(t) = \frac{1}{(t-4)(t^2+3)}.$$

It is not necessary to actually find the Taylor series.

8.8 More Calculations with Power Series

There are two approaches to finding the Taylor series expansion of an analytic function. You can start with the function and its derivatives and calculate the coefficients $a_n = f^{(n)}(t_0)/n!$. This is often quite messy. (Also, it may be quite difficult to show that the series converges to the function in an appropriate range by estimating $R_N(t)$.) Another approach is to start with the series for a related function and then derive the desired series by differentiating, integrating or other manipulations.

Example 169 By substituting $u = -t^2$ in the series expansion

$$\frac{1}{1-u} = 1 + u + u^2 + \dots$$

we may obtain the expansion

$$\frac{1}{1+t^2} = 1 - t^2 + t^4 - t^6 + \dots + (-1)^n t^{2n} + \dots$$

Since the first expansion is valid for $-1 < u < 1$, the second expansion is valid for $0 \leq t^2 < 1$, i.e., $-1 < t < 1$. In this range, we may integrate the series term by term to obtain

$$\int_0^t \frac{1}{1+s^2} ds = t - \frac{t^3}{3} + \frac{t^5}{5} - \dots + (-1)^n \frac{t^{2n+1}}{2n+1} + \dots$$

or after evaluating the left hand side

$$\tan^{-1} t = t - \frac{t^3}{3} + \frac{t^5}{5} - \dots + (-1)^n \frac{t^{2n+1}}{2n+1} + \dots$$

for $-1 < t < 1$. You might try to derive this Taylor series by taking derivatives of the $\tan^{-1} t$. You will find it is a lot harder.

The above series also converges for $t = 1$ by the alternating series test, and its sum is $\tan^{-1}(1)$. (This does not follow simply by the above analysis, since that only applied in the range $-1 < t < 1$, but it can be demonstrated by a more refined analysis.) Thus we get the formula

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots + (-1)^n \frac{1}{2n+1} + \dots$$

In principle, this series may be used to compute π to any desired degree of accuracy. Unfortunately, it converges rather slowly, but similar expansions for other rational multiples of π converge rapidly. Recently, such series have been used to test the capabilities of modern computers by computing π to exceptionally large numbers of decimal places.

The Taylor series does not converge for $t = -1$, since in that case you get

$$-1 - \frac{1}{3} - \frac{1}{5} - \frac{1}{7} - \cdots - \frac{1}{2n+1} - \cdots$$

(Apply the integral test to the negative of the series.) That fact should make you cautious about trying to extend results to the endpoints of the interval of convergence.

Example 170 The quantities $\sin t/t$ and $(1 - \cos t)/t$ occur in a variety of applications.

The limits $\lim_{t \rightarrow 0} \sin t/t = 1$ and $\lim_{t \rightarrow 0} (1 - \cos t)/t = 0$ must be determined in order to calculate the derivatives of the sin and cos functions. We can use the current theory to get more precise information about how the above ratios behave near $t = 0$. We discuss the case of $\sin t/t$. ($(1 - \cos t)/t$ is similar. Start with

$$\sin t = t - \frac{t^3}{3!} + \frac{t^5}{5!} - \cdots + (-1)^n \frac{t^{2n+1}}{(2n+1)!} + \cdots$$

which is valid for *all* t . Multiply by $1/t$ to obtain

$$\frac{\sin t}{t} = 1 - \frac{t^2}{3!} + \frac{t^4}{5!} - \cdots + (-1)^n \frac{t^{2n}}{(2n+1)!} + \cdots$$

which is also valid for all $t \neq 0$. (The analysis breaks down if $t = 0$, but the formula may be considered valid in the sense that the right hand side is 1 and the left hand side has limit 1.) Note that the series is an alternating series so we may estimate the error if we stop after N terms by looking at the *next* term. By taking $n = 1$, we conclude that

$$\frac{\sin t}{t} \approx 1$$

for t small with the difference behaving like $O(t^2)$.

Sometimes one can use Taylor series to evaluate an integral quickly and to high accuracy in a case where it is not possible to determine an elementary anti-derivative.

Example 171 We shall calculate $\int_0^{\pi/2} \frac{\sin x}{x} dx$ so that the error is less than 5×10^{-4} , which *loosely speaking* says is it accurate to three decimal places. Note that the function $\sin x/x$ does not have an elementary anti-derivative. As in the previous example,

$$\frac{\sin x}{x} = 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \cdots + (-1)^n \frac{x^{2n}}{(2n+1)!} + \cdots$$

and we may consider this expansion valid for all x by using the limit 1 for the value of $\sin x/x$ for $x = 0$. Integrating term by term yields

$$\int_0^{\pi/2} \frac{\sin x}{x} dx = \frac{\pi}{2} - \frac{1}{3 \cdot 3!} \left(\frac{\pi}{2}\right)^3 + \frac{1}{5 \cdot 5!} \left(\frac{\pi}{2}\right)^5 - \frac{1}{7 \cdot 7!} \left(\frac{\pi}{2}\right)^7 + \dots$$

Also, it is possible to show that the right hand side is an alternating series *in which the terms decrease*. Indeed, checking the first few values, we have (to 4 decimal places)

$$\begin{aligned} \frac{\pi}{2} &= 1.5708 \\ \frac{1}{18} \left(\frac{\pi}{2}\right)^3 &= 0.2153 \\ \frac{1}{600} \left(\frac{\pi}{2}\right)^5 &= 0.0159 \\ \frac{1}{35280} \left(\frac{\pi}{2}\right)^7 &= 0.0007 \\ &\vdots \end{aligned}$$

The next term is $(1/9 \cdot 9!)(\pi/2)^9 \approx 2 \times 10^{-5}$. Hence, the error in just using the terms listed is certainly small enough that it won't affect the 3rd decimal place. Adding up those terms (including signs) yields 1.371 to 4 decimal places.

The Binomial Theorem An important Taylor series is that for the function $f(t) = (1+t)^a$. That series is called the *binomial series*. Perhaps you recall from high school the formulas

$$\begin{aligned} (1+t)^2 &= 1 + 2t + t^2 \\ (1+t)^3 &= 1 + 3t + 3t^2 + t^3 \\ (1+t)^4 &= 1 + 4t + 6t^2 + 4t^3 + t^4 \\ &\vdots \end{aligned}$$

The general case for a a positive integer is the *binomial theorem* formula

$$(1+t)^a = 1 + at + \frac{a(a-1)}{2}t^2 + \dots + at^{a-1}t = \sum_{n=0}^a \binom{a}{n} t^n,$$

where

$$\binom{a}{n} = \frac{a(a-1)(a-2)\dots(a-n+1)}{n!} = \frac{a!}{n!(a-n)!}.$$

(By convention $\binom{a}{n} = 0$ if $n < 0$ or $n > a$.) If we take a to be a negative integer, a fraction or indeed *any non-zero real number*, then we may still define binomial coefficients

$$\binom{a}{n} = \frac{a(a-1)(a-2)\dots(a-n+1)}{n!}$$

with the convention that $\binom{a}{0} = 1$. Unlike the ordinary binomial coefficients, these quantities aren't necessarily positive, and moreover they don't vanish for $n > a$. Thus, instead of a polynomial, we get a *series expansion*

$$(1+t)^a = \sum_{n=0}^{\infty} \binom{a}{n} t^n \quad (133)$$

which is valid for $-1 < t < 1$.

Example 172 For $a = -2$, the coefficients are

$$\begin{aligned} \binom{-2}{0} &= 1 \\ \binom{-2}{1} &= -2 \\ \binom{-2}{2} &= \frac{(-2)(-3)}{2} = 3 \\ \binom{-2}{3} &= \frac{(-2)(-3)(-4)}{3!} = -4 \\ &\vdots \\ \binom{-2}{n} &= \frac{(-2)(-3)\dots(-n)(-n-1)}{n!} = (-1)^n(n+1) \end{aligned}$$

Hence,

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

(Compare this with the series for $(1+t)^{-2}$ you obtained in the Exercises for Section 6 by differentiating the series for $(1+t)^{-1}$.)

The general binomial theorem (133) was discovered and first proved by Isaac Newton. Since then there have been several different proofs. We shall give a rather tricky proof based on some of the ideas we developed above.

First note that the series $\sum_n \binom{a}{n} t^n$ has radius of convergence 1. For, the ratio

$$\begin{aligned} \frac{|\binom{a}{n+1}| |t|^{n+1}}{|\binom{a}{n}| |t|^n} &= \\ |t| \frac{|a(a-1)\dots(a-n+1)(a-(n+1)+1)|}{(n+1)!} \frac{n!}{|a(a-1)\dots(a-n+1)|} \\ &= |t| \frac{|a-n|}{n+1} \end{aligned}$$

approaches $|t|$ as $n \rightarrow \infty$. Thus, the series converges absolutely for $|t| < 1$. Let

$$f(t) = \sum_{n=0}^{\infty} \binom{a}{n} t^n \quad \text{for } -1 < t < 1.$$

Then

$$f'(t) = \sum_{n=1}^{\infty} \binom{a}{n} n t^{n-1}. \quad (134)$$

Multiply this by t to obtain

$$t f'(t) = \sum_{n=1}^{\infty} \binom{a}{n} n t^n = \sum_{n=0}^{\infty} \binom{a}{n} n t^n,$$

where we put back the $n = 0$ term which is zero in any case. Similarly, putting $n + 1$ for n in (134) yields

$$f'(t) = \sum_{n=0}^{\infty} \binom{a}{n+1} (n+1) t^{n-1},$$

so

$$f'(t) + t f'(t) = \sum_{n=0}^{\infty} \left(\binom{a}{n+1} (n+1) + \binom{a}{n} n \right) t^n.$$

However,

$$\begin{aligned} & \binom{a}{n+1} (n+1) + \binom{a}{n} n \\ &= \frac{a(a-1)\dots(a-(n+1)+1)}{(n+1)!} (n+1) + \frac{a(a-1)\dots(a-n+1)}{n!} n \\ &= \frac{a(a-1)\dots(a-n+1)}{n!} (a-n) + \frac{a(a-1)\dots(a-n+1)}{n!} n \\ &= \frac{a(a-1)\dots(a-n+1)}{n!} a = \binom{a}{n} a. \end{aligned}$$

Hence,

$$(1+t)f'(t) = \sum_{n=0}^{\infty} a \binom{a}{n} t^n = a \sum_{n=0}^{\infty} \binom{a}{n} t^n = a f(t).$$

In other words, $f(t)$ satisfies the differential equation

$$f'(t) - \frac{a}{1+t} f(t) = 0.$$

This is a linear equation, and we know the solution

$$f(t) = C e^{\int \frac{a}{1+t} dt} = C e^{a \ln(1+t)} = C(1+t)^a.$$

To determine C , note that it follows from the definition of f that $f(0) = 1$. Hence, $1 = C1^a = C$, and

$$f(t) = (1+t)^a \quad \text{for } -1 < t < 1$$

as claimed.

Other Manipulations Other manipulations with series are possible. For example, series may be added, subtracted, multiplied or even divided, but determining the radius of convergence of the resulting series may be somewhat tricky.

Example 173 We find the Taylor series expansion for $e^{2t} \cos t$ for $t_0 = 0$. We have

$$\begin{aligned} e^{2t} &= 1 + 2t + 2t^2 + \frac{4}{3}t^3 + \dots \\ \cos t &= 1 - \frac{1}{2}t^2 + \frac{1}{24}t^4 - \dots \end{aligned}$$

so

$$\begin{aligned} e^{2t} \cos t &= 1 + 2t + 2t^2 + \frac{4}{3}t^3 + \dots \\ &\quad - \frac{1}{2}t^2 - t^3 - \dots \end{aligned}$$

where we have listed only the terms of degree ≤ 3 . Combining terms, we obtain

$$e^{2t} \cos t = 1 + 2t + \frac{3}{2}t^2 + \frac{1}{3}t^3 + \dots$$

Example 174 We know that

$$\frac{1}{1+u} = 1 - u + u^2 - u^3 + \dots \quad \text{for } |u| < 1.$$

Since $(1+t)^2 = 1 + 2t + t^2$, we may substitute $u = 2t + t^2$ in the above equation to get

$$\begin{aligned} \frac{1}{(1+t)^2} &= \frac{1}{1+2t+t^2} = 1 - (2t+t^2) + (2t+t^2)^2 - (2t+t^2)^3 + \dots \\ &= 1 - 2t - t^2 + 4t^2 + 4t^3 + t^4 - 8t^3 - 12t^4 - 6t^5 - t^6 + \dots \\ &= 1 - 2t + 3t^2 - 4t^3 + \dots \end{aligned}$$

Note that I stopped including terms where I could be sure that the higher degree terms would not contribute further to the given power of t . (The term $(2t+t^2)^4$ would contribute a term involving t^4 .) Note also that it is not clear from the above computations what radius of convergence to specify for the last expansion to be valid.

You should compare the above expansion with what you would obtain by using the binomial theorem for $(1+t)^{-2}$.

One consequence of the fact that we may manipulate series in this way is that the sum, difference, or product of analytic functions is again analytic. Similarly, the quotient of two analytic functions is also analytic, at least if we exclude from the domain points where the denominator vanishes. Finally, it is even possible to

substitute one series in another, so the composition of two analytic functions is generally analytic.

Exercises for 8.8.

- Using the expansion

$$\tan^{-1}(t) = \sum_{n=0}^{\infty} \frac{(-1)^n t^{2n+1}}{2n+1} \quad \text{for} \quad -1 < t < 1$$

calculate $\tan^{-1}(.01)$ to within 5×10^{-4} . (Note that the series is an alternating series, so you can use the next term criterion to estimate the error after n terms.)

- Find Taylor series expansions by the methods of this section in each of the following cases.
 - e^{-t^2} at $t_0 = 0$.
 - e^t at $t_0 = 1$. Hint: $e^t = e^1 e^{t-1}$.
 - $(1+t)^{1/2}$ at $t_0 = 0$.
 - $(1-t)^{-3}$ at $t_0 = 0$.
 - $\frac{\ln(1+t)}{t}$ at $t_0 = 0$.
 - $\frac{\tan^{-1}(t)}{t}$ at $t_0 = 0$.
 - $t^2 e^{-t}$ at $t_0 = 0$.
 - $\frac{1+t}{1-t}$ at $t_0 = 0$.
- Calculate $\int_0^{0.1} e^{-t^2} dt$ accurately to within 5×10^{-5} .
- Calculate $\int_0^1 \frac{1 - \cos t}{t^2} dt$ accurately to within 5×10^{-4} .
- A very thin tunnel is drilled in a straight line through the Earth connecting two points $1/4$ mile apart (great circle distance). Assume the Earth is a perfect sphere with radius exactly 4000 miles. Find the maximum depth of the tunnel in feet accurately to 10 significant figures. Hint: The distance is $R(1 - \cos(\theta/2))$ where θ is the angle in radians at the center of the sphere subtended by the great circle arc.
- The quantity $\sqrt{1 - \frac{v^2}{c^2}}$ plays an important role in the special theory of relativity.

- (a) Using the first order term in the binomial expansion, derive the approximation

$$\sqrt{1 - \frac{v^2}{c^2}} \approx 1 - \frac{v^2}{2c^2}.$$

- (b) If v is 1 accurate the above approximation is.
 (c) Does a similar analysis work for

$$\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \approx 1 + \frac{v^2}{2c^2}?$$

7. (Optional) Try to obtain a power series expansion for $\sec t$ by putting $u = \frac{t^2}{2} - \frac{t^4}{4!} + \frac{t^6}{6!} - \dots$ in

$$\frac{1}{1-u} = 1 + u + u^2 + u^3 + \dots$$

Try to include at least terms up to degree 6.

8.9 Multidimensional Taylor Series

We want to extend the notions introduced in the previous sections to functions of more than one variable, i.e., $f(\mathbf{r})$ with \mathbf{r} in \mathbf{R}^n . In this section, we shall concentrate entirely on the case $n = 2$ because the notational difficulties get progressively worse as the number of variables increases. However, with a good understanding of that case, it is not hard to see how to proceed in higher dimensional cases.

Suppose then that $n = 2$, and a scalar valued function is given by $f(\mathbf{r}) = f(x, y)$. We want to expand this function in a multidimensional power series. From our previous discussion of linear approximation, we know it should start

$$f(x, y) = f(0, 0) + f_x(0, 0)x + f_y(0, 0)y + \dots,$$

and we need to figure out what the higher degree terms should be.

To this end, consider series of the form

$$a_{00} + a_{10}x + a_{01}y + a_{20}x^2 + a_{11}xy + a_{02}y^2 + \dots = \sum_{n=0}^{\infty} \left(\sum_{i+j=n} a_{ij} x^i y^j \right).$$

Notice the way the terms are arranged. A typical term will be some multiple of a *monomial* $x^i y^j$. The coefficient of that monomial is denoted a_{ij} where the subscripts

specify the powers of x and y in the monomial. Moreover, we call $n = i + j$ the *total degree* of the monomial $x^i y^j$, and we group all monomials of the same degree together as $\sum_{i+j=n} a_{ij} x^i y^j$.

The above series is centered at $(0, 0)$. We can similarly discuss a multidimensional power series of the form

$$\sum_{n=0}^{\infty} \left(\sum_{i+j=n} a_{ij} (x - x_0)^i (y - y_0)^j \right)$$

centered at the point (x_0, y_0) . For the moment we will concentrate on series centered at $(0, 0)$ in order to save writing. Once you understand that, you can get the general case by substituting $x - x_0$ and $y - y_0$ for x and y and making other appropriate changes.

A function $f(x, y)$ with domain some open set in \mathbf{R}^2 is said to be *analytic* if at each point in its domain it may be expanded in a power series centered at that point in some neighborhood of the point. Suppose then that

$$f(x, y) = a_{00} + a_{10}x + a_{01}y + a_{20}x^2 + a_{11}xy + a_{02}y^2 + \cdots = \sum_{n=0}^{\infty} \left(\sum_{i+j=n} a_{ij} x^i y^j \right)$$

is valid in some neighborhood of $(0, 0)$. By putting $x = 0, y = 0$ we see that

$$f(0, 0) = a_{00}.$$

Just as in the case of power series in one variable, term by term differentiation (or integration) is valid. Hence, we have

$$\frac{\partial f}{\partial x} = a_{10} + 2a_{20}x + a_{11}y + \cdots = \sum_{n=1}^{\infty} \left(\sum_{i+j=n} i a_{ij} x^{i-1} y^j \right). \quad (135)$$

Putting $x = 0, y = 0$ yields

$$\frac{\partial f}{\partial x}(0, 0) = a_{10},$$

just as we expected. Similarly, $\frac{\partial f}{\partial y}(0, 0) = a_{01}$.

Take yet another derivative to get

$$\frac{\partial^2 f}{\partial x^2} = 2a_{20} + \cdots = \sum_{n=2}^{\infty} \sum_{i+j=n} i(i-1) a_{ij} x^{i-2} y^j.$$

Hence,

$$\frac{\partial^2 f}{\partial x^2}(0, 0) = 2a_{20}.$$

Similar reasoning applies to partials with respect to y , and continuing in this way, we discover that

$$\begin{aligned} a_{n0} &= \frac{1}{n!} \frac{\partial^n f}{\partial x^n}(0, 0) \\ a_{0n} &= \frac{1}{n!} \frac{\partial^n f}{\partial y^n}(0, 0). \end{aligned}$$

However, this still leaves the bulk of the coefficients undetermined. To find these, we need to find some *mixed partials*. Thus, from equation (135), we get

$$\frac{\partial^2 f}{\partial x \partial y} = a_{11} + \cdots = \sum_{n=2}^{\infty} \left(\sum_{i+j=n} i j a_{ij} x^{i-1} y^{j-1} \right).$$

Thus,

$$\frac{\partial^2 f}{\partial x \partial y}(0, 0) = a_{11}.$$

Here the total degree is $1 + 1 = 2$. Suppose we consider the case of total degree $r + s = k$ where r and s are the subdegrees for x and y respectively. Then, it is possible to show that

$$\frac{\partial^k f}{\partial x^r \partial y^s} = \sum_{n=k}^{\infty} \sum_{i+j=n} i(i-1)(i-2)\cdots(i-r+1) j(j-1)(j-2)\cdots(j-s+1) a_{ij} x^{i-r} y^{j-s}.$$

The leading term in the expansion on the right for $n = k$ has only one non-zero term, the one with $i = r$ and $j = s$, i.e.,

$$r!s!a_{rs}.$$

Hence, since $k = r + s$,

$$a_{rs} = \frac{1}{r!s!} \frac{\partial^{r+s} f}{\partial x^r \partial y^s}(0, 0).$$

(You should do all the calculations for a_{21} and a_{12} to convince yourself that it really works this way. If you don't quite see why the calculations work as claimed in the general case, you should probably just accept them.)

Thus, the power series expansion centered at $(0, 0)$ is

$$f(x, y) = \sum_{n=0}^{\infty} \left(\sum_{i+j=n} \frac{1}{i!j!} \frac{\partial^n f}{\partial x^i \partial y^j}(0, 0) x^i y^j \right).$$

Similarly, the power series expansion centered at (x_0, y_0) is

$$f(x, y) = \sum_{n=0}^{\infty} \sum_{i+j=n} \frac{1}{i!j!} \frac{\partial^n f}{\partial x^i \partial y^j}(x_0, y_0) (x - x_0)^i (y - y_0)^j.$$

The right hand side is called the (multidimensional) *Taylor series* for the function centered at (x_0, y_0) .

There is another way to express the series. Consider the terms of total degree n

$$\sum_{i+j=n} \frac{1}{i!j!} \frac{\partial^n f}{\partial x^i \partial y^j} (x-x_0)^i (y-y_0)^j,$$

where to save writing we omit explicit mention of the fact that the partial derivatives are to be evaluated at (x_0, y_0) . To make this look a bit more like the case of one variable, we divide by a common factor of $n!$. Of course, we must then also multiply each term by $n!$ to compensate, and the extra $n!$ may be incorporated with the other factorials to obtain

$$\frac{n!}{i!j!} = \binom{n}{i}.$$

Hence, the terms of degree n may be expressed

$$\frac{1}{n!} \sum_{i+j=n} \binom{n}{i} \frac{\partial^n f}{\partial x^i \partial y^j} (x-x_0)^i (y-y_0)^j. \quad (136)$$

Example 175 Let $f(x, y) = e^{x+y}$ and $x_0 = 0, y_0 = 0$. Then it is not hard to check that

$$\frac{\partial^n f}{\partial x^i \partial y^j} = e^{x+y} = e^0 = 1 \quad \text{at } (0, 0).$$

Hence, the series expansion is

$$e^{x+y} = \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{i+j=n} \binom{n}{i} x^i y^j.$$

An easier way to derive this same series is to start with

$$e^u = \sum_{n=0}^{\infty} \frac{1}{n!} u^n$$

and put $u = x + y$. Since

$$(x+y)^n = \sum_{i+j=n} \binom{n}{i} x^i y^j,$$

we get the same answer.

There is yet another way to express formula (136). To save writing, put $\Delta x = x - x_0, \Delta y = y - y_0$. Consider the operator $D = \Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y}$. Then, symbolically,

$$\begin{aligned} D^n &= \left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y} \right)^n = \sum_{i+j=n} \binom{n}{i} \left(\Delta x \frac{\partial}{\partial x} \right)^i \left(\Delta y \frac{\partial}{\partial y} \right)^j \\ &= \sum_{i+j=n} \binom{n}{i} \Delta x^i \Delta y^j \frac{\partial^n}{\partial x^i \partial y^j}. \end{aligned}$$

Hence,

$$\sum_{n=0}^{\infty} \frac{1}{n!} (D^n f)(x_0, y_0) = \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{i+j=n} \binom{n}{i} \Delta x^i \Delta y^j \frac{\partial^n f}{\partial x^i \partial y^j}(x_0, y_0),$$

which is the expression arrived at earlier in formula (136).

The Error One can analyze the error R_N made if you stop with terms of degree N for multidimensional Taylor series in a manner similar to the case of ordinary Taylor series. The exact form of this remainder R_N is not as important as the fact that *as a function of* $|\Delta \mathbf{r}| = \sqrt{\Delta x^2 + \Delta y^2}$ it is $O(|\Delta \mathbf{r}|^{N+1})$. Thus, we may write in general

$$f(x + \Delta x, y + \Delta y) = \sum_{n=0}^N \frac{1}{n!} \sum_{i+j=n} \binom{n}{i} \frac{\partial^n f}{\partial x^i \partial y^j} \Delta x^i \Delta y^j + O(|\Delta \mathbf{r}|^{N+1}).$$

One important case is $N = 1$ which gives

$$f(x + \Delta x, y + \Delta y) = f(x, y) + f_x \Delta x + f_y \Delta y + O(|\Delta \mathbf{r}|^2).$$

(You should compare this with our previous discussion of the linear approximation in Chapter III.)

The case $N = 2$ is also worth writing out explicitly. It may be put in the following form

$$\begin{aligned} f(x + \Delta x, y + \Delta y) \\ = f(x, y) + f_x \Delta x + f_y \Delta y + \frac{1}{2} (f_{xx} \Delta x^2 + 2f_{xy} \Delta x \Delta y + f_{yy} \Delta y^2) + O(|\Delta \mathbf{r}|^3). \end{aligned}$$

We shall use this relation in the next section.

Derivation of the Remainder Term You may want to skip the following discussion.

There is a multidimensional analogue of Taylor's formula *with remainder*. The easiest way to get it is to derive it from the one dimensional case as follows. We concentrate, as above, on the case $n = 2$. Consider the function $f(\mathbf{r}) = f(x, y)$ near the point $\mathbf{r}_0 = (x_0, y_0)$ and put $\Delta \mathbf{r} = \langle \Delta x, \Delta y \rangle = \langle x - x_0, y - y_0 \rangle$. Consider the line segment from \mathbf{r}_0 to $\mathbf{r} = \mathbf{r}_0 + \Delta \mathbf{r}$ parameterized by

$$\mathbf{r}_0 + t\Delta \mathbf{r}, \quad \text{where } 0 \leq t \leq 1.$$

Define a function of one variable

$$g(t) = f(\mathbf{r}_0 + t\Delta \mathbf{r}), \quad 0 \leq t \leq 1.$$

By the 1-dimensional Taylor's formula, we have

$$g(t) = g(0) + g'(0)t + \frac{1}{2}g''(0)t^2 + \cdots + \frac{1}{N!}g^{(N)}(0)t^N + R_N(t)$$

where the remainder $R_N(t)$ is given by a certain integral. Putting $t = 1$, we obtain

$$g(1) = g(0) + g'(0) + \frac{1}{2}g''(0) + \cdots + \frac{1}{N!}g^{(N)}(0) + R_N(1)$$

where

$$R_N(1) = \frac{1}{N!} \int_0^1 (1-s)^N g^{(N+1)}(s) ds.$$

We need to express the above quantities in terms of the function f and its partial derivatives. This is not hard if we make use of the chain rule

$$\frac{dg}{dt} = \nabla f \cdot \frac{d}{dt}(\mathbf{r}_0 + t\Delta\mathbf{r}) = \nabla f \cdot \Delta\mathbf{r}.$$

We can write this as a symbolic relation between operators

$$\frac{d}{dt}g = \Delta\mathbf{r} \cdot \nabla f = (\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y})f = Df.$$

Hence,

$$\frac{d^n}{dt^n}g = D^n f$$

so evaluating at $t = 0, \mathbf{r} = \mathbf{r}_0$ yields

$$f(x, y) = \sum_{n=0}^N \frac{1}{n!} D^n f(x_0, y_0) + R_N(1).$$

To estimate the remainder, consider

$$\frac{d^{N+1}}{ds^{N+1}}g(s) = \sum_{i+j=N+1} \binom{N+1}{i} \Delta x^i \Delta y^j \frac{\partial^{N+1}f}{\partial x^i \partial y^j}(\mathbf{r}_0 + s\Delta\mathbf{r}).$$

Assume that within the indicated range

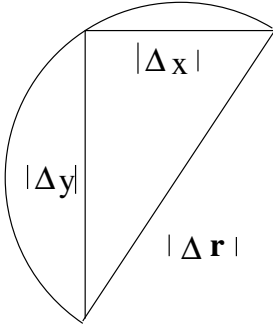
$$\left| \frac{\partial^{N+1}f}{\partial x^i \partial y^j} \right| \leq M.$$

Then

$$\left| \frac{d^{N+1}}{ds^{N+1}}g(s) \right| \leq \sum_{i+j=N+1} \binom{N+1}{i} |\Delta x|^i |\Delta y|^j M = M(|\Delta x| + |\Delta y|)^{N+1}.$$

The quantity $|\Delta x| + |\Delta y|$ represents the sum of the lengths of the legs of a right triangle with hypotenuse $|\Delta\mathbf{r}|$. A little plane geometry will convince you that $|\Delta x| + |\Delta y| \leq \sqrt{2}|\Delta\mathbf{r}|$. Hence,

$$\left| \frac{d^{N+1}}{ds^{N+1}}g(s) \right| \leq M(\sqrt{2}|\Delta\mathbf{r}|)^{N+1} = 2^{(N+1)/2} M |\Delta\mathbf{r}|^{N+1}.$$



Hence,

$$|R_N(1)| \leq \frac{1}{N!} 2^{(N+1)/2} M |\Delta \mathbf{r}|^{N+1} \int_0^1 (1-s)^N ds$$

or

$$|R_N(1)| \leq \frac{2^{(N+1)/2} M}{(N+1)!} |\Delta \mathbf{r}|^{N+1}. \quad (138)$$

A similar analysis works for functions $f : \mathbf{R}^n \rightarrow \mathbf{R}$, but the factor in the numerator will be more complicated for more than two variables.

Exercises for 8.9.

1. Find the Taylor series expansion for each of the following functions up to and including terms of degree 2.
 - (a) $f(x, y) = \sqrt{1 + x^2 + y^2}$ at $(x_0, y_0) = (0, 0)$.
 - (b) $f(x, y) = \frac{1+x}{1+y}$ at $(x_0, y_0) = (0, 0)$.
 - (c) $f(x, y) = x^2 + 3xy + y^3 + 2y^2 - 4y + 6$ at $(x_0, y_0) = (1, -2)$.
2. Find the Taylor series expansion for $f(x, y) = \sin(x + y)$ at $(0, 0)$
 - (a) by the general formula for the Taylor series of a function of two variables,
 - (b) by substituting $u = x + y$ in the Taylor series for $\sin u$.
3. Find the Taylor series expansion for $f(x, y) = e^{-(x^2+y^2)}$ at $(0, 0)$ up to and including terms of degree 3
 - (a) by the general formula for the Taylor series of a function of two variables,
 - (b) by substituting $u = -(x^2 + y^2)$ in the Taylor series for e^u ,
 - (c) by multiplying the Taylor series for e^{-x^2} and e^{-y^2} .

8.10 Local Behavior of Functions and Extremal Points

In your one variable calculus course, you learned how to find maximum and minimum points for graphs of functions. We want to generalize the methods you learned

there to functions of several variables. Among such problems, one usually distinguishes *global* problems from *local* problems. A *global* maximum (minimum) point is one at which the function takes on the largest (smallest) possible value for all points in its domain. A *local* maximum (minimum) point is one at which the function is larger (smaller) than its values at all nearby points. Thus, the bottom of the crater in the volcano Mauna Loa is a local minimum but certainly not a global minimum. Generally, one may use the methods of differential calculus to determine local maxima or minima, but usually other considerations must be brought into play to determine the global maximum or minimum. In this section, we shall concern ourselves only with local extremal points.

First, let's review the single variable case. Suppose $f : \mathbf{R} \rightarrow \mathbf{R}$ is a function, and we want to determine where it has local minima. If f is sufficiently smooth, we may begin to expand it near a point x in a Taylor series

$$f(x + \Delta x) = f(x) + f'(x)\Delta x + O(\Delta x^2)$$

or

$$f(x + \Delta x) - f(x) = f'(x)\Delta x + O(\Delta x^2).$$

The terms on the right represented by $O(\Delta x^2)$ are generally small compared to the linear term $f'(x)\Delta x$, so provided $f'(x) \neq 0$ and Δx is small, we may write

$$f(x + \Delta x) - f(x) \approx f'(x)\Delta x. \quad (139)$$

On the other hand, at a local minimum, we must have $f(x + \Delta x) - f(x) \geq 0$, and that contradicts (139) since the quantity on the right changes sign when Δx changes sign. The only way out of this dilemma is to conclude that $f'(x) = 0$ at a local minimum. Similar reasoning applies at a local maximum.

Suppose $f'(x) = 0$. Then, the approximation (139) is no longer valid, and we must consider higher order terms. Continuing the Taylor expansion yields

$$\begin{aligned} f(x + \Delta x) - f(x) &= f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2 + O(\Delta x^3) \\ &= \frac{1}{2}f''(x)\Delta x^2 + O(\Delta x^3). \end{aligned}$$

Reasoning as above, we conclude that if $f''(x) \neq 0$ and Δx is small, the quadratic term on the right dominates. That means that $f(x + \Delta x) \geq f(x)$ if $f''(x) > 0$, in which case we conclude that x is a local minimum point. Similarly, $f(x + \Delta x) \leq f(x)$ if $f''(x) < 0$, in which case we conclude that x is a local maximum point. If $f''(x) = 0$, then of course no conclusion is possible. As you know, in that case, x might be a local minimum point, a local maximum point, or a point at which the graph has a point of inflection. Taking the Taylor series one step further might yield additional information, but we will leave that for you to investigate on your own.

We now want to generalize the above analysis to functions $f : \mathbf{R}^2 \rightarrow \mathbf{R}$. If we assume that f is sufficiently smooth, then it may be expanded near a point (x, y)

$$f(x + \Delta x, y + \Delta y) = f(x, y) + \nabla f \cdot \Delta \mathbf{r} + O(|\Delta \mathbf{r}|^2)$$

which may be rewritten

$$f(x + \Delta x, y + \Delta y) - f(x, y) = \nabla f \cdot \Delta \mathbf{r} + O(|\Delta \mathbf{r}|^2).$$

As above, if $|\Delta \mathbf{r}|$ is small enough, and $\nabla f \neq \mathbf{0}$, then the linear term on the right dominates. Since in that case the linear term can be either positive or negative depending on $\Delta \mathbf{r}$, it follows that the quantity on the left cannot be of a single sign. Hence, if the point (x, y) is either a local maximum or a local minimum point, we must necessarily have $\nabla f = 0$. In words, *at a local extremal point of the function f , the gradient $\nabla f = 0$.*

A point (x, y) at which $\nabla f = \mathbf{0}$ is called a *critical point* of the function. For a smooth function, every local maximum or minimum is a critical point, but the converse is not generally true. (This parallels the single variable case.) All the assertion $\nabla f = 0$ tells us is that the tangent plane to the graph of the function is horizontal at the critical point.

Examples Let $f(x, y) = x^2 + 2x + y^2 - 6y$. Then

$$\nabla f = \langle f_x, f_y \rangle = \langle 2x + 2, 2y - 6 \rangle.$$

Setting this to zero yields $2x + 2 = 0, 2y - 6 = 0$ or $x = -1, y = 3$. Hence, $(-1, 3)$ is a critical point. We can see that this point is actually a local minimum point by completing the square.

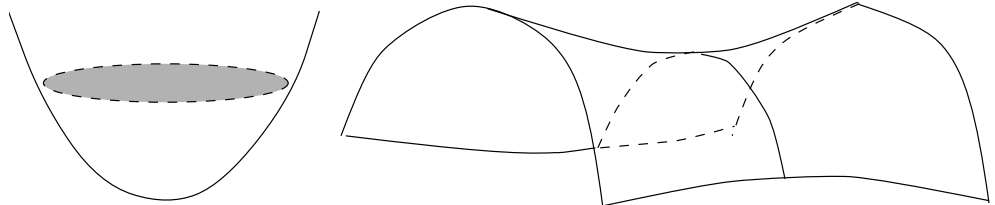
$$f(x, y) = x^2 + 2x + y^2 - 6y = x^2 + 2x + 1 + y^2 - 6y + 9 - 10 = (x + 1)^2 + (y - 3)^2 - 10.$$

The graph of this function is a circular paraboloid (bowl) with vertex $(-1, 3, -10)$, and it certainly has a minimum at $(-1, 3)$. (It is even a global minimum!)

Consider on the other hand, the function $f(x, y) = x^2 - y^2$. Setting $\nabla f = \mathbf{0}$ yields

$$\langle 2x, -2y \rangle = \langle 0, 0 \rangle$$

or $x = y = 0$. Hence, the origin is a critical point. However, we know the graph of the function is a hyperbolic paraboloid with a *saddle point* at $(0, 0)$. Near a saddle point, the function increases in some directions and decreases in others, so such a point is neither a local maximum nor a local minimum.



As in the single variable case, the analysis may be carried further by considering second derivatives and quadratic terms. Suppose $\nabla f = \mathbf{0}$, and extend the Taylor expansion to include the quadratic terms. We have

$$f(x + \Delta x, y + \Delta y) - f(x, y) = \frac{1}{2}(f_{xx}\Delta x^2 + 2f_{xy}\Delta x\Delta y + f_{yy}\Delta y^2) + O(|\Delta \mathbf{r}|^3).$$

In most circumstances, the quadratic terms will dominate, so we will be able to determine the local behavior by examining them. However, because the expression is so much more complicated than in the single variable case, the theory is more complicated. To save writing, put $A = f_{xx}$, $B = f_{xy}$, $C = f_{yy}$, $u = \Delta x$, and $v = \Delta y$. Then we want to look at the graph of

$$z = Q(u, v) = Au^2 + 2Buv + Cv^2$$

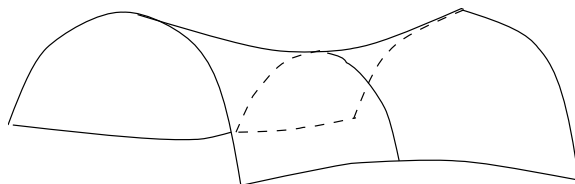
in the neighborhood of the origin. Such an expression is called a *quadratic form*.

Before considering the general case we consider some examples.

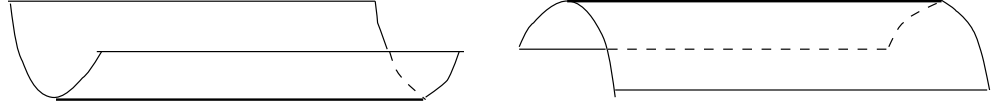
(a) $z = 2u^2 + v^2$. The graph is a bowl shaped surface which opens upward, so the origin is a local minimum. Similarly, the graph of $z = -u^2 - 5v^2$ is a bowl shaped surface opening downward, and the origin is a local maximum.



(b) $z = uv$. The graph is a saddle shaped surface, and the origin is neither a local minimum nor a local maximum.



(c) $z = u^2$. The graph is a 'trough' centered on the u -axis and opening upward. The entire v -axis is a line of local minimum points. Similarly, the graph of $z = -v^2$ is a trough centered on the u -axis and opening downward.



We now consider the general case. Suppose first that $A \neq 0$. Multiply through by A and then complete the square

$$\begin{aligned} AQ(u, v) &= A^2u^2 + 2ABuv + B^2v^2 - B^2v^2 + ACv^2 \\ &= (Au + Bv)^2 + (AC - B^2)v^2. \end{aligned}$$

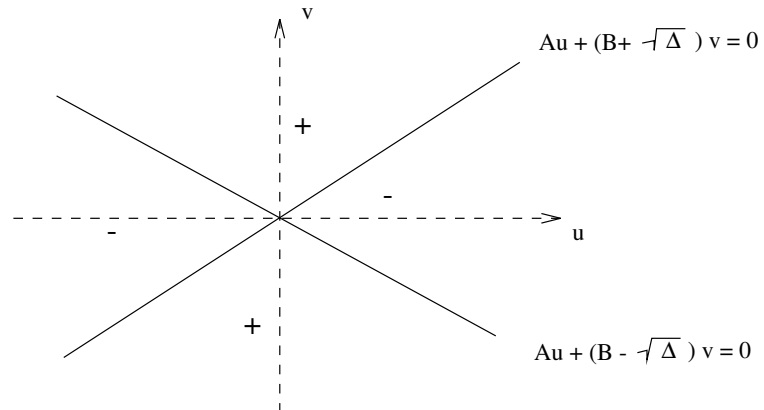
The first term is a square, so it is always non-negative, so the signed behavior of the expression depends on the quantity $\Delta = AC - B^2$. Δ is called the *discriminant of the quadratic form*.

If $\Delta > 0$, the expression is a sum of squares so it is always non-negative. Even better, it can't be zero unless $u = v = 0$. For, if the sum is zero, it follows $Au + Bv = 0$ and also $v = 0$. Since we assumed $A \neq 0$, that means that $u = 0$. Hence, if $A \neq 0$ and $\Delta > 0$, then $AQ(u, v)$ is always positive (except for $u = v = 0$), and $Q(u, v)$ has the *same sign as A*. Hence, the origin is a local minimum if $A > 0$, and it is a local maximum if $A < 0$. In this case the form is called *definite*.

If $\Delta < 0$, the expression is a difference of squares, so it is sometimes positive and sometimes negative. The graph of $z = AQ(u, v) = (Au + Bv)^2 - |\Delta|v^2$ is a saddle shaped surface which intersects the u, v -plane in the locus

$$Au + Bv = \pm\sqrt{|\Delta|}v,$$

which is a pair of lines intersecting at the origin. These lines divide the u, v -plane into four regions with the surface above it in two of the regions and below it in the other two. In this case, the form is called *indefinite*.



If $\Delta = 0$,

$$z = AQ(u, v) = (Au + Bv)^2$$

defines a trough shaped surface which intersects the u, v -plane along the line $Au + Bv = 0$. The graph of $z = Q(u, v)$ either lies above or below the u, v -plane, depending on the sign of A . If $\Delta = AC - B^2 = 0$, we say that the quadratic form is *degenerate*.

If $A = 0$ and $B \neq 0$, then $\Delta = AC - B^2 = -B^2 < 0$, and

$$Q(u, v) = 2Buv + Cv^2 = (2Bu + Cv)v$$

can be positive or negative. This just amounts to a special case of the analysis above for an indefinite form.

If $A = B = 0$ and $C \neq 0$, then $\Delta = 0$, and

$$Q(u, v) = Cv^2$$

and this is a special case of a degenerate form.

If $A = B = C = 0$, there isn't really anything to consider since we don't really have a quadratic form.

The above analysis tells us how the quadratic terms in

$$f(x + \Delta x, y + \Delta y) - f(x, y) = \frac{1}{2}(f_{xx}\Delta x^2 + 2f_{xy}\Delta x\Delta y + f_{yy}\Delta y^2) + O(|\Delta \mathbf{r}|^3)$$

behave. If $|\Delta \mathbf{r}|$ is small, the behavior on the right is determined primarily by $Q(\Delta x, \Delta y)$, with the cubic and higher order providing a slight distortion. In particular we have

Sufficiency Conditions for Local Extrema Assume $f : \mathbf{R}^2 \rightarrow \mathbf{R}$ has continuous second partials, and that (x, y) is a critical point. Let

$$\Delta = f_{xx}f_{yy} - f_{xy}^2$$

where everything is evaluated at the critical point.

(a) If $\Delta > 0$, then the critical point is a local minimum if $f_{xx} > 0$, or a local maximum if $f_{xx} < 0$. (The quadratic approximation is an elliptic paraboloid.)

(b) If $\Delta < 0$, then the critical point is neither a local maximum nor a local minimum. (The quadratic approximation is a saddle.)

(c) If $\Delta = 0$, any of the above possibilities might occur.

The proofs of these assertions require a careful analysis in which the different order terms are compared to one another. We shall not go into these matters here in detail,

but a few remarks might be enlightening. The reason why case (c) is ambiguous is not hard to see. If $\Delta = 0$, the graph of the degenerate quadratic form is a trough which contains a line of minimum (or maximum) points. *Along that line*, the cubic order terms will control what happens. We could have either a local maximum or a local minimum or neither depending on the shape of the trough and the contribution from the cubic terms. In case $\Delta < 0$, it is harder to see how the cubic terms perturb the saddle shaped graph for the quadratic form, but it may still be thought of as a saddle.

There is one slightly confusing point in the use of the above formulas. It appears that f_{xx} is playing a special role in (a). However, it is in fact true in case (a) that f_{xx} and f_{yy} have the same sign. Otherwise, $\Delta = f_{xx}f_{yy} - f_{xy}^2 < 0$.

Example 176 We shall classify the critical points of the function defined by $f(x, y) = x \sin y$. First, to find the critical points, solve

$$\begin{aligned}\frac{\partial f}{\partial x} &= \sin y = 0 \\ \frac{\partial f}{\partial y} &= x \cos y = 0.\end{aligned}$$

From the first equation, we get $y = k\pi$ where k is any integer. Since $\cos(k\pi) \neq 0$, the second equation yields $x = 0$. Hence, there are infinitely many critical points $(0, k\pi)$ where k ranges over all possible integers.

To apply the criteria above, we need to calculate the discriminant. We have

$$\begin{aligned}f_{xx} &= 0 \\ f_{yy} &= -x \sin y \\ f_{xy} &= \cos y \\ \Delta &= 0 - \cos^2 y = -\cos^2 y.\end{aligned}$$

At $(0, k\pi)$, we have $\Delta = -(\pm 1)^2 = -1 < 0$. Hence, (b) applies and every critical point is a saddle point.

Example 177 Let

$$f(x, y) = x^3 + y^3 - 3x^2 + 3y^2 + 2.$$

To find the critical points, solve

$$\begin{aligned}f_x &= 3x^2 - 6x = 0 \\ f_y &= 3y^2 + 6y = 0.\end{aligned}$$

The solutions are $x = 0, 2$ and $y = 0, -2$, so the critical points are

$$(0, 0), (0, -2), (2, 0), (2, -2).$$

To classify these, calculate

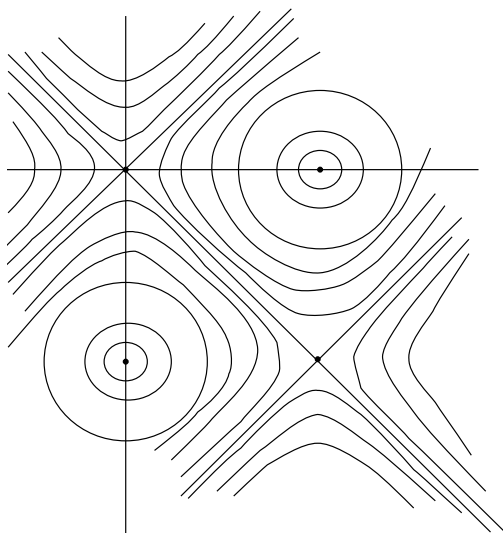
$$\begin{aligned}f_{xx} &= 6x - 6 \\f_{yy} &= 6y + 6 \\f_{xy} &= 0 \\ \Delta &= 36(x-1)(y+1).\end{aligned}$$

At $(0, 0)$, $\Delta = -36 < 0$, so $(0, 0)$ is a saddle point.

At $(0, -2)$, $\Delta = 36 > 0$. Since, $f_{xx} = -6 < 0$, $(0, -2)$ is a local maximum.

At $(2, 0)$, $\Delta = 36 > 0$. Since, $f_{xx} = 6 > 0$, $(2, 0)$ is a local minimum.

At $(2, -2)$, $\Delta = -36 < 0$, so $(2, -2)$ is a saddle point.



Example 178 Let

$$f(x, y) = x^3 + y^2.$$

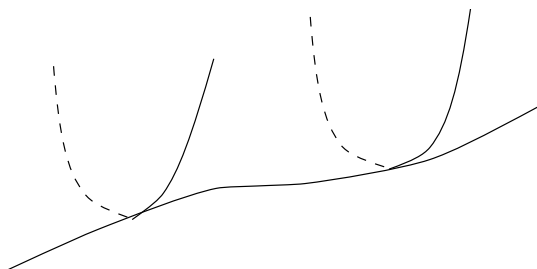
We have

$$f_x = 3x^2 \quad f_y = 2y,$$

so $(0, 0)$ is the only critical point. Moreover

$$f_{xx} = 6x, \quad f_{yy} = 2, \quad \text{and } f_{xy} = 0.$$

Hence, $\Delta = 12x = 0$ at $(0, 0)$. Hence, the criteria yield no information in this case. In fact, the point is not a maximum, a minimum, or a saddle point.



The example $f(x, y) = x^4 + y^2$ is very similar. It also has a degenerate critical point at $(0, 0)$ but in this case it is a minimum point.

If the quadratic terms vanish altogether, then the cubic terms dominate and various interesting possibilities arise. One of these is called a ‘monkey saddle’, and you can imagine its shape from its name. (See the Exercises.)

Higher Dimensional Cases For a *smooth* function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ with $n > 2$, many of the above considerations still apply. If $\nabla f = 0$ at a point \mathbf{r} , then the point is called a *critical point*. Local maxima and minima occur at critical points, but there are many other possibilities. Examination of the quadratic terms may allow one to determine precisely what happens, but even in the case $n = 3$ this can be much more complicated than the case of functions of two variables.

Exercises for 8.10.

1. Find all the critical points of each of the following functions
 - (a) $f(x, y) = x^2 + y^2 - 6x + 4y - 3$.
 - (b) $f(x, y) = x^2 - 2x + y^3$.
 - (c) $f(x, y) = e^{x^2+2x-y^2+6y}$.
 - (d) $f(x, y) = \sin x \cos y$.
 - (e) $f(x, y, z) = x^2 + 2xy + y^2 + z^2 - 4z$.
2. For each of the following functions find and classify each of its critical points. (If the theory described in the section gives no information, don’t try to analyze the critical point further.)
 - (a) $f(x, y) = x^2 + y^2 - 6x + 4y - 3$.
 - (b) $f(x, y) = 1 - 2xy - 2x - 2y - 3y^2$.
 - (c) $f(x, y) = x^3 + y^3 - 3xy$.
 - (d) $f(x, y) = 10 - 3x^2 - 3y^2 + 2xy - 4x - 4y$.
 - (e) $f(x, y) = x^2 + y^2 + 2xy + y^3$.

3. Sketch the graph of $f(x, y) = x^4 + y^2$. Pay particular attention to the neighborhood of its one critical point at the origin. Is that critical point a local maximum or minimum? Explain.
4. The origin is the only critical point of $f(x, y) = x^2 + y^5$. Is it a local maximum or minimum? Explain.
5. (a) Show that $(0, 0)$ is the only critical point of $f(x, y) = x^3 - xy^2$.
(b) Show that the discriminant $\Delta = 0$.
(c) By writing $f(x, y) = x(x - y)(x + y)$, show that the plane is divided into 6 wedge shaped regions by the lines $x = 0$, $x = y$, and $x = -y$. Examine the sign of f in each of these regions. The graph of this function is often called a 'monkey saddle'.
6. Find and classify the critical points of $f(x, y) = x^4 - 2x^2 + y^4 - 8y^2$.
7. Consider the problem of minimizing the square of the distance from the point $(0, 0, 1)$ to the point (x, y, z) on the hyperbolic paraboloid $z = x^2 - y^2$. See what you can conclude about that problem by the methods of this section.

Chapter 9

Series Solution of Differential Equations

9.1 Power Series Solutions at Ordinary Points

We want to solve equations of the form

$$y'' + p(t)y' + q(t)y = 0 \quad (140)$$

where $p(t)$ and $q(t)$ are analytic functions on some interval in \mathbf{R} . We saw several examples of such equations in the previous chapter, and we shall see others. They arise commonly in solving physical problems.

The functions $p(t)$ and $q(t)$ are supposed to be analytic on their domains, but they often will have singularities at other points. We assume that these singularities occur at *isolated* points of \mathbf{R} . Points where the functions are analytic are called *ordinary* points and other points are called *singular* points.

Example 179 The coefficients of Legendre's equation

$$y'' - \frac{2t}{1-t^2}y' + \frac{\alpha(\alpha+1)}{1-t^2}y = 0$$

are analytic on any interval not containing the points $t = \pm 1$. ± 1 are singular points.

Let t_0 be an ordinary point of (140). It makes sense to look for a solution $y = y(t)$ which is analytic in an interval containing t_0 . Such a solution will have a power series centered at t_0

$$y = \sum_{n=0}^{\infty} a_n(t-t_0)^n \quad (141)$$

with a positive radius of convergence R . Hence, one way to try to solve the equation is to substitute the series (141) in the equation (140) and see if we can determine the coefficients a_n . Generally, these coefficients may be expressed in terms of the first two, $a_0 = y(t_0)$ and $a_1 = y'(t_0)$. (See Chapter VIII, Section 1.)

We illustrate the method by an example.

Example 179, continued Consider

$$y'' - \frac{2t}{1-t^2}y' + \frac{\alpha(\alpha+1)}{1-t^2}y = 0$$

in the neighborhood of the point $t_0 = 0$. Before trying a series solution, it is better in this case to multiply through by the factor $1-t^2$ so as to avoid fractions. (This is not absolutely necessary, but in most cases it simplifies the calculations.) This yields

$$(1-t^2)y'' - 2ty' + \alpha(\alpha+1)y = 0. \quad (142)$$

We try a power series of the form $y = \sum_{n=0}^{\infty} a_n t^n$ and calculate

$$\begin{aligned} y &= \sum_{n=0}^{\infty} a_n t^n \\ y' &= \sum_{n=1}^{\infty} n a_n t^{n-1} \\ y'' &= \sum_{n=2}^{\infty} n(n-1) a_n t^{n-2}. \end{aligned}$$

We now write down the terms in the differential equation and *renumber the indices* so that we may collect coefficients of common powers of t .

$$\begin{aligned} y'' &= \sum_{n=2}^{\infty} n(n-1) a_n t^{n-2} &= \sum_{n+2=2}^{\infty} (n+2)(n+2-1) a_{n+2} t^{n+2-2} \\ & &\text{replacing } n \text{ by } n+2 \\ & &= \sum_{n=0}^{\infty} (n+2)(n+1) a_{n+2} t^n \\ -t^2 y'' &= \sum_{n=2}^{\infty} (-1)n(n-1) a_n t^{2+n-2} &= \sum_{n=0}^{\infty} (-1)n(n-1) a_n t^n \\ & &\text{terms for } n=0, 1 \text{ are } 0 \\ -2ty' &= \sum_{n=1}^{\infty} (-2)n a_n t^{1+n-1} &= \sum_{n=0}^{\infty} (-2)n a_n t^n \\ & &\text{term for } n=0 \text{ is } 0 \\ \alpha(\alpha+1)y &= \sum_{n=0}^{\infty} \alpha(\alpha+1) a_n t^n. \end{aligned}$$

(Make sure you understand each step!) Now add everything up. On the left side, we get zero, and on the right side we collect terms involving the same power of t .

$$0 = \sum_{n=0}^{\infty} [(n+2)(n+1)a_{n+2} - n(n-1)a_n - 2na_n + \alpha(\alpha+1)a_n]t^n.$$

A power series in t is zero if and only if the coefficient of each power of t is zero, so we obtain

$$\begin{aligned} (n+2)(n+1)a_{n+2} - n(n-1)a_n - 2na_n + \alpha(\alpha+1)a_n &= 0 \quad \text{for } n \geq 0 \\ (n+2)(n+1)a_{n+2} &= [n(n-1) + 2n - \alpha(\alpha+1)]a_n \\ (n+2)(n+1)a_{n+2} &= [n^2 - n + 2n - \alpha(\alpha+1)]a_n = [n^2 + n - \alpha(\alpha+1)]a_n \\ (n+2)(n+1)a_{n+2} &= (n+\alpha+1)(n-\alpha)a_n \\ a_{n+2} &= \frac{(n+\alpha+1)(n-\alpha)}{(n+2)(n+1)}a_n \quad \text{for } n \geq 0. \end{aligned}$$

The last equation is an example of what is called a *recurrence relation*. Once a_0 and a_1 are known, it is possible iteratively to determine any a_n with $n \geq 2$.

It is better, of course, to find a general *formula* for a_n , but this is not always possible. In the present example, it is possible to find such a formula, but it is very complicated. (See *Differential Equations* by G. F. Simmons, Section 27.) We derive the formula for two particular values of α . First take $\alpha = 1$. The recurrence relation is

$$a_{n+2} = \frac{(n+2)(n-1)}{(n+2)(n+1)}a_n = \frac{n-1}{n+1}a_n \quad \text{for } n \geq 0.$$

Thus, for even n ,

$$\begin{aligned} n=0 & \quad a_2 = -a_0 \\ n=2 & \quad a_4 = \frac{1}{3}a_2 = -\frac{1}{3}a_0 \\ n=4 & \quad a_6 = \frac{3}{5}a_4 = \frac{3}{5}\left(-\frac{1}{3}\right)a_0 = -\frac{1}{5}a_0 \\ n=6 & \quad a_8 = \frac{5}{7}a_6 = \frac{5}{7}\left(-\frac{1}{5}\right)a_0 = -\frac{1}{7}a_0. \end{aligned}$$

The general rule is now clear

$$a_n = -\frac{1}{n-1}a_0 \quad \text{for } n = 2, 4, 6, \dots$$

This could also be written

$$a_{2k} = -\frac{1}{2k-1}a_0 \quad \text{for } k = 1, 2, 3, \dots$$

For n odd, we get for $n = 1$,

$$a_3 = \frac{0}{2}a_1 = 0,$$

so $a_n = 0$ for every odd $n \geq 3$.

We may now write out the general solution

$$\begin{aligned} y &= \sum_{n=0}^{\infty} a_n t^n \\ &= a_0 + \underbrace{\sum_{k=1}^{\infty} \left(-\frac{1}{2k-1} \right) a_0 t^{2k}}_{\text{even } n} + a_1 t \\ &= a_0 \left(1 - \sum_{k=1}^{\infty} \frac{t^{2k}}{2k-1} \right) + a_1 t. \end{aligned}$$

Define $y_1 = 1 - \sum_{k=1}^{\infty} \frac{t^{2k}}{2k-1}$ and $y_2 = t$. Then, we have shown that any solution may be written

$$y = a_0 y_1 + a_1 y_2. \quad (143)$$

Moreover it is clear that y_1 and y_2 form a linearly independent pair. (Even if it weren't obvious by inspection, we could always verify it by calculating the Wronskian at $t = 0$.)

Hence, (143) gives a general solution of Legendre's differential equation for $\alpha = 1$.

(Since $y_2 = t$ is a solution, you could use the method of reduction of order to find a second independent solution. This yields

$$y = 1 - \frac{t}{2} \ln \left(\frac{1+t}{1-t} \right) = 1 - \frac{t}{2} (\ln(1+t) - \ln(1-t)).$$

See if you can expand this out and get y_1 .)

Let's see what happens for $\alpha = -\frac{1}{2}$. The recurrence relation may be rewritten

$$\begin{aligned} a_{n+2} &= \frac{(n+1/2)^2}{(n+2)(n+1)} a_n \\ &= \frac{(2n+1)^2}{4(n+2)(n+1)} a_n \quad \text{for } n \geq 0. \end{aligned}$$

(Both numerator and denominator were multiplied by 4.) Thus, for n even, we have

$$\begin{array}{ll} n = 0 & a_2 = \frac{1^2}{4 \cdot 2} a_0 \\ n = 2 & a_4 = \frac{5^2}{4 \cdot 4 \cdot 3} a_2 = \frac{(5 \cdot 1)^2}{4^2 \cdot 4!} a_0 \\ n = 4 & a_6 = \frac{9^2}{4 \cdot 6 \cdot 5} a_4 = \frac{(9 \cdot 5 \cdot 1)^2}{4^3 \cdot 6!} a_0 \\ n = 6 & a_8 = \frac{13^2}{4 \cdot 8 \cdot 7} a_6 = \frac{(13 \cdot 9 \cdot 5 \cdot 1)^2}{4^4 \cdot 8!} a_0. \end{array}$$

Note that $4^3 6! = 2^6 6!$ and $4^4 8! = 2^8 8!$. We begin see the following rule for n even,

$$a_n = \frac{[(2n-3)(2n-7)\dots 5\cdot 1]^2}{2^n n!} a_0 \quad \text{for } n = 2, 4, 6, \dots$$

In the numerator, we start with $2n-3$, reduce successively by 4 until we get down to 1, multiply all those numbers together, and square the whole thing.

For n odd, the same analysis yields a similar result.

$$a_n = \frac{[(2n-3)(2n-7)\dots 3]^2}{2^{n-1} n!} a_1 \quad \text{for } n = 3, 5, 7, \dots$$

As above, we may define

$$y_1 = 1 + \sum_{\substack{n \text{ even} \\ n > 0}} \frac{[(2n-3)(2n-7)\dots 5\cdot 1]^2}{2^n n!} t^n$$

$$y_2 = t + \sum_{\substack{n \text{ odd} \\ n > 1}} \frac{[(2n-3)(2n-7)\dots 3]^2}{2^{n-1} n!} t^n$$

and

$$y = \sum_{n=0}^{\infty} a_n t^n = a_0 y_1(t) + a_1 y_2(t).$$

Neither y_1 nor y_2 is a polynomial. For any power series in t , the constant term is the value of the sum at $t = 0$ and the coefficient of t is the value of its derivative at $t = 0$. Hence,

$$\begin{array}{ll} y_1(0) = 1 & y_1'(0) = 0 \\ y_2(0) = 0 & y_2'(0) = 1. \end{array}$$

(Why?) Hence, the Wronskian $W(0) = 1 \cdot 1 - 0 \cdot 0 = 1 \neq 0$, and it follows that the two solutions form a linearly independent pair. (Note also, that it is fairly clear that neither is a constant multiple of the other since y_1 starts with 1 and y_2 starts with t .)

Example 180 Consider the equation $y'' - 2ty' + 2y = 0$. The coefficients $p(t) = -2t$ and $q(t) = 2$ have no singularities, so every point is an ordinary point. To make the problem a trifle more interesting, we find a series solution centered at $t_0 = 1$, i.e., we try a series of the form $y = \sum_{n=0}^{\infty} a_n (t-1)^n$. To simplify the algebra, we introduce a new variable $s = t - 1$. Then, since $t = s + 1$, the equation may be rewritten

$$y'' - 2(s+1)y' + 2y = 0.$$

Subtle point: $y' = \frac{dy}{dt} = \frac{dy}{ds}$ and similarly for y'' . (Why?) Hence, we are not begging the question in the above equation by treating s as the independent variable!

We proceed exactly as before, but we shall skip some steps.

$$\begin{aligned}
y'' &= \sum_{n=2}^{\infty} n(n-1)a_n s^{n-2} &= \sum_{n=0}^{\infty} (n+2)(n+1)a_{n+2} s^n \\
-2sy' &= \sum_{n=1}^{\infty} (-2na_n)s^{1+n-1} &= \sum_{n=0}^{\infty} (-2na_n)s^n \\
-2y' &= \sum_{n=1}^{\infty} (-2na_n)s^{n-1} &= \sum_{n=0}^{\infty} (-2(n+1)a_{n+1})s^n \\
2y &= &= \sum_{n=0}^{\infty} (2a_n)s^n.
\end{aligned}$$

Adding up and comparing corresponding coefficients of s^n yields

$$\begin{aligned}
0 &= (n+2)(n+1)a_{n+2} - 2na_n - 2(n+1)a_{n+1} + 2a_n \\
(n+2)(n+1)a_{n+2} &= 2(n+1)a_{n+1} + (2n-2)a_n \\
a_{n+2} &= 2 \frac{(n+1)a_{n+1} + (n-1)a_n}{(n+2)(n+1)} \quad n \geq 0.
\end{aligned}$$

In this case, we cannot separate the terms into even and odd terms, and the general term is not so easy to determine. Here are some of the terms

$$\begin{aligned}
n=0 & \quad a_2 = 2 \frac{1}{2}(a_1 - a_0) = a_1 - a_0 \\
n=1 & \quad a_3 = 2 \frac{2a_2 + 0 \cdot a_1}{3 \cdot 2} = \frac{2}{3}(a_1 - a_0) \\
n=2 & \quad a_4 = 2 \frac{3a_3 + 1 \cdot a_2}{4 \cdot 3} = \frac{1}{6}(2(a_1 - a_0) + (a_1 - a_0)) \\
& \quad = \frac{1}{2}(a_1 - a_0) \\
n=3 & \quad a_5 = 2 \frac{4a_4 + 2a_3}{5 \cdot 4} = \dots = \frac{1}{3}(a_1 - a_0) \\
& \quad \vdots
\end{aligned}$$

I gave up trying to find the general terms. The general solution is

$$\begin{aligned}
y &= \sum_{n=0}^{\infty} a_n s^n \\
&= a_0 + a_1 s + (a_1 - a_0)s^2 + \frac{2}{3}(a_1 - a_0)s^3 + \frac{1}{2}(a_1 - a_0)s^4 + \frac{1}{3}(a_1 - a_0)s^5 + \dots \\
&= a_0 \underbrace{\left(1 - s^2 - \frac{2}{3}s^3 - \frac{1}{2}s^4 - \frac{1}{3}s^5 - \dots\right)}_{y_1} \\
&\quad + a_1 \underbrace{\left(s + s^2 + \frac{2}{3}s^3 + \frac{1}{2}s^4 + \frac{1}{3}s^5 + \dots\right)}_{y_1}.
\end{aligned}$$

Hence, if we put back $s = t - 1$, we may write

$$y = a_0 y_1(t) + a_1 y_2(t)$$

where

$$\begin{aligned} y_1(t) &= 1 - (t-1)^2 - \frac{2}{3}(t-1)^3 - \frac{1}{2}(t-1)^4 - \frac{1}{3}(t-1)^5 - \dots \\ y_2(t) &= (t-1) + (t-1)^2 + \frac{2}{3}(t-1)^3 + \frac{1}{2}(t-1)^4 + \frac{1}{3}(t-1)^5 + \dots \end{aligned}$$

The radius of convergence of a series solution An extension of the *basic existence and uniqueness theorem*—which we won't try to prove in this course—tells us that a solution of $y'' + p(t)y' + q(t)y = 0$ is analytic *at least* where the coefficients $p(t)$ and $q(t)$ are analytic. (The solution could be analytic on an even larger domain.) We may use this fact to study the radius of convergence of a series solution of a linear differential equation. Namely, in Chapter VIII, Section 8,

we described a rule for determining the radius of convergence of the Taylor series of a function: calculate the distance to the nearest singularity. Unfortunately, there was an important caveat to keep in mind. You have to look also at singularities in the complex plane.

Example 179, revisited Legendre's equation

$$y'' - \frac{2t}{1-t^2}y' + \frac{\alpha(\alpha+1)}{1-t^2}y = 0$$

has singularities when $1 - t^2 = 0$, i.e., at $t = \pm 1$. The distance from $t_0 = 0$ to the nearest singularity is 1. Hence, the radius of convergence of a power series solution centered at $t_0 = 0$ is at least 1, but it may be larger. Consider in particular the case $\alpha = 1$. One of the solutions $y_2(t) = t$ is a polynomial and so it converges for all t . Its radius of convergence is infinite. The other solution is

$$y_1(t) = 1 - \sum_{k=1}^{\infty} \frac{t^{2k}}{2k-1}$$

and it is easy to check by the ratio test that its radius of convergence is precisely 1.

Example 181 Consider

$$y'' + \frac{3t}{1+t^2}y' + \frac{1}{1+t^2}y = 0.$$

The coefficients have singularities where $1 + t^2 = 0$, i.e., at $t = \pm i$. The distance from $t_0 = 0$ to $\pm i$ in the complex plane is 1. Hence, series solutions of this equation centered at $t_0 = 0$ will have radius of convergence at least 1.

(See Braun, Section 2.8, Example 2 for solutions of this equation.)

More complicated coefficients In all the above examples, the coefficients were quite simple. They were what we call *rational functions*, i.e., quotients of polynomials. For such functions, it is possible to multiply through by a common denominator and thereby deal only with polynomial coefficients. Of course, in general that may not be possible. For example, for the equation

$$y'' + (\sin t)y' + (\cos t)y = 0$$

the simplification employed previously won't work. However, the method still works. Namely, we may put the expansions

$$\begin{aligned}\sin t &= \sum_{n=0}^{\infty} (-1)^n \frac{t^{2n+1}}{(2n+1)!} \\ \cos t &= \sum_{n=0}^{\infty} (-1)^n \frac{t^{2n}}{(2n)!}\end{aligned}$$

along with the expansions

$$\begin{aligned}y &= \sum_{n=0}^{\infty} a_n t^n \\ y &= \sum_{n=1}^{\infty} n a_n t^{n-1} \\ y &= \sum_{n=2}^{\infty} n(n-1) a_n t^{n-2}\end{aligned}$$

in the differential equation, multiply everything out, and collect coefficients as before. Although this is theoretically feasible, you wouldn't want to try it unless it were absolutely necessary. Fortunately, that is seldom the case.

Exercises for 9.1.

1. Find a general solution of $y'' + ty' + y = 0$ in the form $a_0 y_1(t) + a_1 y_2(t)$ where $y_1(t)$ and $y_2(t)$ are appropriate power series centered at $t_0 = 0$.
2. The equation $y'' - 2ty' + 2\alpha y = 0$ is called *Hermite's equation*. (It arises among other places in the study of the quantum mechanical analogue of a harmonic oscillator.) Find a general series solution as above in the case $\alpha = 1$. One of the two series is a polynomial.
3. The equation $(1 - t^2)y'' - ty' + \alpha^2 y = 0$ is called *Chebyshev's equation*. Its solutions are used in algorithms for calculating functions in computers and electronic calculators. Find a general series solution as above for $\alpha = 1$. One of the two series is a polynomial.

4. Write a computer program which computes the coefficients a_n from the recurrence relation

$$a_{n+2} = 2 \frac{(n+1)a_{n+1} + (n-1)a_n}{(n+2)(n+1)} \quad n \geq 0.$$

Use the program to determine a_{100} given $a_0 = 1, a_1 = 0$.

5. (a) Show that

$$1 - \frac{t}{2}(\ln(1+t) - \ln(1-t)) = 1 - \sum_{k=1}^{\infty} \frac{t^{2k}}{2k-1}.$$

The right hand side is the series for one of the solutions of Legendre's Equation with $\alpha = 1$. (The other solution is t , and the left hand side is the solution obtained from t by reduction of order.)

(b) Show that the radius of convergence of the above series is 1.

(c) Using the recurrence relation for Legendre's equation, show that both the even and odd series have radius of convergence 1 except in the case that α is a positive integer, in which case one of the two is a polynomial.

6. Without finding the solutions, find a lower bound on the radius of convergence of a power series solution for each of the following equations with the series centered at the indicated point.

(a) $(2+t)(3-t)y'' + 2y' + 3t^2y = 0$ at $t_0 = 0$.

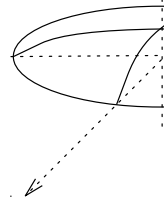
(b) $(2+t^2)y'' - ty' - 3y = 0$ at $t_0 = 1$.

7. Let $y_1(t)$ be the power series solution to $y'' + p(t)y' + q(t)y'' = 0$ obtained at an ordinary point by setting $a_0 = 1, a_1 = 0$. Similarly, let $y_2(t)$ be the solution obtained by setting $a_0 = 0, a_1 = 1$. How can you conclude that the pair $\{y_1, y_2\}$ is linearly independent?

9.2 Partial Differential Equations

You probably have learned by now that certain partial differential equations such as Laplace's Equation or the Wave Equation govern the behavior of important physical systems. The solution of such equations leads directly to the consideration of second order linear differential equations, and it is this fact that lends such equations much of their importance. In this section, we show how an interesting physical problem leads to Bessel's equation

$$t^2y'' + ty' + (t^2 - m^2)y = 0 \quad \text{where } m = 0, 1, 2, \dots$$



In the sections that follow we shall discuss methods for solving Bessel's equation and related equations by use of infinite series.

The physical problem we shall consider is that of determining the possible vibrations of a circular drum head. We model such a drum head as a disk of radius a in the x, y -plane centered at the origin. We suppose that the circumference of the disk is fixed, but that other points may be displaced upward or downward in the z -direction. The displacement z is a function of both position (x, y) and time t . If we assume that the displacement is small, then to a high degree of approximation, this function satisfies the wave equation

$$\frac{1}{v^2} \frac{\partial^2 z}{\partial t^2} = \nabla^2 z$$

where v is a constant and ∇^2 is the Laplace operator

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}.$$

(You will study this equation shortly in physics if you have not done so already.)

Because the problem exhibits circular symmetry, it is appropriate to switch to polar coordinates, and then the Laplace operator takes the form

$$\nabla^2 = \frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r} + \frac{1}{r^2} \frac{\partial^2}{\partial \theta^2}.$$

(See the exercises for Chapter V, Section 13.) Thus the wave equation may be rewritten

$$\frac{1}{v^2} \frac{\partial^2 z}{\partial t^2} = \frac{\partial^2 z}{\partial r^2} + \frac{1}{r} \frac{\partial z}{\partial r} + \frac{1}{r^2} \frac{\partial^2 z}{\partial \theta^2}. \quad (144)$$

Since the circumference of the disk is fixed, we must add the *boundary* condition $z(a, \theta, t) = 0$ for all θ and all t .

A complete study of such equations will be undertaken in your course next year in Fourier series and boundary value problems. For the moment we shall consider only solutions which can be expressed

$$z = T(t)R(r)\Theta(\theta) \quad (145)$$

where the variables have been *separated* out in three functions, each of which depends only on one of the variables. (The method employed here is called *separation of variables*. It is similar in spirit to the method employed previously for ordinary differential equations, but of course the context is entirely different, so the two methods should not be confused with one another.) It should be emphasized that the general solution of the wave equation cannot be so expressed, but as you shall see next year, it can be expressed as a sum (usually infinite) of such functions. The boundary condition for a separated solution in this case is simply $R(a) = 0$.

If we substitute (145) in equation (144), the partial derivatives become ordinary derivatives of the relevant functions and we obtain

$$\frac{1}{v^2} T'' R \Theta = T R'' \Theta + \frac{1}{r} T R' \Theta + \frac{1}{r^2} T R \Theta''.$$

Divide through by $z = T R \Theta$ to obtain

$$\frac{1}{v^2} \frac{T''}{T} = \frac{R''}{R} + \frac{1}{r} \frac{R'}{R} + \frac{1}{r^2} \frac{\Theta''}{\Theta}.$$

In this equation, the left hand side $\frac{1}{v^2} \frac{T''}{T}$ depends only on t , and the right hand side does not depend on t . Hence, both equal the same constant γ , i.e.,

$$\begin{aligned} \frac{1}{v^2} \frac{T''}{T} &= \gamma \\ \frac{R''}{R} + \frac{1}{r} \frac{R'}{R} + \frac{1}{r^2} \frac{\Theta''}{\Theta} &= \gamma. \end{aligned}$$

The second of these equations may be rewritten

$$\frac{\Theta''}{\Theta} = r^2 (\text{an expression depending only on } r),$$

so by similar reasoning, it must be equal to a constant μ . Thus,

$$\begin{aligned} \frac{\Theta''}{\Theta} &= \mu \\ \Theta'' - \mu \Theta &= 0. \end{aligned}$$

This is a simple second order equation with known solutions. If μ is positive, the general solution has the form $C_1 e^{\sqrt{\mu} \theta} + C_2 e^{-\sqrt{\mu} \theta}$. However, the function Θ must satisfy the *periodicity* condition

$$\Theta(\theta + 2\pi) = \Theta(\theta) \quad \text{for every } \theta$$

since adding 2π to the polar angle θ does not change the point represented. The solution listed above does not satisfy this condition so we conclude that $\mu \leq 0$. In that case, the general solution is $\Theta = C_1 \cos \sqrt{|\mu|} \theta + C_2 \sin \sqrt{|\mu|} \theta$. Moreover, the periodicity condition tells us that $\sqrt{|\mu|}$ must be an integer. Thus we may take $\mu = -m^2$ where m is an integer. We may even assume that that $m \geq 0$ since changing the sign of m makes no essential difference in the form of the general solution

$$\Theta = C_1 \cos m\theta + C_2 \sin m\theta \quad m = 0, 1, 2, \dots$$

If we now put $\Theta''/\Theta = -m^2$ back into the separated equation, we obtain

$$\frac{1}{v^2} \frac{T''}{T} = \frac{R''}{R} + \frac{1}{r} \frac{R'}{R} + \frac{1}{r^2} (-m^2) = \gamma$$

where γ is a constant. It turns out that $\gamma < 0$, but the argument is somewhat more complicated than that given above for μ . One way to approach this would be as follows. The above equation gives the following equation for T

$$T'' - \gamma v^2 T = 0.$$

We know by observation (and common sense) that the motion of the drum head is oscillatory. If $\gamma > 0$, this equation has non-periodic exponential solutions (as in the previous argument for Θ). Similarly, $\gamma = 0$ implies that $T = c_1 t + c_2$, which would make sense only if $c_1 = c_2 = 0$. That corresponds to the solution in which the drum does not vibrate at all, and it is not very interesting. Hence, the only remaining possibility is $\gamma < 0$, in which case we get periodic oscillations:

$$T(t) = C_1 \cos(\sqrt{|\gamma|} vt) + C_2 \sin(\sqrt{|\gamma|} vt).$$

This argument is a bit unsatisfactory for the following reason. We should be able to derive *as a conclusion* the fact that the solution is periodic in time. After all, the purpose of a physical theory is to predict as much as we can with as few assumptions as possible. It is in fact possible to show that $\gamma < 0$ by another argument (discussed in the Exercises) which only uses the fact that the drum head is fixed at its circumference.

Write $\gamma = -\lambda$ where $\lambda = |\gamma|$, and consider the equation for R

$$\begin{aligned} \frac{R''}{R} + \frac{1}{r} \frac{R'}{R} + \frac{1}{r^2} (-m^2) &= -\lambda \\ r^2 R'' + r R' + (\lambda r^2 - m^2) R &= 0 \end{aligned}$$

where $m = 0, 1, 2, \dots$, $\lambda > 0$, and R satisfies the boundary condition $R(a) = 0$. It is usual to make one last transformation to simplify this equation, namely introduce a new variable $s = \sqrt{\lambda} r$. Then if $S(s) = R(r)$, we have

$$\begin{aligned} \frac{dR}{dr} &= \frac{dS}{ds} \frac{ds}{dr} = \sqrt{\lambda} \frac{dS}{ds} \\ \frac{d^2 R}{dr^2} &= \frac{d}{dr} \frac{dR}{dr} = \sqrt{\lambda} \frac{d}{dr} \frac{dS}{ds} = \sqrt{\lambda} \sqrt{\lambda} \frac{d}{ds} \frac{dS}{ds} = \lambda \frac{d^2 S}{ds^2}. \end{aligned}$$

Thus

$$\begin{aligned} \lambda r^2 S'' + \sqrt{\lambda} r S' + (\lambda r^2 - m^2) S &= 0 \\ \text{or} \quad s^2 S'' + s S' + (s^2 - m^2) S &= 0. \end{aligned}$$

The last equation is just *Bessel's Equation*, and we shall see how to solve it and similar equations in the next sections in this chapter. Note that in terms of the function $S(s)$, the boundary condition becomes

$$S(\sqrt{\lambda} a) = 0.$$

On the other hand, as mentioned above, $\sqrt{\lambda} = \sqrt{|\gamma|}$ is a factor in determining the frequency of oscillation of the drum. Hence, finding the roots of the equation $S(x) = 0$ is a matter of some interest.

If we switch back to calling the independent variable t and the dependent variable y , Bessel's Equation takes the form used earlier

$$t^2 y'' + ty' + (t^2 - m^2)y = 0$$

or

$$y'' + \frac{1}{t}y' + \frac{t^2 - m^2}{t^2}y = 0.$$

Of course, t here need not bear any relation to 'time'. In fact, in the above analysis, t came from the polar coordinate r . That means we have the following quandary. The value $t = 0$ (the origin in the above discussion) may be a specially interesting point. However, it is also a singular point for the differential equation. Thus, we may have to use series solutions *centered at a singular point*, and that is rather different from what we did previously for ordinary points.

Exercises for 9.2.

1. Consider solutions of Laplace's equation $\nabla^2 z = \frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = 0$ of the form $z = X(x)Y(y)$.

(a) Derive the equation

$$\frac{X''(x)}{X(x)} + \frac{Y''(y)}{Y(y)} = 0.$$

(b) Conclude that X and Y satisfy the second order differential equations

$$X'' + cX = 0 \quad \text{and} \quad Y'' - cY = 0$$

where c is some constant.

(c) Suppose we want to find a non-zero solution of the above form on the unit square in the first quadrant under the assumption that $z(x, 0) = z(0, y) = z(1, y) = 0$. Show that this leads to the conditions $X(0) = X(1) = 0$ and $Y(0) = 0$.

(d) Under these assumptions show that c cannot be negative. Hint: Solve the differential equation for X under the assumption that $c < 0$ and conclude that the solution cannot vanish for $x = 0$ and $x = 1$ unless it is identically zero.

(e) Show that $c = (k\pi)^2$ with $k = 0, 1, 2, \dots$. Find the general form of $X(x)$ and $Y(y)$.

2. (Optional) Show that $\gamma < 0$, where γ is the quantity discussed in the Section. Use the following argument.

(a) Put $z = T(t)U(r, \theta)$ in $\frac{1}{v^2} \frac{\partial^2 z}{\partial t^2} = \nabla^2 z$ and derive the equation

$$\frac{1}{v^2} \frac{T''}{T} = \frac{\nabla^2 U}{U} = \gamma.$$

(In the notation of the section, $U(r, \theta) = R(r)\Theta(\theta)$. γ is the same as before.) Thus, we have

$$\nabla^2 U = \gamma U. \quad (\text{A})$$

(b) A necessary detour: Apply the normal form of Green's Theorem

$$\int_{\partial D} \mathbf{F} \cdot \mathbf{N} ds = \iint_D \nabla \cdot \mathbf{F} dA$$

to the vector field $\mathbf{F} = U \nabla U$ to derive the formula

$$\int_{\partial D} U \nabla U \cdot \mathbf{N} ds = \iint_D (|\nabla U|^2 + U \nabla^2 U) dA. \quad (\text{B})$$

(c) Let D be a disk of radius a centered at the origin. Assume that

$$U(a, \theta) = 0 \quad (\text{C})$$

for all θ , i.e., that z vanishes on the boundary of D . Use (A), (B), and (C) to derive a contradiction to the assumption that $\gamma \geq 0$. What conclusion can you draw if $\gamma = 0$?

9.3 Regular Singular Points and the Method of Frobenius

The general problem we want to consider now is how to solve an equation of the form $y'' + p(t)y' + q(t)y = 0$ by expanding in a series *centered at a singular point*. To understand the process, we start by reviewing the simplest case which is *Euler's Equation*

$$y'' + \frac{\alpha}{t} y' + \frac{\beta}{t^2} y = 0$$

where α and β are constants. To solve it in the vicinity of the singular point $t = 0$, we try a solution of the form $y = t^r$. Then, $y' = r t^{r-1}$ and $y'' = r(r-1)t^{r-2}$, so

the equation becomes

$$\begin{aligned} r(r-1)t^{r-2} + \alpha r \frac{t^{r-1}}{t} + \beta \frac{t^r}{t^2} &= 0 \\ (r(r-1) + \alpha r + \beta)t^{r-2} &= 0 \\ r(r-1) + \alpha r + \beta &= 0 \\ r^2 + (\alpha - 1)r + \beta &= 0. \end{aligned}$$

This is a quadratic equation with two roots r_1, r_2 , so we get two solutions $y_1 = t^{r_1}$ and $y_2 = t^{r_2}$. If the roots are different, these form a linearly independent pair, and the general solution is

$$y = C_1 t^{r_1} + C_2 t^{r_2}.$$

If the roots are equal, i.e., $r_1 = r_2 = r$, then $y_1 = t^r$ is one solution, and we may use reduction of order to find another solution. It turns out to be $y_2 = t^r \ln t$. (See the Exercises.) The general solution is

$$y = C_1 t^r + C_2 t^r \ln t.$$

There are a couple of things to notice about the above process. First, the method depended critically on the fact that t occurred precisely to the right powers in the two denominators. Otherwise, we might not have ended up with the common factor t^{r-2} . Secondly, the solutions of the quadratic equation need not be positive integers; they could be negative integers, fractions, or worse. In such cases t^r may exhibit some singular behavior at $t = 0$. This will certainly be the case if $r < 0$, since in that case $t^r = 1/t^{|r|}$ blows up as $t \rightarrow 0$. If $r > 0$ but r is not an integer, then t^r is continuous at $t = 0$, but it may have discontinuous derivatives of some order. For example, if $y = t^{5/3}$, then $y' = (5/3)t^{2/3}$ and $y'' = (10/9)t^{-1/3}$, which is not continuous at $t = 0$. Hence, the singularity at $t = 0$ in the differential equation tends to show up in some way in the solution. (In general, the roots r could even be complex numbers, which complicates the matter even more. In this course, we shall ignore that possibility.)

Consider now the general equation

$$y'' + p(t)y' + q(t)y = 0,$$

and suppose $p(t)$ or $q(t)$ is singular at $t = 0$. We say that $t = 0$ is a *regular singular point* if we can write

$$p(t) = \frac{\bar{p}(t)}{t} \quad q(t) = \frac{\bar{q}(t)}{t^2}$$

where $\bar{p}(t)$ and $\bar{q}(t)$ are *analytic* in the vicinity of $t = 0$. This means that the differential equation has the form

$$\begin{aligned} y'' + \frac{\bar{p}(t)}{t}y' + \frac{\bar{q}(t)}{t^2}y &= 0 \\ \text{or} \quad t^2 y'' + t \bar{p}(t)y' + \bar{q}(t)y &= 0. \end{aligned}$$

(This is what we get if we replace the constants in Euler's Equation by analytic functions.)

More generally, we say that $t = t_0$ is a regular singular point of the differential equation if it may be rewritten

$$y'' + \frac{\bar{p}(t)}{t - t_0}y' + \frac{\bar{q}(t)}{(t - t_0)^2}y = 0$$

or $(t - t_0)^2y'' + (t - t_0)\bar{p}(t)y' + \bar{q}(t)y = 0$

where $\bar{p}(t)$ and $\bar{q}(t)$ are analytic functions in the vicinity of $t = t_0$.

Example 182 Bessel's Equation

$$y'' + \frac{1}{t}y' + \frac{t^2 - m^2}{t^2}y = 0$$

has a regular singular point at $t = 0$.

Example 183 Legendre's Equation

$$y'' - \frac{2t}{1 - t^2}y' + \frac{\alpha(\alpha + 1)}{1 - t^2}y = 0$$

has regular singular points both at $t = 1$ and $t = -1$.

For, at $t = 1$, we may write

$$p(t) = \frac{-2t}{1 - t^2} = \frac{2t}{(t - 1)(t + 1)} = \frac{2t/(t + 1)}{t - 1}$$

$$q(t) = \frac{\alpha(\alpha + 1)}{1 - t^2} = \frac{-\alpha(\alpha + 1)}{(t - 1)(t + 1)} = \frac{-\alpha(\alpha + 1)(t - 1)/(t + 1)}{(t - 1)^2}$$

and

$$\bar{p}(t) = \frac{2t}{t + 1}$$

$$\bar{q}(t) = \frac{-\alpha(\alpha + 1)(t - 1)}{t + 1}$$

are analytic functions near $t = 1$. (They are of course singular at $t = -1$, but that is far enough away not to matter.)

A similar argument which reverses the roles of $t - 1$ and $t + 1$ shows that $t = -1$ is also a regular singular point.

Example 184 The equation

$$y'' - \frac{2}{t^2}y' + 5y = 0$$

has an *irregular* singular point at $t = 0$. In this case, the best we can do with $p(t)$ is

$$\frac{-2}{t^2} = \frac{-2/t}{t}$$

and $-2/t$ is certainly not analytic at $t = 0$.

To solve an equation with a regular singular point at $t = t_0$, we allow for the possibility that the solution is singular at $t = t_0$, but we hope that the singularity won't be worse than a negative or fractional power of $t - t_0$, as in the case of Euler's Equation. That is, we try for a solution of the form

$$y = (t - t_0)^r \sum_{n=0}^{\infty} a_n (t - t_0)^n.$$

Since the power series is analytic, this amounts to trying for a solution of the form $y = (t - t_0)^r g(t)$ where $g(t)$ is analytic near t_0 . This method is called the *method of Frobenius*.

There is one technical problem with the method of Frobenius. If r is not an integer, then *by definition* $(t - t_0)^r = e^{r \ln(t - t_0)}$. Unfortunately, $\ln(t - t_0)$ is undefined for $t - t_0 < 0$. Fortunately, since $t = t_0$ is a singular point of the differential equation, one is usually interested either in the case $t > t_0$ or $t < t_0$, but one does not usually have to worry about going from one to the other. We shall concentrate in this course on the case $t > t_0$. For the case $t < t_0$, similar methods work except that you should use $|t - t_0|^r$ instead of $(t - t_0)^r$.

The Method of Frobenius for Bessel's Equation

We want to solve

$$t^2 y'' + t y' + (t^2 - m^2)y = 0$$

near the regular singular point $t_0 = 0$. We look for a solution defined for $t > 0$. Clearly, we may assume m is non-negative since its square is what appears in the equation. In interesting applications m is a non-negative integer, but for the moment we make no assumptions about m except that $m \geq 0$. The method of Frobenius suggests trying

$$y = t^r \sum_{n=0}^{\infty} a_n t^n = \sum_{n=0}^{\infty} a_n t^{n+r}.$$

Note that we may assume here that $a_0 \neq 0$ since if it were zero that would mean the series would start with a positive power of t which could be factored out from each term and absorbed in t^r by increasing r .

As before, we calculate

$$y' = \sum_{n=0}^{\infty} (n+r) a_n t^{n+r-1} \quad \text{and} \quad y'' = \sum_{n=0}^{\infty} (n+r)(n+r-1) a_n t^{n+r-2}.$$

(Note however that we can't play any games with the lower index as we did for ordinary points.) Thus,

$$\begin{aligned}
 t^2 y'' &= \sum_{n=0}^{\infty} (n+r)(n+r-1)a_n t^{n+r} \\
 t y' &= \sum_{n=0}^{\infty} (n+r)a_n t^{n+r} \\
 t^2 y &= \sum_{n=0}^{\infty} a_n t^{n+r+2} = \sum_{n-2=0}^{\infty} a_{n-2} t^{n+r} \\
 &= \sum_{n=2}^{\infty} a_{n-2} t^{n+r} \\
 -m^2 y &= \sum_{n=0}^{\infty} (-m^2 a_n) t^{n+r}.
 \end{aligned}$$

Note that after the adjustments, one of the sums starts at $n = 2$. That means that when we add up the terms for each n , we have to consider those for $n = 0$ and $n = 1$ separately since they don't involve terms from the aforementioned sum. Thus, for $n = 0$, we get

$$0 = r(r-1)a_0 + ra_0 - m^2 a_0$$

while for $n = 1$, we get

$$0 = (r+1)ra_1 + (r+1)a_1 - m^2 a_1.$$

The general rule starts with $n = 2$

$$0 = (n+r)(n+r-1)a_n + (n+r)a_n + a_{n-2} - m^2 a_n \quad \text{for } n \geq 2.$$

These relations may be simplified. For, $n = 0$, we get

$$(r^2 - m^2)a_0 = 0.$$

However, since by assumption $a_0 \neq 0$, we get

$$r^2 - m^2 = 0.$$

This quadratic equation in r is called the *indicial equation*. In this particular case, it is easy to solve

$$r = \pm m.$$

(Note that if $m = 0$, this is a double root!) For $n = 1$, (using $r^2 = m^2$) we get

$$((r+1)^2 - m^2)a_1 = (2r+1)a_1 = 0 \tag{147}$$

Except in the case $r = -1/2$, this implies that $a_1 = 0$. Finally, for $n \geq 2$, the coefficient of a_n is $(n+r)(n+r-1) + (n+r) - m^2 = (n+r)^2 - m^2$, so we get

$$[(n+r)^2 - m^2]a_n + a_{n-2} = 0. \tag{148}$$

We shall now consider separately what happens for each root $r = \pm m$ of the indicial equation. (Some people prefer to see how far they can get without specifying r , and then they put r equal to each of the roots when they can't proceed further.)

Solution of Bessel's Equation for the Positive Root of the Indicial Equation

Take $r = m \geq 0$. Then, $r \neq -1/2$, so $a_1 = 0$. For $n \geq 2$, the coefficient of a_n is $(n+m)^2 - m^2 = n^2 + 2nm = n(n+2m)$, and we may solve

$$a_n = -\frac{a_{n-2}}{n(n+2m)}.$$

This recurrence relation allows us to determine a_n for all $n > 0$. First of all, since $a_1 = 0$, it follows that $a_3 = a_5 = \cdots = 0$, i.e., all odd numbered coefficients are zero. For n even, we have

$$\begin{aligned} n=2 \quad a_2 &= -\frac{a_0}{2(2+2m)} = -\frac{a_0}{2^2(m+1)} \\ n=4 \quad a_4 &= -\frac{a_2}{4(4+2m)} = \frac{a_0}{2^5(m+2)(m+1)} \\ n=6 \quad a_6 &= -\frac{a_4}{6(6+2m)} = -\frac{a_0}{3 \cdot 2^7(m+3)(m+2)(m+1)} \\ n=8 \quad a_8 &= -\frac{a_4}{6(6+2m)} = \frac{a_0}{4 \cdot 3 \cdot 2^9(m+4)(m+3)(m+2)(m+1)} \\ &= \frac{a_0}{4!2^8(m+4)(m+3)(m+2)(m+1)} \\ &\vdots \end{aligned}$$

The general rule may be written

$$a_{2k} = (-1)^k \frac{a_0}{k!2^{2k}(m+k)(m+k-1)\cdots(m+2)(m+1)} \quad k = 0, 1, 2, \dots$$

Note the quantity $(m+k)(m+k-1)\cdots(m+2)(m+1)$ is similar to a factorial in which each term has an extra addend m . If we interpret this product to be 1 if $k = 0$, and recall that $0! = 1$, then the formula is valid, as indicated, for $k = 0$.

The corresponding series solution is

$$y = t^m \sum_{n=0}^{\infty} a_n t^n = a_0 t^m \sum_{k=0}^{\infty} \frac{(-1)^k}{k!2^{2k}(m+k)(m+k-1)\cdots(m+2)(m+1)} t^{2k}.$$

Note that since $m \geq 0$, this solution is actually continuous at $t = 0$. (It is the product of the continuous function t^m with the sum of a power series, i.e., an analytic function.) If m is a non-negative integer, the solution is even analytic at $t = 0$. If m is positive but not an integer, the solution is definitely not analytic since a sufficiently high derivative of t^m will involve a negative power and so fail to exist at $t = 0$. (A function which is analytic at $t = 0$ has derivatives of every order at $t = 0$. They are just the coefficients (except for factorials) of its Taylor series centered at $t = 0$.)

The constant a_0 is arbitrary and determined by initial conditions. However, we may pick out one specific solution and any other such solution will be a constant multiple of it. It is often useful to adjust the constant a_0 for the distinguished solution so that the formulas work out nicely. If m is a non-negative integer the most common choice is $a_0 = 1/(2^m m!)$. This yields

$$\begin{aligned} y &= \frac{1}{2^m m!} t^m \sum_{k=0}^{\infty} \frac{(-1)^k}{2^{2k} k! (m+k)(m+k-1) \dots (m+2)(m+1)} t^{2k} \\ &= \sum_{k=0}^{\infty} \frac{(-1)^k}{2^{2k+m} k! (m+k)!} t^{2k+m} \\ &= \sum_{k=0}^{\infty} \frac{(-1)^k}{k! (m+k)!} \left(\frac{t}{2}\right)^{2k+m}. \end{aligned}$$

For m a non-negative integer, this solution is called a *Bessel Function of the first kind*, and it is denoted

$$J_m(t) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! (m+k)!} \left(\frac{t}{2}\right)^{2k+m}.$$

The ratio test shows that the series converges for all t , so its sum is an analytic function for all t . Two interesting cases are

$$\begin{aligned} J_0(t) &= \sum_{k=0}^{\infty} \frac{(-1)^k}{(k!)^2} \left(\frac{t}{2}\right)^{2k} \\ J_1(t) &= \sum_{k=0}^{\infty} \frac{(-1)^k}{(k+1)! k!} \left(\frac{t}{2}\right)^{2k+1}. \end{aligned}$$

If $m > 0$ but m is not an integer, everything above works except that the resulting function is not analytic, but as mentioned above, it is bounded and continuous as $t \rightarrow 0$. The choice of the constant a_0 for a distinguished solution is trickier. The term 2^m poses no problems, but we need an analogue of $m!$ for m a positive real number which is not an integer. It turns out that there is a real valued function $\Gamma(x)$ of the real variable x , which is even analytic for non-negative values of x , and which satisfies the rules

$$\Gamma(x+1) = x\Gamma(x) \quad \Gamma(1) = 1. \quad (149)$$

It is called appropriately enough the *Gamma function*. (See the Exercises for its definition.) It follows from the rules (149) that if m is a positive integer,

$$\begin{aligned} \Gamma(m) &= (m-1)\Gamma(m-1) = (m-1)(m-2)\Gamma(m-2) = \dots \\ &= (m-1) \dots 2 \cdot 1 \Gamma(1) = (m-1)!. \end{aligned}$$

This is usually rewritten with m replaced by $m+1$

$$\Gamma(m+1) = m! \quad \text{for } m = 0, 1, 2, \dots$$

You will study the Gamma function in more detail in your complex variables course.

Using the Gamma function, we may take $a_0 = \frac{1}{2^m \Gamma(m+1)}$. Then, $2^m 2^{2k} = 2^{2k+m}$, and

$$\Gamma(m+1)(m+1)(m+2)\dots(m+k) = \Gamma(m+k+1).$$

So, combining a_0 with a_{2k} , we obtain the solution

$$\begin{aligned} J_m(t) &= \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(m+k+1)} \left(\frac{t}{2}\right)^{2k+m} \\ &= \left(\frac{t}{2}\right)^m \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(m+k+1)} \left(\frac{t}{2}\right)^{2k} \end{aligned}$$

where the fractional power has been put in front in the second equation to emphasize the non-analytic part of the solution.

One interesting case is $m = 1/2$.

$$\begin{aligned} J_{1/2}(t) &= \frac{t^{1/2}}{2^{1/2} \Gamma(3/2)} \sum_{k=0}^{\infty} \frac{(-1)^k}{2^{2k} k! (1/2+k)(1/2+k-1)\dots(1/2+1)} t^{2k} \\ &= \frac{1}{t^{1/2} 2^{1/2} \Gamma(3/2)} \sum_{k=0}^{\infty} (-1)^k \frac{t^{2k+1}}{2^k k! (2k+1)(2k-1)\dots 3} \\ &= \frac{1}{t^{1/2} 2^{1/2} \Gamma(3/2)} \sum_{k=0}^{\infty} (-1)^k \frac{t^{2k+1}}{(2k)(2k-2)\dots 4 \cdot 2 (2k+1)(2k-1)\dots 3} \\ &= \frac{1}{t^{1/2} 2^{1/2} \Gamma(3/2)} \sum_{k=0}^{\infty} (-1)^k \frac{t^{2k+1}}{(2k+1)!} \\ &= \frac{1}{t^{1/2} 2^{1/2} \Gamma(3/2)} \sin t. \end{aligned}$$

However, $\Gamma(3/2) = (1/2)\Gamma(1/2)$, and it may be shown that $\Gamma(1/2) = \sqrt{\pi}$. Thus, after some algebra, we get

$$J_{1/2}(t) = \sqrt{\frac{2}{\pi}} \frac{\sin t}{\sqrt{t}}.$$

Solution of Bessel's Equation for the Negative Root of the Indicial Equation Suppose $m > 0$. By considering the positive root $r = m$ of the indicial equation $r^2 - m^2 = 0$, we found one solution of Bessel's Equation. We now attempt to find a second linearly independent solution by considering the negative root $r = -m$. For this root, equation (147) for $n = 1$ becomes

$$(2r+1)a_1 = (1-2m)a_1 = 0,$$

which, as earlier, implies that $a_1 = 0$ except in the case $m = 1/2, r = -1/2$. However, even in that case we need only find one additional solution, so in any event we shall concentrate on the *even numbered* coefficients.

For $n \geq 2$, equation (148) becomes

$$(n^2 + 2nr)a_n + a_{n-2} = (n^2 - 2nm)a_n + a_{n-2} = 0$$

which may be solved to obtain

$$a_n = \frac{-a_{n-2}}{n(n-2m)} \quad \text{for } n \geq 2,$$

provided $n - 2m \neq 0$. Thus, we may use the recurrence relation to generate coefficients (for n even) for a second solution as long as $m \neq n/2$, i.e., m is not an integer. Assume that is the case. Then we get

$$\begin{aligned} n = 2 & \quad a_2 = -\frac{a_0}{2(2-2m)} = -\frac{a_0}{2^2(1-m)} \\ n = 4 & \quad a_4 = -\frac{a_2}{4(4-2m)} = -\frac{a_0}{2^4 2(1-m)(2-m)} \\ n = 6 & \quad a_6 = -\frac{a_4}{6(6-2m)} = -\frac{a_0}{2^6 3 \cdot 2(1-m)(2-m)(3-m)} \\ & \quad \vdots \end{aligned}$$

Putting $n = 2k$, we get the general rule

$$a_{2k} = \frac{(-1)^k a_0}{2^{2k} k! (1-m)(2-m) \dots (k-m)} \quad k \geq 0.$$

The corresponding solution is

$$y = a_0 t^{-m} \sum_{k=0}^{\infty} \frac{(-1)^k a_0}{k! (1-m)(2-m) \dots (k-m)} \left(\frac{t}{2}\right)^{2k}.$$

If $m > 0$ is not a positive integer, we may set $a_0 = \frac{1}{2^{-m} \Gamma(-m+1)}$. That is similar to what we did above, except that m is replaced by $-m$. The resulting solution is

$$J_{-m}(t) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(-m+k+1)} \left(\frac{t}{2}\right)^{2k-m}.$$

Note that this solution has a singularity for $t = 0$ because of the common factor t^{-m} , so it is certainly not a constant multiple of $J_m(t)$. Hence, we have a linearly independent pair of solutions and we conclude that *if m is not an integer*, the general solution of Bessel's equation is

$$y = C_1 J_m(t) + C_2 J_{-m}(t).$$

The case $m = 1/2$ is interesting. The series can be rewritten

$$\begin{aligned}
 J_{-1/2}(t) &= \frac{1}{2^{-1/2}\Gamma(1/2)} \sum_{k=0}^{\infty} (-1)^k \frac{1}{k!(1-1/2)(2-1/2)\dots(k-1/2)2^{2k}} t^{2k-1/2} \\
 J_{-1/2}(t) &= \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{t}} \sum_{k=0}^{\infty} (-1)^k \frac{1}{k!(1-1/2)(2-1/2)\dots(k-1/2)2^{2k}} t^{2k} \\
 &= \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{t}} \sum_{k=0}^{\infty} (-1)^k \frac{t^{2k}}{2^k k!(2-1)(4-1)\dots(2k-1)} t^{2k} \\
 &= \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{t}} \sum_{k=0}^{\infty} (-1)^k \frac{t^{2k}}{(2k)!} \\
 &= \sqrt{\frac{2}{\pi}} \frac{\cos t}{\sqrt{t}}.
 \end{aligned}$$

Note that we still have to deal with the case that m is an integer since the above method breaks down.

Exercises for 9.3.

- Find general solutions for each of the following equations. (You may find it helpful first to review Chapter VII, Section 4, Exercise 6 and Section 8, Exercise 3.)
 - $t^2 y'' + 3ty' - 3y = 0$.
 - $t^2 y'' - 5ty' + 9y = 0$.
- Assume $r^2 + (\alpha - 1)r + \beta = 0$ has a double root $r = (1 - \alpha)/2$. Then, $y_1 = t^r$ is one solution of Euler's Equation. Verify that the method of reduction of order yields the second solution $y_2 = t^r \ln t$.
- In each of the following cases, tell if the given value of t is a regular singular point for the indicated differential equation.
 - $t(t+1)^2 y'' - 2ty' + 7y = 0$, $t = 0$.
 - $(t-1)^2 ty'' + 3(t-1)y' - y = 0$, $t = 1$.
 - $(t+2)^2 y'' + (t+1)y' + (t+2)y = 0$, $t = -2$.
 - $(\sin t)y'' + 2y' - (\cos t)y = 0$, $t = 0$.
- Find one solution of $ty'' + y' + ty = 0$ at $t_0 = 0$ by the method of Frobenius. You should get $J_0(t)$ up to a multiplicative constant.
 - Show that $\frac{d}{dt} J_0(t) = -J_1(t)$.
 - Show that $\frac{d}{dt} (tJ_1(t)) = tJ_0(t)$.

6. Show that the power series in the definition

$$J_m(t) = \frac{(t/2)^m}{\Gamma(m+1)} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(m+k)(m+k-1)\dots(m+1)} \left(\frac{t}{2}\right)^{2k}$$

has infinite radius of convergence.

7. Calculate $J_0(t)$ to 4 decimal places for $t = 0, 0.1, 0.2, \dots, 0.9, 1.0$.
8. Show by direct substitution in the differential equation $t^2 y'' + ty' + (t^2 - 1/4)y = 0$ that $y_1 = (\sin t)/\sqrt{t}$ and $y_2 = (\cos t)/\sqrt{t}$ are solutions.
9. (Optional) The Gamma function is defined by

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt \quad \text{for } x > 0.$$

(It may also be extended to negative values of x and indeed to complex values of x . The resulting function has singularities at $x = 0, -1, -2, \dots$)

- (a) Show $\Gamma(1) = \int_0^{\infty} e^{-t} dt = 1$
- (b) Show $\Gamma(x+1) = x\Gamma(x)$. Hint: Apply integration by parts to $\Gamma(x+1) = \int_0^{\infty} t^x e^{-t} dt$.
- (c) Show $\Gamma(1/2) = \int_0^{\infty} t^{1/2} e^{-t} dt = \sqrt{\pi}$. Hint: Substitute $t = s^2$ to obtain $\Gamma(1/2) = 2 \int_0^{\infty} e^{-s^2} ds$. Now use the calculation from Chapter IV, Section 8 of $\int_0^{\infty} e^{-u^2/2} du = \sqrt{\pi/2}$.

9.4 The Method of Frobenius. General Theory

In the previous section, we applied the method of Frobenius to Bessel's Equation. For some cases (m not an integer), the method gave a complete solution, but for other cases (m a non-negative integer), it gave only one solution. In the 'bad' cases, we need another method to find a linearly independent pair of solutions from which we can form a general solution of the differential equation.

Before attempting to deal with the 'bad' cases, we should discuss how the method of Frobenius works for other differential equations.

Example 185 We shall try to solve

$$t^2 y'' + 2ty' - (2+t)y = 0 \quad t > 0$$

by a series of the form $y = t^r \sum_{n=0}^{\infty} a_n t^n = \sum_{n=0}^{\infty} a_n t^{n+r}$. Calculating as previously, we have

$$\begin{aligned} t^2 y'' &= \sum_{n=0}^{\infty} (n+r)(n+r-1) a_n t^{n+r} \\ 2ty' &= \sum_{n=0}^{\infty} 2(n+r) a_n t^{n+r} \\ -2y &= \sum_{n=0}^{\infty} (-2a_n) t^{n+r} \\ -ty &= \sum_{n=0}^{\infty} (-a_n) t^{n+r+1} = \sum_{n-1=0}^{\infty} (-a_{n-1}) t^{n-1+r+1} \\ &= \sum_{n=1}^{\infty} (-a_{n-1}) t^{n+r}. \end{aligned}$$

Adding up corresponding powers of t yields for $n = 0$

$$[r(r-1) + 2r - 2]a_0 = 0 \quad (150)$$

and for $n \geq 1$

$$[(n+r)(n+r-1) + 2(n+r) - 2]a_n - a_{n-1} = 0. \quad (151)$$

Since $a_0 \neq 0$, (150) yields the *indicial equation*

$$f(r) = r(r-1) + 2r - 2 = r^2 + r - 2 = (r-1)(r+2) = 0.$$

Note also that for $n \geq 1$, (151) has the form

$$f(n+r)a_n - a_{n-1} = 0$$

where

$$f(n+r) = (n+r)(n+r-1) + 2(n+r) - 2 = (n+r)^2 + (n+r) - 2 = (n+r-1)(n+r+2).$$

(You should look back at the previous section at this point to see what happened for Bessel's equation. The indicial equation was $f(r) = r^2 - m^2 = 0$, while the coefficient of a_n in each of the equations for $n \geq 1$ was $f(r) = (n+r)^2 - m^2$.)

The roots of the indicial equation are $r = 1$ and $r = -2$. Consider first $r = 1$. For $n \geq 1$, $f(n+1) = n(n+3)$ so (151) becomes

$$n(n+3)a_n - a_{n-1} = 0$$

which may be solved to obtain the recurrence relation

$$a_n = \frac{a_{n-1}}{n(n+3)} \quad \text{for } n \geq 1. \quad (153)$$

Note that there is no problem with this relation since the denominator never vanishes. This is not an accident. We shall see below why it happened.

Clearly, this recurrence relation may be used to determine all the coefficients a_n in terms of a_0 . (We leave it to the student to actually do that.) Thus, we obtain one solution of the differential equation which is in fact uniquely determined except for the *non-zero* multiplicative constant a_0 .

Consider next the root $r = -2$ of the indicial equation. For $n \geq 1$, $f(n-2) = (n-3)n$, so (151) becomes

$$(n-3)na_n = a_{n-1}.$$

This can be solved for $n = 1$ and $n = 2$ to obtain

$$\begin{aligned} a_1 &= \frac{a_0}{(-2)1} = -\frac{a_0}{2} \\ a_2 &= \frac{a_1}{(-1)2} = \frac{a_0}{4}. \end{aligned} \tag{154}$$

However, for $n = 3$ we encounter a problem. The equation becomes

$$0 \cdot a_3 = a_2$$

which is consistent only if $a_2 = 0$, and by (154) that contradicts the assumption that $a_0 \neq 0$. Hence, for the root $r = -2$ of the indicial equation, the process breaks down.

Let's see what was going on both in our discussion of Bessel's equation and in the last example. In each case, we had a quadratic indicial equation

$$f(r) = 0.$$

In addition, we had for $n \geq 1$ equations of the form

$$f(n+r)a_n = \text{an expression involving } a_j \text{ with } j < n.$$

(The expression on the right might be zero as for Bessel's equation with $n = 1$.) Suppose $r_1 \geq r_2$ are the two roots of the indicial equation. If $n \geq 1$, it can never be the case that $f(n+r_1) = 0$ since the only other root of the equation $f(r) = 0$ is r_2 which is not larger than r_1 . Hence, we may always solve the above equation to obtain recurrence relations

$$a_n = \frac{\text{expression involving } a_j \text{ with } j < n}{f(n+r_1)}, \quad n \geq 1.$$

For the other root, r_2 the situation is more complicated. If $r_2 + n$ is never a root of the quadratic equation $f(r) = 0$, the above reasoning applies and we obtain recurrence relations

$$a_n = \frac{\text{expression involving } a_j \text{ with } j < n}{f(n+r_2)}, \quad n \geq 1.$$

Thus, we obtain a second solution, and it is not hard to see that the two solutions form a linearly independent pair. (See the appendix to this section.)

There is the possibility, however, that for some integer $n = k$, $r_2 + k = r_1$ is the *larger* root of the indicial equation, i.e., $f(r_2 + k) = 0$. In that case, the k th recursion relation becomes

$$0 = f(k + r_2)a_k = \text{an expression involving } a_j \text{ with } j < k,$$

so the process will break down *unless we are incredibly lucky and the expression on the right happens to be zero*. In that case, we can seize on our great fortune, and set a_k equal to any convenient value. Usually, we just set $a_k = 0$. In any case, the process may continue unimpeded for $n > k$ as soon as we successfully get past the 'barrier' at k .

The upshot of the above analysis is that the method of Frobenius may break down in the 'bad case' that $r_1 - r_2$ is a positive integer. Of course, if $r_1 = r_2$, the method only gives one solution in any case. Hence, you must be on guard whenever $r_1 - r_2$ is a non-negative integer.

If you look back at the previous section, you will see that Bessel's Equation exhibits the phenomena we just described. The roots are $\pm m$, so $m = 0$ is the case of equal roots, and the method of Frobenius only gives one solution. If $m > 0$, the larger of the two roots is $r_1 = m$, and this gives a solution in any case. The smaller of the two roots is $r_2 = -m$, and provided $m - (-m) = 2m$ is not a positive integer, the method of Frobenius also generates a solution for $r_2 = -m$. On the other hand, if $2m$ is a positive integer, then the recurrence relations for $r_2 = -m$ take the form

$$\begin{aligned} 1(1 - 2m)a_1 &= 0 & \text{for } n = 1 \\ n(n - 2m)a_n &= -a_{n-2} & \text{for } n \geq 2. \end{aligned}$$

The first equation ($n = 1$) implies that $a_1 = 0$ except in the case $m = 1/2$. Even in that case, our luck holds out, and we may take $a_1 = 0$ since the right hand side of the equation is zero. Similarly, if $m = k/2$ for some odd integer $k > 1$, then the recursion relation will imply that $a_n = 0$ for every odd $n < k$, and the k th recurrence relation will read

$$k(0)a_k = -a_{k-2} = 0,$$

so we may set $a_k = 0$. It follows that if m is half an odd integer, then we may assume all the odd numbered a_n are zero. For even n , the coefficient $n(n - 2m) = n(n - k)$ is never zero, so there is no problem determining the a_n as previously.

The only remaining case is when m is itself a positive integer. In this case, for $n = 2m$, the coefficient on the left $n(n - 2m)$ is zero, but the quantity a_{n-1} is not, so there is no way to recover. Hence, we must find another method to generate a second solution.

The General Theory We shall explain here why the method of Frobenius behaves the way it does. This section may be omitted your first time through the material, but you should come back a look at it after you have worked some more examples.

Consider the differential equation

$$t^2 y'' + t\bar{p}(t)y' + \bar{q}(t)y = 0$$

where $\bar{p}(t)$ and $\bar{q}(t)$ are analytic functions in a neighborhood of $t = 0$. (That is what it means to say $t = 0$ is a regular singular point.) We shall try to find a solution of the form $y = t^r \sum_{n=0}^{\infty} a_n t^n$ for $t > 0$. (As mentioned earlier, if you want a solution for $t < 0$, replace t^r by $|t|^r$.) Since $\bar{p}(t)$ and $\bar{q}(t)$ are analytic at $t = 0$, they have power series expansions

$$\bar{p}(t) = p_0 + p_1 t + p_2 t^2 + \dots$$

$$\bar{q}(t) = q_0 + q_1 t + q_2 t^2 + \dots$$

Putting these in the differential equation, one term at a time, we have

$$\begin{aligned} t^2 y'' &= \sum_{n=0}^{\infty} a_n (n+r)(n+r-1) t^{n+r} \\ p_0 t y' &= \sum_{n=0}^{\infty} p_0 a_n (n+r) t^{n+r} \\ p_1 t t y' &= p_1 t^2 y' = \sum_{n=0}^{\infty} p_1 a_n (n+r) t^{n+r+1} = \sum_{n=1}^{\infty} p_1 a_{n-1} (n+r-1) t^{n+r} \\ p_2 t^2 t y' &= p_2 t^3 y' = \sum_{n=0}^{\infty} p_2 a_n (n+r) t^{n+r+2} = \sum_{n=2}^{\infty} p_2 a_{n-2} (n+r-2) t^{n+r} \\ &\vdots \quad \text{sums starting with } n = 3, 4, \dots \\ q_0 y &= \sum_{n=0}^{\infty} q_0 a_n t^{n+r} \\ q_1 t y &= \sum_{n=0}^{\infty} q_1 a_n t^{n+r+1} = \sum_{n=1}^{\infty} q_1 a_{n-1} t^{n+r} \\ q_2 t^2 y &= \sum_{n=0}^{\infty} q_2 a_n t^{n+r+2} = \sum_{n=2}^{\infty} q_2 a_{n-2} t^{n+r} \\ &\vdots \quad \text{sums starting with } n = 3, 4, \dots \end{aligned}$$

Adding up coefficients of corresponding powers of t yields for $n = 0$

$$[r(r-1) + p_0 r + q_0] a_0 = 0,$$

and since by assumption $a_0 \neq 0$, we get the indicial equation

$$f(r) = r(r-1) + p_0 r + q_0 = 0.$$

(Notice the similarity to the equation obtained for Euler's equation.) For $n \geq 1$, we obtain equations of the form

$$[(n+r)(n+r-1) + p_0(n+r) + q_0]a_n + \text{lesser numbered terms} = 0.$$

The coefficient of a_n will always be

$$f(n+r) = (n+r)(n+r-1) + p_0(n+r) + q_0,$$

and the additional terms will depend in general on the exact nature of the coefficients $p_1, p_2, \dots, q_1, q_2, \dots$ (You should work out the cases $n = 1$ and $n = 2$ to make sure you understand the argument!)

The above calculation justifies *in general* the conclusions drawn earlier by looking at examples. However, there is still one point that has not been addressed. Even if the method unambiguously generates the coefficients of the series (in terms of a_0), it won't be of much use unless that series has a positive radius of convergence. Determining the radius of convergence of the series generated by the method of Frobenius in any specific case is usually not difficult, but showing that it is not zero in general is hard. We shall leave that for you to investigate by yourself at a future date if you are sufficiently interested. *Introduction to Differential Equations* by Simmons has a good treatment of the question. **Appendix. Why the solution**

pair $\{y_1, y_2\}$ is linearly independent when $r_1 - r_2$ is not an integer We have

$$y_1 = t^{r_1} g_1(t) \quad \text{and} \quad y_2 = t^{r_2} g_2(t)$$

where $g_1(t)$ and $g_2(t)$ are the sums of the series obtained in each case. Since by assumption, the leading coefficient $a_0 \neq 0$, neither of these functions vanishes. It follows that the quotient of the solutions has the form

$$\frac{y_1}{y_2} = \frac{t^{r_1} g_1(t)}{t^{r_2} g_2(t)} = t^{r_1-r_2} g(t)$$

where $g(t) = g_1(t)/g_2(t)$ is a quotient of two analytic functions, neither of which vanishes at $t = 0$, so $g(t)$ is also analytic at $t = 0$ and does not vanish there. If this were constant, we could write

$$t^{r_1-r_2} = \frac{c}{g(t)}$$

and since $g(0) \neq 0$, the right hand side is analytic. On the other hand, if $r_1 - r_2$ is not a positive integer, $t^{r_1-r_2}$ is definitely not analytic. (If it is a positive integer, then it vanishes at $t = 0$, but the right hand side does not.)

Exercises for 9.4.

All the problems in this section concern linear second order homogeneous differential equations with $t = 0$ a regular singular point.

1. Assuming $a_0 = 1$, use the recurrence relation

$$a_n = \frac{a_{n-1}}{n(n+3)} \quad \text{for } n \geq 1$$

to obtain a general formula for a_n . What is the radius of convergence of the series $t \sum_{n=0}^{\infty} a_n t^n$?

2. In each of the following cases, the two roots of the indicial equation do not differ by a non-negative integer. Find a pair of linearly independent solutions by the method of Frobenius

(a) $4t^2 y'' + 2ty' - (2+t)y = 0$.

(b) $2ty'' + 3y' + y = 0$.

3. In each of the following cases, the indicial equation has two equal roots. Find one solution by the method of Frobenius.

(a) $t^2 y'' - ty' + (1-t)y = 0$.

(b) $t^2 y'' + 3ty' + (1+2t)y = 0$.

4. In each of the following cases, the two roots of the indicial equation differ by a positive integer. Find a solution for the larger root by the method of Frobenius. Determine if you can also find a solution for the smaller root by setting the appropriate $a_k = 0$.

(a) $t^2 y'' - (2+t)y = 0$.

(b) $t^2 y'' - (4t + t^2)y' + 6y = 0$.

(c) $t^2 y'' + (t - t^2)y' - y = 0$.

5. (Optional) Suppose the two roots $r_1 \geq r_2$ of the indicial equation differ by a positive integer $k = r_1 - r_2$. Let $y_1(t)$ be the solution obtained for r_1 by the method of Frobenius with $a_0 = 1$. Suppose that we are in the fortunate situation that the method of Frobenius also yields a solution $y_2(t)$ for the root r_2 , but that we do not assume $a_k = 0$. Show that $y_2(t) - a_k y_1(t)$ may be expanded in a series of the form $\sum_{n=0}^{\infty} b_n t^{n+r_2}$ with $b_k = 0$.

9.5 Second Solutions in the Bad Cases

We consider next what to do when the method of Frobenius fails to produce a second solution. We suppose that we are working at $t = 0$ which is assumed to be a regular singular point. It is clear how to modify the formulas and arguments for an arbitrary regular singular point t_0 .

The Case of Equal Roots Suppose first that the indicial equation

$$f(r) = 0$$

has a double root r . Let $y_1(t)$ denote the solution obtained by the method of Frobenius for this root. Then, we may apply the method of reduction of order to obtain a second solution. It turns out that this second solution always has the form

$$y_2(t) = y_1(t) \ln t + \sum_{n=1}^{\infty} c_n t^{n+r}. \quad (155)$$

Note that the summation starts with $n = 1$. You should compare (155) with the second solution for Euler's equation in the case of a double root. In that case, we had $y_1 = t^r$ and $y_2 = t^r \ln t = y_1 \ln t$, so the logarithmic term is not a complete surprise.

We shall see later how the method of reduction of order leads to such a solution, but first let's look at an example.

Example 186 Consider Bessel's equation for $m = 0$

$$t^2 y'' + ty' + (t^2 - 0^2)y = t^2 y'' + ty' + t^2 y = 0.$$

$r = 0$ is a double root and the first solution is given by

$$y_1 = J_0(t) = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(k!)^2} \left(\frac{t}{2}\right)^{2k}.$$

As suggested above, let's try a second solution of the form

$$y = y_1(t) \ln t + \sum_{n=1}^{\infty} c_n t^n.$$

Then

$$\begin{aligned} y' &= y_1'(t) \ln t + \frac{y_1(t)}{t} + \sum_{n=1}^{\infty} n c_n t^{n-1} \\ y'' &= y_1''(t) \ln t + 2 \frac{y_1'(t)}{t} - \frac{y_1(t)}{t^2} + \sum_{n=2}^{\infty} n(n-1) c_n t^{n-2}. \end{aligned}$$

Thus, renumbering as needed, we get

$$\begin{array}{llll} t^2 y'' = t^2 y_1'' \ln t & + & 2t y_1' & - & y_1 & + & \sum_{n=1}^{\infty} n(n-1) c_n t^n \\ ty' = t y_1' \ln t & & & + & y_1 & + & \sum_{n=1}^{\infty} n c_n t^n \\ t^2 y = t^2 y_1 \ln t & & & & & + & \sum_{n=3}^{\infty} c_{n-2} t^n \end{array}$$

Add this all up to get zero. The right side includes the terms

$$t^2 y_1'' \ln t + t y_1' \ln t + t^2 y_1 \ln t = (t^2 y_1'' + t y_1' + t^2 y_1) \ln t = 0$$

since y_1 is a solution of the differential equation. The terms $-y_1$ and $+y_1$ cancel, so we are left only with $2ty_1'(t)$ and the summation terms. The coefficient of t^n for $n \geq 3$ is

$$n(n-1)c_n + nc_n + c_{n-2} = n^2 c_n + c_{n-2},$$

but we also have two additional terms, those for $n = 1$ and $n = 2$. Putting this all together, we get

$$2ty_1'(t) + c_1 t + 4c_2 t^2 + \sum_{n=3}^{\infty} [n^2 c_n + c_{n-2}] t^n = 0,$$

or

$$c_1 t + 4c_2 t^2 + \sum_{n=3}^{\infty} [n^2 c_n + c_{n-2}] t^n = -2ty_1'(t).$$

However, we may evaluate $ty_1'(t)$ without too much difficulty.

$$y_1 = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(k!)^2 2^{2k}} t^{2k}$$

so

$$\begin{aligned} -2ty_1' &= -2t \sum_{k=1}^{\infty} (-1)^k \frac{1}{(k!)^2 2^{2k}} 2kt^{2k-1} \\ &= \sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k!(k-1)! 2^{2k-2}} t^{2k}. \end{aligned}$$

Thus,

$$c_1 t + 4c_2 t^2 + \sum_{n=3}^{\infty} [n^2 c_n + c_{n-2}] t^n = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k!(k-1)! 2^{2k-2}} t^{2k}.$$

Now compare corresponding powers of t on the two sides, and remember that only even powers appear on the right. For $n = 1$, there are no terms on the right, so

$$c_1 = 0.$$

For $n = 2, k = 1$, we have

$$\begin{aligned} 4c_2 &= + \frac{1}{1!0!2^0} = 1 \\ c_2 &= \frac{1}{4}. \end{aligned}$$

For $n = 3$, we have

$$\begin{aligned} 9c_3 + c_1 &= 0 \\ c_3 &= 0. \end{aligned}$$

For $n = 4, k = 2$, we have

$$\begin{aligned} 16c_4 + c_2 &= -\frac{1}{2!1!2^2} = -\frac{1}{8} \\ c_4 &= -\frac{1}{16}\left(\frac{1}{4} + \frac{1}{8}\right) = -\frac{1}{64}\left(1 + \frac{1}{2}\right). \end{aligned}$$

This process may be continued indefinitely to generate coefficients. Clearly, all the odd numbered coefficients will end up being zero. The general rule for the even numbered coefficients is not at all obvious. It turns out to be

$$c_{2k} = (-1)^{k+1} \frac{1}{(k!)^2 2^{2k}} \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k}\right) \quad k \geq 1.$$

Hence, the second solution of Bessel's equation with $m = 0$ is

$$y_2 = J_0(t) \ln t + \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{(k!)^2} (1 + 1/2 + \cdots + 1/k) \left(\frac{t}{2}\right)^{2k}.$$

It is not hard to see that $\{y_1, y_2\}$ is a linearly independent pair of solutions.

It should be noted that one seldom needs to know the exact form of the second solution. The most important thing about it is that it is not continuous as $t \rightarrow 0$ because of the logarithmic term. The first solution $y_1 = J_0(t)$ is continuous. This has the following important consequence. Suppose we know on physical grounds that the solution is continuous and bounded as $t \rightarrow 0$. Then it follows that in

$$y = C_1 y_1(t) + C_2 y_2(t)$$

the coefficient of y_2 must vanish. For example, this must be the case for the vibrating membrane. The displacement must be bounded at the origin ($r = 0$), so we know the solution can involve only the Bessel function of the first kind.

The Case of Roots Differing by a Positive Integer Suppose the roots r_1, r_2 of the indicial equation satisfy $r_1 - r_2 = k$ where k is a positive integer. If $y_1(t)$ is a solution obtained by the method of Frobenius for $r = r_1$, the larger of the two roots, then the method of reduction of order yields a second solution of the form

$$y_2(t) = ay_1(t) \ln t + t^{r_2} \sum_{n=0}^{\infty} c_n t^n. \quad (156)$$

Note that the summation starts with $n = 0$. It is possible that the method of Frobenius works for the smaller root r_2 (because crucial terms vanish), and in that

case we would have $a = 0$ in (156); the series in the second term is what the method of Frobenius generates. Otherwise, $a \neq 0$, and, since we may always adjust a solution by a non-zero multiplicative constant, we may take $a = 1$. In any case the solution definitely exhibits a logarithmic singularity. In applying the method, you should first see if the method of Frobenius can be made to work for the lesser root. If it doesn't work (because the recursion at some point unavoidably yields a non-zero numerator and a zero denominator), then try a solution of the form (156) with $a = 1$. This will yield a set of recursion rules for the coefficients c_0, c_1, c_2, \dots roughly as in Example 186. To find these rules, you will have to evaluate some expression involving the series for the first solution $y_1(t)$ and its derivative $y_1'(t)$. See the Exercises for some examples.

Bessel's equation with $m > 0$ and $2m$ an integer illustrates the above principles. If m is half an odd integer, then the logarithmic term is missing and we may take $y_2 = J_{-m}(t)$. However, if m is a positive integer, then the logarithmic term is definitely present and we must take

$$y_2 = J_m(t) \ln t + t^{-m} \sum_{n=0}^{\infty} c_n t^n, \quad (157)$$

where the coefficients c_n are determined by the method outlined above. Such solutions are called *Bessel Functions of the second kind*. For technical reasons, it is sometimes better to add to (157) an appropriate multiple of $J_m(t)$ (a solution of the first kind). There is one particular family of such solutions called *Neumann functions* and denoted $Y_m(t)$. We won't study these in this course, but we mention them in case you encounter the notation. The most important thing about Neumann functions is that they have logarithmic singularities. It is possible to see in general that $\{y_1, y_2\}$ is a linearly independent pair of solutions, so the general solution of the differential equation has the form

$$y = C_1 y_1(t) + C_2 y_2(t).$$

If $r_1 > 0$, then $y_1(t) = t^{r_1} \sum_{n=0}^{\infty} a_n t^n$ is bounded as $t \rightarrow 0$. Usually, the second solution y_2 does not have this property. If the logarithmic term is present or if $r_2 < 0$, y_2 will not be bounded as $t \rightarrow 0$. In those cases, if we know the solution is bounded by physical considerations, we may conclude that $C_2 = 0$.

Using Reduction of Order for the Second Solution You might want to skip this section the first time through.

There are a variety of ways to show the second solution has the desired form in each of the 'bad' cases. We shall use the method of reduction of order. There is an alternate method based on solving the recursion relations in terms of r before setting r equal to either of the roots of the indicial equation. It is possible thereby to derive a set of formulas for the coefficients c_n . (See *Braun*, Section 2.8.3 for a discussion of this method.) However, no method gives a really practical approach

for finding the coefficients, so in most cases it is enough to know the form of the solution and then try to find the coefficients by substituting in the equation as we did above.

First, assume r is a double root of the indicial equation $f(r) = 0$ and $y_1 = t^r g(t)$ (where $g(t) = \sum_{n=0}^{\infty} a_n t^n$) is a solution obtained by the method of Frobenius. (Note that $g(t)$ is analytic at $t = 0$ and $g(0) \neq 0$.) The indicial equation has the form

$$f(r) = r(r-1) + p_0 r + q_0 = r^2 + (p_0 - 1)r + q_0 = 0$$

where $p(t) = \bar{p}(t)/t = (p_0 + p_1 t + \dots)/t$ and $q(t) = \bar{q}(t)/t^2 = (q_0 + q_1 t + \dots)/t^2$. Since r is double root, we have $(p_0 - 1)^2 - 4q_0 = 0$ and

$$r = -\frac{p_0 - 1}{2}.$$

The method of reduction of order tells us that there is a second solution of the form $y_2 = y_1 v$ where

$$v' = \frac{1}{y_1^2} e^{-\int p(t) dt}.$$

However, as above,

$$\begin{aligned} p(t) &= \frac{\bar{p}(t)}{t} = \frac{1}{t}(p_0 + p_1 t + p_2 t^2 + \dots) \\ &= \frac{p_0}{t} + p_1 + p_2 t + \dots + p_n t^{n-1} + \dots \end{aligned}$$

Hence,

$$\int p(t) dt = p_0 \ln t + p_1 t + \frac{p_2}{2} t^2 + \dots,$$

so

$$e^{-\int p(t) dt} = e^{-p_0 \ln t} e^{-p_1 t - \dots} = t^{-p_0} h_1(t)$$

where $h_1(t) = e^{-p_1 t - \dots}$ is an analytic function at $t = 0$ such that $h_1(0) \neq 0$. On the other hand

$$\frac{1}{y_1^2} = \frac{1}{t^{2r} g(t)^2} = t^{-2r} h_2(t)$$

where, since $g(0) \neq 0$, $h_2(t) = 1/g(t)^2$ is also analytic at $t = 0$ and $h_2(0) \neq 0$. It follows that

$$v' = t^{-2r} h_2(t) t^{-p_0} h_1(t) = t^{-(2r+p_0)} h_1(t) h_2(t).$$

However, $r = -(p_0 - 1)/2$, so $2r + p_0 = 1$. Moreover, $h(t) = h_1(t) h_2(t)$ is analytic and $h(0) \neq 0$, so it may be expanded in a power series

$$h(t) = h_0 + h_1 t + h_2 t^2 + \dots$$

with $h_0 \neq 0$. Thus,

$$\begin{aligned} v' &= t^{-1} (h_0 + h_1 t + h_2 t^2 + \dots) \\ &= \frac{h_0}{t} + h_1 + \frac{h_2}{t} + \dots \\ v &= h_0 \ln t + h_1 t + \frac{h_2}{2} t^2 + \dots \end{aligned}$$

It follows that

$$y_2 = y_1 v = h_0 y_1(t) \ln t + y_1(t) \left(h_1 t + \frac{h_2}{2} t^2 + \dots \right).$$

The solution generated by this process may always be modified by multiplying by a constant. Hence, since we know that $h_0 \neq 0$, we may multiply by its reciprocal, so in effect we may assume $h_0 = 1$. Moreover, since $y_1 = t^r g(t)$, the second term has the form

$$t^r g(t) \left(h_1 t + \frac{h_2}{2} t^2 + \dots \right).$$

Since $g(t)$ and $h_1 t + \frac{h_2}{2} t^2 + \dots$ are analytic, so is their product which can be expressed as a power series $\sum_{n=0}^{\infty} c_n t^n$. Moreover, since the ‘h series’ does not have a constant term, its sum vanishes at $t = 0$, and the same can be said for the product. That implies $c_0 = 0$, i.e., the summation may be assumed to start with $n = 1$. Thus we see that the solution obtained by reduction of order has the form

$$y_2 = y_1(t) \ln t + t^r \sum_{n=1}^{\infty} c_n t^n$$

as claimed.

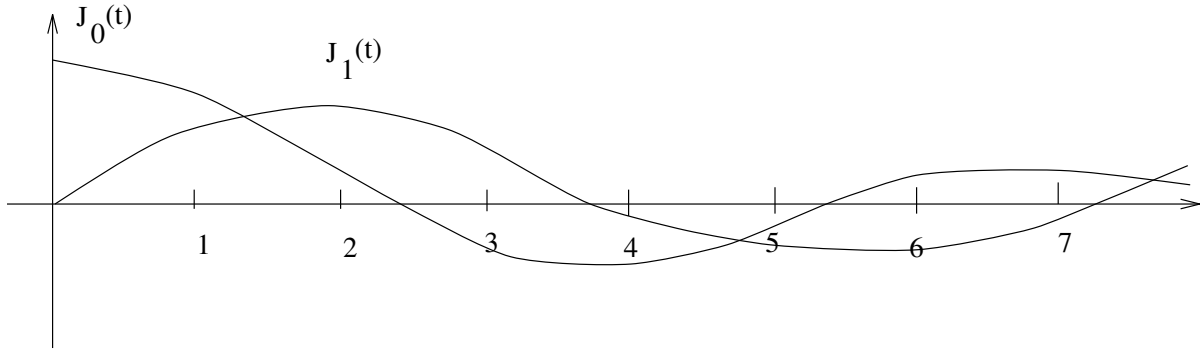
A very similar analysis of the process of reduction of order works in the case $r_1 - r_2$ is a positive integer.

Exercises for 9.5.

- Find a general solution of Laguerre’s Equation $ty'' + (1-t)y' + \lambda y = 0$ as follows.
 - Show that the indicial equation is $r^2 = 0$.
 - Find a solution for $r = 0$ by the method of Frobenius by assuming $a_0 = 1$. Note that this solution is analytic in general and is actually a polynomial if λ is a non-negative integer.
 - Set up the procedure for finding a second solution of the form $y = y_1 \ln t + \sum_{n=1}^{\infty} c_n t^n$. In the case $\lambda = 0$, find c_1 , a recurrence relation for c_n , the general value of c_n and the solution.
- Find a second solution of Bessel’s Equation for $m = 1$ by trying a solution of the form $y = J_1(t) \ln t + \sum_{n=0}^{\infty} c_n t^{n-1}$. You may not be able to determine a general formula for c_n but determine coefficients at least up to c_4 . You should discover that the recursion rule does not completely determine c_2 , so you can set it equal to anything you want. You should set it equal to zero for this problem. (However, it turns out that zero is not usually the best choice. By an appropriate non-zero choice, one gets the Neumann function $Y_1(t)$ discussed in the the text.)

9.6 More about Bessel Functions

We saw that $J_{1/2}(t) = \sqrt{2/\pi} \sin t/\sqrt{t}$ and $J_{-1/2}(t) = \sqrt{2/\pi} \cos t/\sqrt{t}$. In general, it turns out that the Bessel functions $J_m(t)$ all behave like sines and cosines with an ‘amplitude’ $1/\sqrt{t}$.



To see why this is the case, put $y = u/\sqrt{t}$ in the differential equation

$$t^2 y'' + t y' + (t^2 - m^2)y = 0.$$

We have

$$\begin{aligned} y' &= \frac{u'}{\sqrt{t}} - \frac{1}{2} \frac{u}{t^{3/2}} \\ y'' &= \frac{u''}{\sqrt{t}} - \frac{u'}{t^{3/2}} + \frac{3}{4} \frac{u}{t^{5/2}}. \end{aligned}$$

Hence,

$$\begin{aligned} t^2 y'' + t y' + (t^2 - m^2)y &= t^{3/2} u'' - t^{1/2} u' + \frac{3}{4} \frac{u}{t^{1/2}} + t^{1/2} u' - \frac{1}{2} \frac{u}{t^{1/2}} + t^{3/2} u - m^2 \frac{u}{t^{1/2}} \\ &= t^{3/2} u'' + t^{3/2} u + \left(\frac{1}{4} - m^2\right) \frac{u}{t^{1/2}}. \end{aligned}$$

Divide the last expression by $t^{3/2}$. We see thereby that if y is a solution of Bessel's equation, then $u = y\sqrt{t}$ satisfies the differential equation

$$u'' + u + \frac{1/4 - m^2}{t^2} u = 0. \quad (158)$$

Let $t \rightarrow \infty$. Then $\frac{1/4 - m^2}{t^2} \rightarrow 0$, so for large t , the equation is close to

$$u'' + u = 0. \quad (159)$$

The general solution of (159) is

$$u = C_1 \cos t + C_2 \sin t = A \cos(t + \delta).$$

Hence, it is *plausible* that for large t , the solution of (158) should look very similar. That means that for large t , any solution $y = u/\sqrt{t}$ of Bessel's equation should look like

$$y \approx A \frac{\cos(t + \delta)}{\sqrt{t}}.$$

(It is not clear in general that the limiting behavior of the solution of a differential equation is the same as a solution of the limit of the differential equation, but in this particular case, it happens to be true.)

Since $J_m(t) \approx A \frac{\cos t}{\sqrt{t}}$ for large t , we would expect it to oscillate with approximate period 2π . That means that successive roots of the equation

$$J_m(t) = 0$$

should differ roughly by π for large t . This is indeed the case. As mentioned earlier, the roots of this equation are important in determining the frequencies of the basic vibrations of a vibrating drum (and similar physical systems). The first root of the equation $J_0(t) = 0$ is approximately 2.4048.

Functions Related to Bessel Functions The differential equation 160

$$t^2 z'' + 2tz' + (t^2 - p^2)z = 0 \tag{160}$$

often arises in solving problems with spherical symmetry. (Note that the coefficient of z' is $2t$ rather than t .) This equation is related to Bessel's equation as follows. In

$$t^2 y'' + ty' + (t^2 - m^2)y = 0$$

put $y = \sqrt{t}z$. Then

$$\begin{aligned} y' &= \sqrt{t}z' + \frac{1}{2} \frac{z}{t^{1/2}} \\ y'' &= \sqrt{t}z'' + 2 \frac{1}{2} \frac{z'}{t^{1/2}} - \frac{1}{4} \frac{z}{t^{3/2}}. \end{aligned}$$

We obtain

$$t^{5/2}z'' + t^{3/2}z' - \frac{1}{4}t^{1/2}z + t^{3/2}z' + \frac{1}{2}t^{1/2}z + (t^2 - m^2)t^{1/2}z = 0.$$

If we divide through by $t^{1/2}$ and rearrange the terms, we obtain

$$t^2 z'' + 2tz' + (t^2 - m^2 + 1/4)z = 0.$$

If we put $p^2 = m^2 - 1/4$, then we obtain equation (160). It follows that solutions of (160) may be obtained by solving Bessel's equation with $m = \sqrt{p^2 + 1/4}$. Thus, Bessel functions for such m (where p is a non-negative integer) are often called *spherical Bessel functions*.

The solutions of the equation

$$z'' + tz' - (t^2 + m^2)z = 0$$

are called *modified Bessel functions*, and they also arise in important physical applications. (Note that the coefficient of z is $-t^2 - m^2$ rather than $+t^2 - m^2$.) There are two methods for obtaining solutions. First, apply the method of Frobenius. The process is almost identical to that in the case of the

ordinary Bessel functions. The indicial equation is also $r^2 - m^2 = 0$, and the only real difference is that the sign $(-1)^k$ does not occur. In particular, if m is a non-negative integer, a normalized solution obtained for the root $r = m$ is

$$I_m(t) = \sum_{k=0}^{\infty} \frac{1}{k!(m+k)!} \left(\frac{t}{2}\right)^{2k+m}.$$

For m a non-negative integer, a second solution is obtained for the root $r = -m$ by reduction of order and has a logarithmic singularity.

An alternate approach to the modified Bessel functions is to replace t by it in Bessel's equation where $i = \sqrt{-1}$. Then, we see that $I_m(t)$ and $J_m(it)$ just differ by a multiplicative constant. (Check that for yourself!) You may learn more about this in your complex variables course.

One could go on almost without end discussing properties of Bessel functions and of the other *special functions* of mathematical physics. We leave such pleasures for other courses in mathematics, physics, geophysics, etc.

Exercises for 9.6.

1. Find the second positive root of the equation $J_0(t) = 0$. If you want you can just look up the answer in a book, but you will have to figure out which book to look in. You might find it more illuminating to graph $J_0(t)$ using an appropriate graphics program such as Maple or Mathematica and try to zoom in on the root. Such a program will also give you a numerical value for the root if you ask it nicely.
2. $J_m(t)$ where $m^2 = p^2 + 1/4$ is a spherical Bessel function. Show that if $p^2 = k(k+1)$, then m is half an odd integer. That explains why Bessel functions of fractional order m are interesting.

3. (a) Apply the method of Frobenius to the differential equation

$$t^2 y'' + ty' - (t^2 + m^2)y = 0.$$

Find a first solution in the case m is a positive integer. Normalize it by putting $a_0 = 1/(2^m m!)$.

- (b) Compare your answer with $J_m(it)$ where $i^2 = -1$.

Part III

Linear Algebra

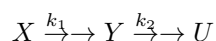
Chapter 10

Linear Algebra, Basic Notation

10.1 Systems of Differential Equations

Usually, in physical problems there are several variables, and the rate of change of each variable depends not only on it, but also on the other variables. This gives rise to a *system of differential equations*.

Example 187 Suppose that in a chemical reaction there are two substances X and Y that we need to keep track of. Suppose the *kinetics* of the reaction is such that X decomposes into Y at a rate proportional to the amount of X present, and suppose Y decomposes into uninteresting byproducts U at a rate proportional to the amount of Y present. We may indicate this schematically by



where k_1 and k_2 are rate constants. We may translate the above description into mathematics as follows. Let $x(t)$ and $y(t)$ denote the amounts of X and Y respectively present at time t . Then

$$\begin{aligned}\frac{dx}{dt} &= -k_1x \\ \frac{dy}{dt} &= k_1x - k_2y.\end{aligned}$$

This is simple example of a system of differential equations. It is not very hard to solve. The first equation involves only x , and its solution is $x = x_0e^{-k_1t}$ where $x_0 = x(0)$ is the amount of X present initially. This may be substituted in the

second equation to obtain

$$\frac{dy}{dt} + k_2y = k_1x = k_1x_0e^{-k_1t},$$

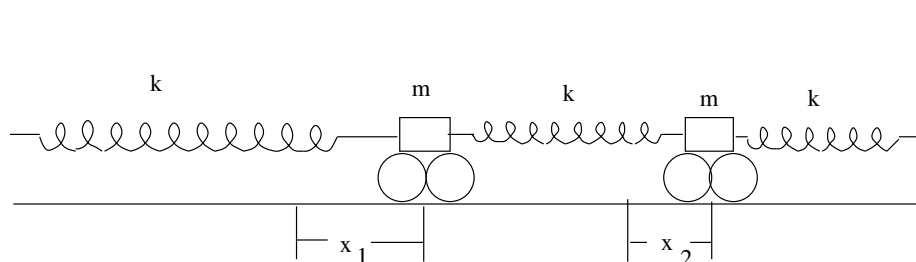
which is a first order linear equation which may be solved by the method in Chapter VI, Section 3.

Not every system is so easy to solve. Suppose for example that the substance Y in addition to decomposing into uninteresting substances also recombines with a substrate to form X at a rate depending on y . We would then have to modify the above system to one of the form

$$\begin{aligned}\frac{dx}{dt} &= -k_1x + k_3y \\ \frac{dy}{dt} &= k_1x - (k_2 + k_3)y.\end{aligned}$$

It is not immediately clear how to go about solving this system!

Example 188 Consider two identical masses m on a track connected by springs as indicated below. Suppose all the springs have the same spring constant k .



The masses will be at rest in certain equilibrium positions, but if they are displaced from those positions, the resulting system will oscillate in some very complicated way. Let x_1 and x_2 denote the displacements of the masses from equilibrium. The force exerted on the first mass by the spring to its left will be $-kx_1$ since that spring will be stretched by x_1 . On the other hand, the spring in the middle will be stretched by $x_1 - x_2$, so the force it exerts of the first mass is $-k(x_1 - x_2)$. Thus the total force on the first particle is the sum, so by Newton's Second Law

$$m \frac{d^2x_1}{dt^2} = -kx_1 - k(x_1 - x_2).$$

By a similar argument, we get for the second mass

$$m \frac{d^2x_2}{dt^2} = -kx_2 - k(x_2 - x_1).$$

(You should check both these relations to be sure you agree the forces are being exerted in the proper directions.) These equations may be simplified algebraically to yield

$$\begin{aligned} m \frac{d^2 x_1}{dt^2} &= -2kx_1 + kx_2 \\ m \frac{d^2 x_2}{dt^2} &= kx_1 - 2kx_2. \end{aligned} \quad (161)$$

Of course, to determine the motion completely, it is necessary to specify the initial positions $x_1(t_0), x_2(t_0)$ and the initial velocities $x'_1(t_0), x'_2(t_0)$ of *both* masses.

The above example is typical of many interesting physical systems. For example, a molecule consists of several atoms with attractive forces between them which to a first approximation may be treated mathematically as simple springs.

Higher Order Equations as Systems There is a simple trick which, by introducing new variables, allows us to reduce a differential equation of *any order* to a *first order system*.

Example 189 Consider the differential equation

$$y''' + 2y'' + 3y' - 4y = e^t. \quad (162)$$

There is an elaborate theory of such equations which generalizes what we did in Chapter VII for second order equations. However, there is another approach which replaces the equation by a *first order system*. Introduce new variables as follows:

$$\begin{aligned} x_1 &= y \\ x_2 &= y' = x'_1 \\ x_3 &= y'' = x'_2. \end{aligned}$$

From the differential equation,

$$x'_3 = y''' = 4y - 3y' - 2y'' + e^t.$$

Hence, we may replace the single equation (162) by the system

$$\begin{aligned} x'_1 &= x_2 \\ x'_2 &= x_3 \\ x'_3 &= 4x_1 - 3x_2 - 2x_3 + e^t. \end{aligned}$$

The same analysis may in fact be applied to any system of any order in any number of variables.

Example 188, revisited Introduce additional variables $x_3 = x'_1$ and $x_4 = x'_2$. Then from (??), we have

$$\begin{aligned}x'_3 &= x''_1 = -\frac{2k}{m}x_1 + \frac{k}{m}x_2 \\x'_4 &= x''_2 = \frac{k}{m}x_2 - \frac{2k}{m}x_2.\end{aligned}$$

Putting this all together yields the first order system in 4 variables

$$\begin{aligned}x'_1 &= x_3 \\x'_2 &= x_4 \\x'_3 &= -\frac{2k}{m}x_1 + \frac{k}{m}x_2 \\x'_4 &= \frac{k}{m}x_2 - \frac{2k}{m}x_2.\end{aligned}$$

To completely specify the solution, we need (in the new notation) to specify the 4 initial values $x_1(0), x_2(0), x_3(0) = x'_1(0)$ and $x_4(0) = x'_2(0)$.

By similar reasoning, any system may be reduced to a first order system involving variables x_1, x_2, \dots, x_n each of which is a function of the independent variable t . Using the notation of \mathbf{R}^n , such a collection of variables may be combined in a single *vector* variable

$$\mathbf{x} = \langle x_1, x_2, \dots, x_n \rangle$$

which is assumed to be a vector valued function $\mathbf{x}(t)$ of t . For each component x_i , its derivative is supposed to be a function

$$\frac{dx_i}{dt} = f_i(x_1, x_2, \dots, x_n, t) = f_i(\mathbf{x}, t).$$

We may summarize this in a single vector equation

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, t)$$

where the vector valued function \mathbf{f} has components the scalar functions f_i for $i = 1, 2, \dots, n$.

The most important special case is that of *linear systems*. In this case the component functions have the special form

$$f_i(x_1, x_2, \dots, x_n, t) = a_{i1}(t)x_1 + a_{i2}(t)x_2 + \dots + a_{in}(t)x_n + g_i(t)$$

for $i = 1, 2, \dots, n$. That is, each component function depends *linearly* on the dependent variables x_1, x_2, \dots, x_n with coefficients $a_{ij}(t)$ which as indicated may depend on t . In this case, the system $\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, t)$ may be written out in ‘longhand’

as

$$\begin{aligned}\frac{dx_1}{dt} &= a_{11}(t)x_1 + a_{12}(t)x_2 + \cdots + a_{1n}(t)x_n + g_1(t) \\ \frac{dx_2}{dt} &= a_{21}(t)x_1 + a_{22}(t)x_2 + \cdots + a_{2n}(t)x_n + g_2(t) \\ &\vdots \\ \frac{dx_n}{dt} &= a_{n1}(t)x_1 + a_{n2}(t)x_2 + \cdots + a_{nn}(t)x_n + g_n(t).\end{aligned}$$

You should compare this with each of the examples discussed in this section. You should see that they are all linear systems.

The above notation is quite cumbersome, and clearly it will be easier to follow if we use the notation of vectors in \mathbf{R}^n as above. To exploit this fully for linear systems we need additional notation and concepts. To this end, we shall study some *linear algebra*, one purpose of which is to make higher dimensional problems look one-dimensional. In the next sections we shall introduce notation so that the above *linear system* may be rewritten

$$\frac{d\mathbf{x}}{dt} = A(t)\mathbf{x} + \mathbf{g}(t).$$

With this notation, much of what we discovered about first order equations in Chapter VI will apply to systems.

Exercises for 10.1.

1. Solve the system

$$\begin{aligned}\frac{dx}{dt} &= -k_1x \\ \frac{dy}{dt} &= k_1x - k_2y\end{aligned}$$

as suggested in the text. The answer should be expressed in terms of $x_0 = x(0)$ and $y_0 = y(0)$.

2. In each of the following cases, reduce the given equation(s) to an appropriate first order system.
 - (a) $y''' + 2y'' - 3y' + 2y = \cos t$.
 - (b) $y''' - 2(y')^2 + y = 0$.
 - (c) $y_1'' = 3y_1 - 2y_2$, $y_2'' = -2y_1 + 4y_2$.
 - (d) $x_1'' + (x_1')^2 + x_1x_2 = 0$, $x_2'' + x_1 - x_2 = 0$.
3. In the previous problem, determine if the first order system you obtained in each part is linear.

10.2 Matrix Algebra

A rectangular array

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

is called an $m \times n$ *matrix*. It has m *rows* and n *columns*. The quantities a_{ij} are called the *entries* of the matrix. The first index i tells you which *row* it is in, and the second index j tells you which *column* it is in.

Examples

$$\begin{array}{ll} \begin{bmatrix} -2k & k \\ k & -2k \end{bmatrix} & \text{is a } 2 \times 2 \text{ matrix} \\ \begin{bmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{bmatrix} & \text{is a } 2 \times 4 \text{ matrix} \\ \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} & \text{is a } 1 \times 4 \text{ matrix} \\ \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} & \text{is a } 4 \times 1 \text{ matrix} \end{array}$$

Matrices of various sizes and shapes arise in many situations. For example, the first matrix listed above is the *matrix of coefficients* on the right hand side of the system

$$\begin{aligned} m \frac{d^2 x_1}{dt^2} &= -2kx_1 + kx_2 \\ m \frac{d^2 x_2}{dt^2} &= kx_1 - 2kx_2. \end{aligned}$$

in Example 188 of the previous section.

In computer programming, a matrix is called a *2-dimensional array* and the entry in row i and column j is usually denoted $a[i, j]$ instead of a_{ij} . As in programming, it is useful to think of the entire array as a single entity, so we use a single letter to denote it

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

There are various different special arrangements which play important roles. A matrix with the same number of rows as columns is called a *square* matrix. Matrices

of coefficients for linear systems of differential equations are usually square. A 1×1 matrix

$$\begin{bmatrix} a \end{bmatrix}$$

is not logically distinguishable from a *scalar*, so we make no distinction between the two concepts. A matrix with one row

$$\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}$$

is called a *row vector* and a matrix with one column

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

is called a *column vector*. Logically, either a $1 \times n$ row vector or an $n \times 1$ column with real entries is just an n -tuple, i.e., an element of \mathbf{R}^n . However, as we shall see, operations with row vectors are sometimes different than with column vectors. We may identify either the set of all row vectors with real entries *or* the set of all column vectors with real entries with the set \mathbf{R}^n . For reasons that will become clear shortly, we shall usually make the latter choice. That is, we shall ordinarily think of an element of \mathbf{R}^n as a column vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

More generally, it should be noted that the information contained in any $m \times n$ matrix has two parts. There are the mn entries which, in the real case, specify, in some order, an element of \mathbf{R}^{mn} , and there is also the arrangement of the entries in rows and columns.

Matrices are denoted in different ways by different authors. Most people use ordinary (non-boldface) capital letters, e.g., A, B, X, Q . However, one sometimes wants to use boldface for row or column vectors, as above, when the relationship to \mathbf{R}^n is being emphasized. One may also use lower case non-boldface letters for row vectors or column vectors. Since there is no consistent rule about this, you should make sure you know when a symbol represents a matrix which is not a scalar.

Matrices may be combined in various useful ways. Two matrices *of the same size and shape* are added by adding corresponding entries. You are not allowed to add matrices with different shapes.

Examples

$$\begin{bmatrix} 1 & -1 \\ 2 & 1 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 0 & 3 \\ -1 & -2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 2 & 4 \\ -1 & -1 \end{bmatrix}$$

$$\begin{bmatrix} x+y \\ y \\ 0 \end{bmatrix} + \begin{bmatrix} -y \\ -y \\ x \end{bmatrix} = \begin{bmatrix} x \\ 0 \\ x \end{bmatrix}.$$

The $m \times n$ matrix with zero entries is called a *zero matrix* and is usually just denoted 0. Since zero matrices with different shapes are not the same, it is sometimes necessary to indicate the shape by using subscripts, as in ‘ 0_{mn} ’, but usually the context makes it clear which zero matrix is needed. The zero matrix of a given shape has the property that if you add it to any matrix A of the same shape, you get the A again as the result.

A matrix may also be multiplied by a scalar by multiplying each entry of the matrix by that scalar. More generally, we may multiply several matrices with the same shape by different scalars and add up the result:

$$c_1 A_1 + c_2 A_2 + \cdots + c_k A_k$$

where c_1, c_2, \dots, c_k are scalars and A_1, A_2, \dots, A_k are $m \times n$ matrices with the same m and n . This process is called *linear combination*.

Example

$$2 \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} + (-1) \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} + 3 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 2 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \\ 0 \\ -1 \end{bmatrix} + \begin{bmatrix} 3 \\ 3 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \\ 5 \\ 2 \end{bmatrix}.$$

Sometimes it is convenient to put the scalar on the other side of the matrix, but the meaning is the same: each entry of the matrix is multiplied by the scalar.

$$cA = Ac.$$

We shall also have occasion to consider matrix valued *functions* $A(t)$ of a scalar variable t . That means that each entry $a_{ij}(t)$ is a function of t . Such functions are differentiated or integrated entry by entry.

Examples

$$\frac{d}{dt} \begin{bmatrix} e^{2t} & e^{-t} \\ 2e^{2t} & -e^{-t} \end{bmatrix} = \begin{bmatrix} 2e^{2t} & -e^{-t} \\ 4e^{2t} & e^{-t} \end{bmatrix}$$

$$\int_0^1 \begin{bmatrix} t \\ t^2 \end{bmatrix} dt = \begin{bmatrix} 1/2 \\ 1/3 \end{bmatrix}$$

There are various ways to *multiply* matrices. For example, one sometimes multiplies matrices of the same shape by multiplying corresponding entries. This is useful only in very special circumstances. Another kind of multiplication generalizes the *dot product* of vectors. If

$$\begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}$$

is a row vector of size n , and

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

is a column vector of the same size n , the row by column product is defined to be the sum of the products of corresponding entries

$$\begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = a_1b_1 + a_2b_2 + \dots + a_nb_n = \sum_{i=1}^n a_ib_i.$$

This product is of course a *scalar*, and except for the distinction between row and column vectors, it is the same as the notion of dot product for elements of \mathbf{R}^n introduced in Chapter I, Section 3. You should be familiar with its properties.

More generally, let A be an $m \times n$ matrix and B an $n \times p$ matrix. Then each row of A has the same size as each column of B . The *matrix product* AB is defined to be the $m \times p$ matrix with i, j entry the row by column product of the i th row of A with the j th column of B . Thus, if $C = AB$, then C has the same number of rows as A , the same number of columns as B , and

$$c_{ij} = \sum_{r=1}^n a_{ir}b_{rj}.$$

Examples

$$\begin{aligned}
\underbrace{\begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix}}_{2 \times 2} \underbrace{\begin{bmatrix} 1 & 0 & 1 \\ -1 & 2 & 1 \end{bmatrix}}_{2 \times 3} &= \underbrace{\begin{bmatrix} 2-1 & 0+2 & 2+1 \\ 1-0 & 0+0 & 1+0 \end{bmatrix}}_{2 \times 3} \\
&= \begin{bmatrix} 1 & 2 & 3 \\ 1 & 0 & 1 \end{bmatrix} \\
\underbrace{\begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 2 & 1 \end{bmatrix}}_{3 \times 2} \underbrace{\begin{bmatrix} x \\ y \end{bmatrix}}_{2 \times 1} &= \underbrace{\begin{bmatrix} x-y \\ x \\ 2x+y \end{bmatrix}}_{3 \times 1}
\end{aligned}$$

The most immediate use for matrix multiplication is a simplification of the notation used to describe a linear system. We have

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \end{bmatrix}.$$

(Note that the matrix on the right is an $n \times 1$ column vector and each entry, although expressed as a complicated sum, is a scalar.) With this notation, the linear system

$$\begin{aligned}
\frac{dx_1}{dt} &= a_{11}(t)x_1 + a_{12}(t)x_2 + \cdots + a_{1n}(t)x_n + g_1(t) \\
\frac{dx_2}{dt} &= a_{21}(t)x_1 + a_{22}(t)x_2 + \cdots + a_{2n}(t)x_n + g_2(t) \\
&\vdots \\
\frac{dx_n}{dt} &= a_{n1}(t)x_1 + a_{n2}(t)x_2 + \cdots + a_{nn}(t)x_n + g_n(t).
\end{aligned}$$

may be rewritten

$$\frac{d\mathbf{x}}{dt} = A(t)\mathbf{x} + \mathbf{g}(t)$$

where

$$\mathbf{x} = \mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix}$$

$$A(t) = \begin{bmatrix} a_{11}(t) & a_{12}(t) & \cdots & a_{1n}(t) \\ a_{21}(t) & a_{22}(t) & \cdots & a_{2n}(t) \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1}(t) & a_{n2}(t) & \cdots & a_{nn}(t) \end{bmatrix}$$

$$\mathbf{g}(t) = \begin{bmatrix} g_1(t) \\ g_2(t) \\ \vdots \\ g_n(t) \end{bmatrix}$$

Example 190 The system

$$\begin{aligned} x_1' &= x_2 \\ x_2' &= x_3 \\ x_3' &= 4x_1 - 3x_2 - 2x_3 + e^t. \end{aligned}$$

may be rewritten

$$\frac{d}{dt} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 4 & -3 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ e^t \end{bmatrix}.$$

Another important application of matrices occurs in the analysis of large systems of simultaneous linear algebraic equations. We shall have much more to say about this later in this chapter. In addition, matrices and linear algebra are used extensively in practically every branch of science and engineering.

Exercises for 10.2.

1. Let

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ -3 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} -2 \\ 1 \\ 3 \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}.$$

Calculate $\mathbf{x} + \mathbf{y}$ and $3\mathbf{x} - 5\mathbf{y} + \mathbf{z}$.

2. Let

$$A = \begin{bmatrix} 2 & 7 & 4 & -3 \\ -3 & 0 & 1 & -2 \\ 1 & 3 & -2 & 3 \\ 0 & 0 & 5 & -5 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} 1 \\ -2 \\ 3 \\ 5 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} -2 \\ 2 \\ 0 \\ 4 \end{bmatrix}.$$

Compute $A\mathbf{x}$, $A\mathbf{y}$, $A\mathbf{x} + A\mathbf{y}$, and $A(\mathbf{x} + \mathbf{y})$.

3. Let

$$A = \begin{bmatrix} 1 & -1 & 3 \\ 0 & -2 & 2 \end{bmatrix}, B = \begin{bmatrix} 1 & 2 \\ 1 & 0 \\ -3 & 2 \end{bmatrix}, C = \begin{bmatrix} -1 & 1 & -3 \\ 0 & 2 & -2 \end{bmatrix}, D = \begin{bmatrix} -1 & -2 & 0 \\ 1 & -2 & 1 \\ 2 & 1 & -4 \end{bmatrix}.$$

Calculate each of the following quantities *if it is defined*: $A + 3B$, $A + C$, $C + 2D$, AB , BA , CD , DC .

4. Suppose
- A
- is a
- 2×2
- matrix such that

$$A \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \quad A \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 4 \end{bmatrix}.$$

Find A .

5. Let
- \mathbf{e}_i
- denote the
- $n \times 1$
- column vector, with all entries zero except the
- i
- th which is 1, e.g., for
- $n = 3$
- ,

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Let A be an arbitrary $m \times n$ matrix. Show that $A\mathbf{e}_i$ is the i th column of A . You may verify this just in the case $n = 3$ and A is 3×3 . That is sufficiently general to understand the general argument.

6. Write each of the following systems in matrix form.

$$(a) \ x'_1 = 2x_1 - 3x_2, \ x'_2 = -4x_1 + 2x_2.$$

$$(b) \ x'_1 = 2x_1 - 3x_2 + 4, \ x'_2 = -4x_1 + 2x_2 - 1.$$

$$(c) \ x'_1 = x_2, \ x'_2 = x_3, \ x'_3 = 2x_1 + 3x_2 - x_3 + \cos t.$$

7. Let
- $y(t)$
- be a solution of the linear second order differential equation

$$t^2 y'' + t\bar{p}(t)y' + \bar{q}(t)y = 0.$$

Put

$$\mathbf{x} = \mathbf{x}(t) = \begin{bmatrix} y(t) \\ ty'(t) \end{bmatrix}$$

and show that \mathbf{x} satisfies the matrix differential equation

$$t \frac{d\mathbf{x}}{dt} = \begin{bmatrix} 0 & 1 \\ -\bar{q}(t) & 1 - \bar{p}(t) \end{bmatrix} \mathbf{x}.$$

8. Matrices are used to create more realistic population models than those we considered in Chapter VI, Section 4. First, divide the population into n age groups for an appropriate positive integer n . Let x_i , $i = 1, 2, \dots, n$ be the number of women in the i th age group, and consider the vector \mathbf{x} with those components. Construct an $n \times n$ matrix A which incorporates information about birth and death rates so that $A\mathbf{x}$ gives the population vector after one unit of time has elapsed. Then $A^n\mathbf{x}$ gives the population vector after n units of time.

Assume a human population is divided into 10 age groups between 0 and 99. Suppose the following table gives the birth and death rates for each age group

Age	BR	DR
0...9	0	.01
10...19	.01	.01
20...29	.04	.01
30...39	.03	.01
40...49	.01	.02
50...59	.001	.03
60...69	0	.04
70...79	0	.10
80...89	0	.30
90...99	0	1.00

Find A .

10.3 Formal Rules

The *usual rules of algebra* apply to matrices with a few exceptions. Here are *some* of these rules and warnings about when they apply.

The *associative law*

$$A(BC) = (AB)C$$

works as long as the shapes of the matrices match. That means that the length of each row of A must be the same as the length of each column of B and the length of each row of B must be the same as the length of each column of C . Otherwise, none of the products in the formula will be defined. The proof of the associative law requires some fiddling with indices and is left for the Exercises.

For each positive integer n , the $n \times n$ matrix

$$I = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

is called the *identity matrix* of degree n . As in the case of the zero matrices, we get a different identity matrix for each n , and if we need to note the dependence on n , we shall use the notation I_n . The identity matrix of degree n has the property $IA = A$ for any matrix A with n rows and the property $BI = B$ for any matrix B with n columns. The entries of the identity matrix are usually denoted δ_{ij} . $\delta_{ij} = 1$ if $i = j$ (the *diagonal entries*) and $\delta_{ij} = 0$ if $i \neq j$. The indexed expression δ_{ij} is often called the *Kronecker δ* .

The *commutative law* $AB = BA$ is *not generally true* for matrix multiplication. First of all, the products won't be defined unless the shapes match. Even if the shapes match on both sides, the resulting products may have different sizes. Thus, if A is $m \times n$ and B is $n \times m$, then AB is $m \times m$ and BA is $n \times n$. Finally, even if the shapes match and the products have the same sizes (if both A and B are $n \times n$), it may still be true that the products are different.

Example 191 Suppose

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

Then

$$AB = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = 0 \quad BA = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \neq 0$$

so $AB \neq BA$. Lest you think that this is a specially concocted example, let me assure you that it is the exception rather than the rule for the commutative law to hold for a randomly chosen pair of square matrices.

Another rule of algebra which holds for scalars but *does not generally hold* for matrices is the *cancellation law*.

Example 192 Let

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \quad C = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then

$$AB = 0 \quad \text{and} \quad AC = 0$$

so we cannot necessarily conclude from $AB = AC$ that $B = C$.

The *distributive laws*

$$\begin{aligned} A(B + C) &= AB + AC \\ (A + B)C &= AC + BC \end{aligned}$$

do hold as long as the operations are defined. Note however that since the commutative law does not hold in general, the distributive law must be stated for both possible orders of multiplication.

Another useful rule is

$$c(AB) = (cA)B = A(cB)$$

where c is a scalar and A and B are matrices whose shapes match so the products are defined.

The rules of calculus apply in general to matrix valued functions except that you have to be careful about orders whenever products are involved. For example, we have

$$\frac{d}{dt}(A(t)B(t)) = \frac{dA(t)}{dt}B(t) + A(t)\frac{dB(t)}{dt}$$

for matrix valued functions $A(t)$ and $B(t)$ with matching shapes.

We have just listed *some* of the rules of algebra and calculus, and we haven't discussed any of the proofs. Generally, you can be confident that matrices can be manipulated like scalars if you are careful about matters like commutativity discussed above. However, in any given case, if things don't seem to be working properly, you should look carefully to see if some operation you are using is valid for matrices.

Exercises for 10.3.

1. Find two 2×2 matrices A and B such that neither has *any* zero entries but such that $AB = 0$.
2. Let A be an $m \times n$ matrix, let \mathbf{x} and \mathbf{y} be $n \times 1$ column vectors, and let a and b be scalars. Using the rules of algebra discussed in Section 3, prove

$$A(a\mathbf{x} + b\mathbf{y}) = a(A\mathbf{x}) + b(A\mathbf{y}).$$

3. Prove the associative law $(AB)C = A(BC)$. Hint: If $D = AB$, then $d_{ik} = \sum_{j=1}^n a_{ij}b_{jk}$, and if $E = BC$ then $e_{jr} = \sum_{k=1}^p b_{jk}c_{kr}$, where A is $m \times n$, B is $n \times p$, and C is $p \times q$.

10.4 Linear Systems of Algebraic Equations

Before studying the problem of solving a linear system of differential equations, we tackle the simpler problem of solving a linear system of simultaneous algebraic equations. This problem is important in its own right, and, as we shall see, we need to be able to solve linear algebraic systems in order to be able to solve linear systems of differential equations.

We start with a problem you ought to be able to solve from what you learned in high school

Example 193 Consider the algebraic system

$$\begin{aligned}x_1 + 2x_2 - x_3 &= 1 \\x_1 - x_2 + x_3 &= 0 \\x_1 + x_2 + 2x_3 &= 1\end{aligned}\tag{163}$$

which is a system of 3 equations in 3 unknowns x_1, x_2, x_3 . This system may also be written more compactly as a matrix equation

$$\begin{bmatrix} 1 & 2 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

The method we shall use to solve (163) is the method of *elimination* of unknowns. Subtract the first equation from each of the other equations to eliminate x_1 from those equations.

$$\begin{aligned}x_1 + 2x_2 - x_3 &= 1 \\-3x_2 + 2x_3 &= -1 \\-x_2 + 3x_3 &= 0\end{aligned}$$

Now subtract 3 times the third equation from the second equation.

$$\begin{aligned}x_1 + 2x_2 - x_3 &= 1 \\-7x_3 &= -1 \\-x_2 + 3x_3 &= 0\end{aligned}$$

which may be reordered to obtain

$$\begin{aligned}x_1 + 2x_2 - x_3 &= 1 \\-x_2 + 3x_3 &= 0 \\7x_3 &= 1.\end{aligned}$$

We may now solve as follows. According to the last equation $x_3 = 1/7$. Putting this in the second equation yields

$$-x_2 + 3/7 = 0 \quad \text{or} \quad x_2 = 3/7.$$

Putting $x_3 = 1/7$ and $x_2 = 3/7$ in the first equation yields

$$x_1 + 2(3/7) - 1/7 = 1 \quad \text{or} \quad x_1 = 1 - 5/7 = 2/7.$$

Hence, we get

$$\begin{aligned}x_1 &= 2/7 \\x_2 &= 3/7 \\x_3 &= 1/7\end{aligned}$$

To check, we calculate

$$\begin{bmatrix} 1 & 2 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 2/7 \\ 3/7 \\ 1/7 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

The above example illustrates the general procedure which may be applied to any system of m equations in n unknowns

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

or, using matrix notation,

$$A\mathbf{x} = \mathbf{b}$$

with

$$\begin{aligned} A &= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \\ \mathbf{x} &= \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \\ \mathbf{b} &= \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}. \end{aligned}$$

As in Example 193, a sequence of elimination steps yields a set of equations each involving at least one fewer unknowns than the one above it. This process is called *Gaussian reduction* after the famous 19th century German mathematician C. F. Gauss. To complete the solution, we start with the last equation and substitute back recursively in each of the previous equations. This process is called appropriately *back-substitution*. The combined process will generally lead to a complete solution, but, as we shall see later, there can be some difficulties.

Row Operations and Gauss-Jordan reduction We now consider the general process of solving a system of equations of the form

$$AX = B$$

where A is an $n \times n$ matrix, X is an $n \times p$ matrix of *unknowns*, and B is an $n \times p$ matrix of known quantities. Usually, p will be 1, so X and B will be column vectors, but the procedure is basically the same for any p . For the moment we emphasize the case in which the coefficient matrix A is *square*, but we shall return later to the general case (m and n possibly different).

If you look carefully at Example 193, you will see that we employed three basic types of operations:

1. adding or subtracting a multiple of one equation from another,
2. multiplying or dividing an equation by a non-zero scalar,
3. interchanging two equations.

These operations correspond when using matrix notation to applying the following operations to the matrices on both sides of the equation $AX = B$:

1. adding or subtracting one row of a matrix to another,
2. multiplying or dividing one row of a matrix by a non-zero scalar,
3. interchanging two rows of a matrix.

These operations are called *elementary row operations*.

An important principle about row operations that we shall use over and over again is the following: *To apply a row operation to a product AX , it suffices to apply the row operation to A and then to multiply the result by X .* It is easy to convince yourself that this rule is valid by looking at examples. Thus, for the product,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix}$$

adding the first row to the second yields

$$\begin{bmatrix} ax + by \\ ax + by + cx + dy \end{bmatrix} = \begin{bmatrix} ax + by \\ (a + c)x + (b + d)y \end{bmatrix}.$$

On the other hand, adding the rows of the coefficient matrix yields

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \rightarrow \begin{bmatrix} a & b \\ a + c & b + d \end{bmatrix},$$

and multiplying the transformed matrix by $\begin{bmatrix} x \\ y \end{bmatrix}$ yields

$$\begin{bmatrix} ax + by \\ (a + c)x + (b + d)y \end{bmatrix}$$

as required. (See the appendix to this section for a general proof.)

It is now clear how to proceed in general to solve a system of the form

$$AX = B.$$

Apply row operations to both sides until we obtain a system which is easy to solve (or for which it is clear there is no solution.) Because of the principle just enunciated, we may apply the row operations on the left just to the matrix A and omit reference to X since that is not changed. For this reason, it is usual to collect A on the left and B on the right in a so-called *augmented matrix*

$$[A | B]$$

where the ‘|’ (or other appropriate divider) separates the two matrices. We illustrate this by considering another system of 3 equations in 3 unknowns.

Example 194

$$\begin{aligned} x_1 + x_2 - x_3 &= 0 \\ 2x_1 \quad \quad + x_3 &= 2 \\ x_1 - x_2 + 3x_3 &= 1 \end{aligned}$$

or

$$\begin{bmatrix} 1 & 1 & -1 \\ 2 & 0 & 1 \\ 1 & -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$$

We first do the Gaussian part of the reduction but for the augmented matrix rather than the original set of equations.

$$\begin{aligned} \left[\begin{array}{ccc|c} 1 & 1 & -1 & 0 \\ 2 & 0 & 1 & 2 \\ 1 & -1 & 3 & 1 \end{array} \right] &\rightarrow \left[\begin{array}{ccc|c} 1 & 1 & -1 & 0 \\ 0 & -2 & 3 & 2 \\ 1 & -1 & 3 & 1 \end{array} \right] & -2[r1] + r2 \\ &\rightarrow \left[\begin{array}{ccc|c} 1 & 1 & -1 & 0 \\ 0 & -2 & 3 & 2 \\ 0 & -2 & 4 & 1 \end{array} \right] & -[r1] + r3 \\ &\rightarrow \left[\begin{array}{ccc|c} 1 & 1 & -1 & 0 \\ 0 & -2 & 3 & 2 \\ 0 & 0 & 1 & -1 \end{array} \right] & -[r2] + r3 \end{aligned}$$

At this point the corresponding system is

$$\begin{aligned} x_1 + x_2 - x_3 &= 0 \\ -2x_2 + 3x_3 &= 2 \\ x_3 &= -1 \end{aligned}$$

so we could now apply back-substitution to find the solution. However, it is better for matrix computation to use an essentially equivalent process. Starting with the last row, use the leading non-zero entry to eliminate the entries above it. (That corresponds to substituting the value of the corresponding unknown in the previous equations.) This process is called *Jordan reduction*.

$$\begin{aligned}
 \left[\begin{array}{ccc|c} 1 & 1 & -1 & 0 \\ 0 & -2 & 3 & 2 \\ 0 & 0 & 1 & -1 \end{array} \right] &\rightarrow \left[\begin{array}{ccc|c} 1 & 1 & -1 & 0 \\ 0 & -2 & 0 & 5 \\ 0 & 0 & 1 & -1 \end{array} \right] && -3[r3] + r2 \\
 &\rightarrow \left[\begin{array}{ccc|c} 1 & 1 & 0 & -1 \\ 0 & -2 & 0 & 5 \\ 0 & 0 & 1 & -1 \end{array} \right] && [r3] + r1 \\
 &\rightarrow \left[\begin{array}{ccc|c} 1 & 1 & 0 & -1 \\ 0 & 1 & 0 & -5/2 \\ 0 & 0 & 1 & -1 \end{array} \right] && -(1/2)[r2] \\
 &\rightarrow \left[\begin{array}{ccc|c} 1 & 0 & 0 & -3/2 \\ 0 & 1 & 0 & -5/2 \\ 0 & 0 & 1 & -1 \end{array} \right] && -[r2] + r1
 \end{aligned}$$

This corresponds to the system

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} X = \begin{bmatrix} -3/2 \\ -5/2 \\ -1 \end{bmatrix} \quad \text{or} \quad X = \begin{bmatrix} -3/2 \\ -5/2 \\ -1 \end{bmatrix}$$

which is the desired solution: $x_1 = -3/2, x_2 = -5/2, x_3 = -1$. (Check it by plugging back into the original matrix equation.)

The combined method employed in the previous example is called *Gauss-Jordan reduction*. The strategy is clear. Use a sequence of row operations to reduce the coefficient matrix A to the identity matrix I . *If this is possible*, the same sequence of row operations will transform the matrix B to a new matrix B' , and the corresponding matrix equation will be

$$IX = B' \quad \text{or} \quad X = B'.$$

It is natural at this point to conclude that $X = B'$ is the solution of the original system, but there is a subtlety involved here. The method outlined above shows the following: *if there is a solution*, and if it is possible to reduce A to I by a sequence of row operations, then the solution is $X = B'$. In essence, this says that if the solution exists, then it is unique. It does not demonstrate that any solution exists. Why are we justified in concluding that we do in fact have a solution when the reduction is possible? To understand that, first note that *every possible row operation is reversible*. Thus, to reverse the effect of adding a multiple of one row to another, just subtract the same multiple of the first row from the (modified) second row. To reverse the effect of multiplying a row by a non-zero scalar, just

multiply the (modified) row by the reciprocal of that scalar. Finally, to reverse the effect of interchanging two rows, just interchange them back. Hence, the effect of any sequence of row operations on a system of equations is to produce an *equivalent* system of equations. Anything which is a solution of the initial system is necessarily a solution of the transformed system and vice-versa. Thus, the system $AX = B$ is equivalent to the system $X = IX = B'$, which is to say $X = B'$ is a solution of $AX = B$.

Elementary Matrices and the Effect of Row Operations on Products

Each of the elementary row operations may be accomplished by multiplying by an appropriate square matrix on the left. Such matrices of course should have the proper size for the matrix being multiplied.

To add c times the j th row of a matrix to the i th row (with $i \neq j$), multiply that matrix on the left by the matrix $E_{ij}(c)$ which has diagonal entries 1, the i, j -entry c , and all other entries 0. This matrix may also be obtained by applying the specified row operation to the identity matrix. You should try out a few examples to convince yourself that it works.

Example For $n = 3$,

$$E_{13}(-4) = \begin{bmatrix} 1 & 0 & -4 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

To multiply the i th row of a matrix by $c \neq 0$, multiply that matrix on the left by the matrix $E_i(c)$ which has diagonal entries 1 except for the i, i -entry which is c and which has all other entries zero. $E_i(c)$ may also be obtained by multiplying the i th row of the identity matrix by c .

Example For $n = 3$,

$$E_2(6) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

To interchange the i th and j th rows of a matrix, with $i \neq j$, multiply by the matrix on the left by the matrix E_{ij} which is obtained from the identity matrix by interchanging its i th and j th rows. The diagonal entries of E_{ij} are 1 except for its i, i , and j, j -entries which are zero. Its i, j and j, i -entries are both 1, and all other entries are zero.

Examples For $n = 3$,

$$E_{12} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad E_{13} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

Matrices of the above type are called *elementary matrices*.

The fact that row operations may be accomplished by matrix multiplication by elementary matrices has many important consequences. Thus, let E be an elementary matrix corresponding to a certain elementary row operation. The associative law tells us

$$E(AB) = (EA)B$$

as long as the shapes match. However, $E(AB)$ is the result of applying the row operation to the product AB and $(EA)B$ is the result of applying the row operation to A and then multiplying by B . This establishes the important principle enunciated earlier in this section and upon which Gauss-Jordan reduction is based.

Exercises for 10.4.

1. Solve each of the following systems by Gauss-Jordan elimination *if there is a solution*.

(a)

$$\begin{aligned}x_1 + 2x_2 + 3x_3 &= 4 \\3x_1 + x_2 + 2x_3 &= -1 \\x_1 &\quad + x_3 = 0\end{aligned}$$

(b)

$$\begin{aligned}x_1 + 2x_2 + 3x_3 &= 4 \\2x_1 + 3x_2 + 2x_3 &= -1 \\x_1 + x_2 - x_3 &= 10\end{aligned}$$

(c)

$$\begin{bmatrix} 1 & 1 & -2 & 3 \\ 2 & 1 & 0 & 1 \\ 1 & -1 & 1 & 0 \\ 3 & 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 9 \\ -18 \\ -9 \\ 9 \end{bmatrix}.$$

2. Use Gaussian elimination to solve

$$\begin{bmatrix} 3 & 2 \\ 2 & 1 \end{bmatrix} X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

where X is an unknown 2×2 matrix.

3. Calculate

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}.$$

Hint: Use the row operations suggested by the first three matrices.

4. What is the effect of multiplying a 2×2 matrix A on the right by the elementary matrix

$$\begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix}?$$

What general rule is this a special case of?

10.5 Singularity, Pivots, and Invertible Matrices

Let A be a square coefficient matrix. Gauss-Jordan reduction will work as indicated in the previous section if A can be reduced by a sequence of elementary row operations to the identity matrix I . A square matrix with this property is called *non-singular* or *invertible*. (The reason for the latter terminology will be clear shortly.) If it cannot be so reduced, it is called *singular*. Clearly, there are singular matrices. For example, the matrix equation

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

is equivalent to the system of 2 equations in 2 unknowns

$$\begin{aligned} x_1 + x_2 &= 1 \\ x_1 + x_2 &= 0 \end{aligned}$$

which is *inconsistent* and has no solution. Thus Gauss-Jordan reduction certainly can't work on its coefficient matrix.

To understand how to tell if a square matrix A is non-singular or not, we look more closely at the Gauss-Jordan reduction process. The basic strategy is the following. Start with the first row, and use type (1) row operations to eliminate all entries in the first column below the 1, 1-position. A leading non-zero entry when used in this way is called a *pivot*. There is one problem with this course of action: the leading non-zero entry in the first row may not be in the 1, 1-position. In that case, *first* interchange the first row with a succeeding row which does have a non-zero entry in the first column. (If you think about it, you may still see a problem. We shall come back to this and related issues later.)

After the first reduction, the coefficient matrix will have been transformed to a matrix of the form

$$\begin{bmatrix} p_1 & * & \dots & * \\ 0 & * & \dots & * \\ \vdots & \vdots & \dots & \vdots \\ 0 & * & \dots & * \end{bmatrix}$$

where p_1 is the (first) pivot. We now do something mathematicians (and computer scientists) love: repeat the same process for the submatrix consisting of the second and subsequent rows. If we are fortunate, we will be able to transform A ultimately by a sequence of elementary row operations into matrix of the form

$$\begin{bmatrix} p_1 & * & * & \dots & * \\ 0 & p_2 & * & \dots & * \\ 0 & 0 & p_3 & \dots & * \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & p_n \end{bmatrix}$$

with pivots on the diagonal and nonzero-entries in those pivot positions. (Such a matrix is also called an *upper triangular matrix* because it has zeroes below the diagonal.) We may now start in the lower right hand corner and apply the Jordan reduction process. In this way each of the entries above the diagonal pivots may be eliminated, so we obtain a diagonal matrix

$$\begin{bmatrix} p_1 & 0 & 0 & \dots & 0 \\ 0 & p_2 & 0 & \dots & 0 \\ 0 & 0 & p_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & p_n \end{bmatrix}$$

with non-zero entries on the diagonal. We may now finish off the process by applying type (2) operations to the rows as needed and finally obtain the identity matrix I as required.

The above analysis makes clear that the placement of the pivots is what is essential to non-singularity. What can go wrong? It may happen for a given row that the leading non-zero entry *is not in the diagonal position*, and there is no way to remedy this by interchanging with a subsequent row. In that case, we just do the best we can. *We use a pivot as far to the left as possible (after suitable row interchange with a subsequent row where necessary).* In the extreme case, it may turn out that the submatrix we are working with consists only of zeroes, and there are no possible pivots to choose, so we stop. For a square matrix, this extreme case must occur, since we will run out of pivot positions before we run out of rows. Thus, the Gaussian reduction will still transform A to an upper triangular matrix A' , but some of the diagonal entries will be zero and some of the last rows (perhaps only the last row) will consist of zeroes. That is the singular case.

Example 195

$$\begin{aligned} \begin{bmatrix} 1 & 2 & -1 \\ 1 & 2 & 0 \\ 1 & 2 & -2 \end{bmatrix} &\rightarrow \begin{bmatrix} 1 & 2 & -1 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \end{bmatrix} && \text{clear 1st column} \\ &\rightarrow \begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} && \text{no pivot in 2, 2 position} \end{aligned}$$

Note that the last row consists of zeroes.

We showed in the previous section that if the $n \times n$ matrix A is non-singular, then every equation of the form $AX = B$ (where both X and B are $n \times p$ matrices) has a solution and also that the solution $X = B'$ is *unique*. On the other hand, if A is *singular*, an equation of the form $AX = B$ *may* have a solution, but *there will certainly be matrices B for which $AX = B$ has no solutions*. This is best illustrated by an example.

Example 196 Consider the system

$$\begin{bmatrix} 1 & 2 & -1 \\ 1 & 2 & 0 \\ 1 & 2 & -2 \end{bmatrix} \mathbf{x} = \mathbf{b}$$

where \mathbf{x} and \mathbf{b} are 3×1 column vectors. Without specifying \mathbf{b} , the reduction of the augmented matrix for this system would follow the scheme

$$\begin{bmatrix} 1 & 2 & -1 & b_1 \\ 1 & 2 & 0 & b_2 \\ 1 & 2 & -2 & b_3 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & -1 & * \\ 0 & 0 & 1 & * \\ 0 & 0 & -1 & * \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & 0 & b'_1 \\ 0 & 0 & 1 & b'_2 \\ 0 & 0 & 0 & b'_3 \end{bmatrix}.$$

Now simply choose $b'_3 = 1$ (or any other non-zero value), so the reduced system is inconsistent. (Its last equation would be $0 = b'_3 \neq 0$.) Since, the two row operations may be reversed, we can now work back to a system with the original coefficient matrix which is also inconsistent. (Check in this case that if you choose $b'_1 = 0, b'_2 = 1, b'_3 = 1$, then reversing the operations yields $b_1 = -1, b_2 = 0, b_3 = -1$.)

The general case is completely analogous. Suppose

$$A \rightarrow \cdots \rightarrow A'$$

is a sequence of elementary row operations which transforms A to a matrix A' for which the last row consists of zeroes. Choose any $n \times p$ matrix B' for which the last row *does not* consist of zeroes. Then the equation

$$A'X = B'$$

cannot be valid since the last row on the left will necessarily consist of zeroes. Now reverse the row operations in the sequence which transformed A to A' . Let B be the effect of this reverse sequence on B' .

$$\begin{aligned} A &\leftarrow \cdots \leftarrow A' \\ B &\leftarrow \cdots \leftarrow B' \end{aligned}$$

Then the equation

$$AX = B$$

cannot be consistent because the equivalent system $A'X = B'$ is not consistent.

We shall see later that when A is a singular $n \times n$ matrix, if $AX = B$ has a solution X for a particular B , then it has infinitely many solutions.

There is one unpleasant possibility we never mentioned. It is conceivable that the standard sequence of elementary row operations transforms A to the identity matrix, so we decide it is non-singular, but some other bizarre sequence of elementary row operations transforms it to a matrix with some rows consisting of zeroes, in which case we should decide it is singular. Fortunately this can never happen because singular matrices and non-singular matrices have diametrically opposed properties. For example, if A is non-singular then $AX = B$ has a solution for every B , while if A is singular, there are many B for which $AX = B$ has no solution. This fact does not depend on the method we use to find solutions.

Inverses of Non-singular Matrices Let A be a non-singular $n \times n$ matrix. According to the above analysis, the equation

$$AX = I$$

(where we take B to be the $n \times n$ identity matrix I) has a unique $n \times n$ solution matrix $X = B'$. This B' is called the *inverse* of A , and it is usually denoted A^{-1} . That explains why non-singular matrices are also called *invertible*.

Example 197 Consider

$$A = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 1 & 0 \\ 1 & 2 & 0 \end{bmatrix}$$

To solve $AX = I$, we reduce the augmented matrix $[A | I]$.

$$\begin{aligned} \left[\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 0 & 0 & 1 \end{array} \right] &\rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & 1 & -1 & 1 & 0 \\ 0 & 2 & 0 & -1 & 0 & 1 \end{array} \right] \\ &\rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & 1 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 & -2 & 1 \end{array} \right] \\ &\rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & 2 & -1 \end{array} \right] \\ &\rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 2 & -1 \\ 0 & 1 & 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 & 2 & -1 \end{array} \right]. \end{aligned}$$

(You should make sure you see which row operations were used in each step.) Thus, the solution is

$$X = A^{-1} = \begin{bmatrix} 0 & 2 & -1 \\ 0 & -1 & 1 \\ -1 & 2 & -1 \end{bmatrix}.$$

Check the answer by calculating

$$A^{-1}A = \begin{bmatrix} 0 & 2 & -1 \\ 0 & -1 & 1 \\ -1 & 2 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 1 & 1 & 0 \\ 1 & 2 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

There is a subtle point about the above calculations. The matrix inverse $X = A^{-1}$ was derived as the *unique* solution of the equation $AX = I$, but we checked it by calculating $A^{-1}A = I$. The definition of A^{-1} told us only that $AA^{-1} = I$. Since matrix multiplication is not generally commutative, how could we be sure that the product *in the other order* would also be the identity I ? The answer is provided by the following tricky argument. Let $Y = A^{-1}A$. Then

$$AY = A(A^{-1}A) = (AA^{-1})A = IA = A$$

so that Y is the *unique* solution of the equation $AY = A$. However, $Y = I$ is *also* a solution of that equation, so we may conclude that $A^{-1}A = Y = I$. The upshot is that for a non-singular square matrix A , we have *both* $AA^{-1} = I$ and $A^{-1}A = I$.

The existence of matrix inverses for non-singular square matrices suggests the following scheme for solving matrix equations of the form

$$AX = B.$$

First, find the matrix inverse A^{-1} , and then take $X = A^{-1}B$. This is indeed the solution since

$$AX = A(A^{-1}B) = (AA^{-1})B = IB = B.$$

However, as easy as this looks, one should not be misled by the formal algebra. Note that the only method we have for finding the matrix inverse is to apply Gauss-Jordan reduction to the augmented matrix $[A | I]$. If B has fewer than n columns, then applying Gauss-Jordan reduction directly to $[A | B]$ would ordinarily involve less computation than finding A^{-1} . Hence, in the most common cases, applying Gauss-Jordan reduction to the original system of equations is the best strategy.

Numerical Considerations in Computation The examples we have chosen to

illustrate the principles employ small matrices for which one may do exact arithmetic. The worst that will happen is that some of the fractions may get a bit messy. In real applications, the matrices are often quite large, and it is not practical to do exact arithmetic. The introduction of rounding and similar numerical approximations complicates the situation, and computer programs for solving systems of equations have to deal with problems which arise from this. If one is not careful in designing such a program, one can easily generate answers which are very far off, and even deciding when an answer is sufficiently accurate sometimes involves rather subtle considerations. Typically, one encounters problems for matrices where the entries differ radically in size. Also, because of rounding, few matrices are ever *exactly* singular since one can never be sure that a very small numerical value at a potential pivot would have been zero if the calculations had been done exactly. On

the other hand, it is not surprising that matrices which are close to being singular can give computer programs indigestion.

If you are interested in such questions, there are many introductory texts which discuss numerical linear algebra. Two such are *Introduction to Linear Algebra* by Johnson, Riess, and Arnold and *Applied Linear Algebra* by Noble and Daniel. One of the computer assignments in your programming course is concerned with some of the problems of numerical linear algebra.

Exercises for 10.5.

1. In each of the following cases, find the matrix inverse if one exists. Check your answer by multiplication.

$$(a) \begin{bmatrix} 1 & -1 & -2 \\ 2 & 1 & 1 \\ 2 & 2 & 2 \end{bmatrix}$$

$$(b) \begin{bmatrix} 1 & 4 & 1 \\ 1 & 1 & 2 \\ 1 & 3 & 1 \end{bmatrix}$$

$$(c) \begin{bmatrix} 1 & 2 & -1 \\ 2 & 3 & 3 \\ 4 & 7 & 1 \end{bmatrix}$$

$$(d) \begin{bmatrix} 2 & 2 & 1 & 1 \\ -1 & 1 & -1 & 0 \\ 1 & 0 & 1 & 2 \\ 2 & 2 & 1 & 2 \end{bmatrix}$$

2. Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, and suppose $\det A = ad - bc \neq 0$. Show that

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Hint: If someone is kind enough to suggest to you what A^{-1} is, you need not ‘find’ it. Just check that it works by multiplication.

3. Let A and B be invertible $n \times n$ matrices. Show that $(AB)^{-1} = B^{-1}A^{-1}$. Note the reversal of order! Hint: As above, if you are given a candidate for an inverse, you needn’t ‘find’ it; you need only check that it works.
4. In the general discussion of Gauss-Jordan reduction, we assumed for simplicity that there was at least one non-zero entry in the first column of the coefficient matrix A . That was done so that we could be sure there would be a non-zero entry in the 1, 1-position (after a suitable row interchange) to use as a pivot. What if the first column consists entirely of zeroes? Does the basic argument (for the singular case) still work?

10.6 Gauss-Jordan Reduction in the General Case

Gauss-Jordan reduction works just as well if the coefficient matrix A is singular or even if it is not a square matrix. Consider the system

$$A\mathbf{x} = \mathbf{b}$$

where the coefficient matrix A is an $m \times n$ matrix. We shall concentrate on the case that \mathbf{x} is an $n \times 1$ column vector of unknowns and \mathbf{b} is a given $m \times 1$ column vector. (This illustrates the principles, and the case $AX = B$ where X and B are $n \times p$ matrices with $p > 1$ works in a similar manner.) The method is to apply elementary row operations to the augmented matrix

$$[A | \mathbf{b}] \rightarrow \cdots \rightarrow [A' | \mathbf{b}']$$

making the best of it with the coefficient matrix A . We may not be able to transform A to the identity matrix, but we can always pick out a set of pivots, one in each non-zero row, and otherwise mimic what we did in the case of a square non-singular A . If we are fortunate, the resulting system $A'\mathbf{x} = \mathbf{b}'$ will have solutions.

Example 198 Consider

$$\begin{bmatrix} 1 & 1 & 2 \\ -1 & -1 & 1 \\ 1 & 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \\ 3 \end{bmatrix}.$$

Reduce the augmented matrix as follows

$$\left[\begin{array}{ccc|c} 1 & 1 & 2 & 1 \\ -1 & -1 & 1 & 5 \\ 1 & 1 & 3 & 3 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 1 & 1 & 2 & 1 \\ 0 & 0 & 3 & 6 \\ 0 & 0 & 1 & 2 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 1 & 1 & 2 & 1 \\ 0 & 0 & 3 & 6 \\ 0 & 0 & 0 & 0 \end{array} \right]$$

This completes the ‘Gaussian’ part of the reduction with pivots in the 1, 1 and 2, 3 positions, and the last row of the transformed coefficient matrix consists of zeroes. Let’s now proceed with the ‘Jordan’ part of the reduction. Use the last pivot to clear the column above it.

$$\left[\begin{array}{ccc|c} 1 & 1 & 2 & 1 \\ 0 & 0 & 3 & 6 \\ 0 & 0 & 0 & 0 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 1 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 1 & 1 & 0 & -3 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array} \right]$$

and the resulting augmented matrix corresponds to the system

$$\begin{aligned} x_1 + x_2 &= -3 \\ x_3 &= 2 \\ 0 &= 0 \end{aligned}$$

Note that the last equation could just as well have read $0 = 6$ (or some other non-zero quantity) in which case the system would be inconsistent and not have a

solution. Fortunately, that is not the case in this example. The second equation tells us $x_3 = 2$, but the first equation only gives a relation $x_1 = -3 - x_2$ between x_1 and x_2 . That means that the solution has the form

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -3 - x_2 \\ x_2 \\ 2 \end{bmatrix} = \begin{bmatrix} -3 \\ 0 \\ 2 \end{bmatrix} + x_2 \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$$

where x_2 can have *any value whatsoever*. We say that x_2 is a free variable, and the fact that it is arbitrary means that there are *infinitely many solutions*. x_1 and x_3 are called *bound* variables.

It is instructive to reinterpret this geometrically in \mathbf{R}^3 . The original system of equations may be written

$$\begin{aligned} x_1 + x_2 + 2x_3 &= 1 \\ -x_1 - x_2 + x_3 &= 5 \\ x_1 + x_2 + 3x_3 &= 3 \end{aligned}$$

which are equations for 3 planes in \mathbf{R}^3 . Solutions

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

correspond to points lying in the common intersection of those planes. Normally, we would expect three planes to intersect in a single point. That would have been the case had the coefficient matrix been non-singular. However, in this case the planes intersect in a line, and the solution obtained above may be interpreted as the vector equation of that line. If we put $x_2 = s$ and rewrite the equation using vector notation, we obtain

$$\mathbf{x} = \langle -3, 0, 2 \rangle + s \langle -1, 1, 0 \rangle.$$

Example 198 illustrates many features of the general procedure. Gauss-Jordan reduction of the coefficient matrix is always possible, but the pivots don't always end up on the diagonal. In any case, the Jordan part of the reduction will yield a 1 in each pivot position with zeroes above and below the pivot in that column. In any given row of the reduced coefficient matrix, the pivot will be on the diagonal or to its right, and all entries *to the left of the pivot* will be zero. (Some of the entries to the right of the pivot may be non-zero.) If the number of pivots is smaller than the number of rows (which will always be the case for a singular square matrix), then some rows of the reduced coefficient matrix will consist entirely of zeroes. If there are non-zero entries in those rows to the right of the divider *in the augmented matrix*, the system is inconsistent and has no solutions. Otherwise, the system does have solutions. Such solutions are obtained by writing out the corresponding system, and transposing all terms *not associated with the pivot position* to the right

side of the equation. Each unknown in a pivot position is then expressed in terms of the non-pivot unknowns (if any). The pivot unknowns are said to be *bound*. The non-pivot unknowns may be assigned any value and are said to be *free*.

Example 199 Consider

$$\begin{bmatrix} 1 & 2 & -1 & 0 \\ 1 & 2 & 1 & 3 \\ 2 & 4 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \quad (164)$$

Reducing the augmented matrix yields

$$\begin{aligned} \left[\begin{array}{cccc|c} 1 & 2 & -1 & 0 & 0 \\ 1 & 2 & 1 & 3 & 0 \\ 2 & 4 & 0 & 3 & 0 \end{array} \right] &\rightarrow \left[\begin{array}{cccc|c} 1 & 2 & -1 & 0 & 0 \\ 0 & 0 & 2 & 3 & 0 \\ 0 & 0 & 2 & 3 & 0 \end{array} \right] \rightarrow \left[\begin{array}{cccc|c} 1 & 2 & -1 & 0 & 0 \\ 0 & 0 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right] \\ &\rightarrow \left[\begin{array}{cccc|c} 1 & 2 & -1 & 0 & 0 \\ 0 & 0 & 1 & 3/2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right] \rightarrow \left[\begin{array}{cccc|c} 1 & 2 & 0 & 3/2 & 0 \\ 0 & 0 & 1 & 3/2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right]. \end{aligned}$$

(Note that since there are zeroes to the right of the divider, we don't have to worry about possible inconsistency in this case.) The system corresponding to the reduced augmented matrix is

$$\begin{aligned} x_1 + 2x_2 &+ (3/2)x_4 = 0 \\ x_3 + (3/2)x_4 &= 0 \\ 0 &= 0 \end{aligned}$$

Thus,

$$\begin{aligned} x_1 &= -2x_2 - (3/2)x_4 \\ x_3 &= -3(2)x_4 \end{aligned}$$

with x_1 and x_3 *bound* and x_2 and x_4 *free*. A general solution has the form

$$\begin{aligned} \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} &= \begin{bmatrix} -2x_2 - (3/2)x_4 \\ x_2 \\ -3(2)x_4 \\ x_4 \end{bmatrix} = \begin{bmatrix} -2x_2 \\ x_2 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -(3/2)x_4 \\ 0 \\ -(3/2)x_4 \\ 0 \end{bmatrix} \\ \mathbf{x} &= x_2 \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} -3/2 \\ 0 \\ -3/2 \\ 0 \end{bmatrix} \end{aligned}$$

where x_2 and x_4 can assume any value.

This solution may also be interpreted geometrically in \mathbf{R}^4 . Introduce two vectors

$$\begin{aligned}\mathbf{v}_1 &= \langle -2, 1, 0, 0 \rangle \\ \mathbf{v}_2 &= \langle -3/2, 0, -3/2, 0 \rangle\end{aligned}$$

in \mathbf{R}^4 . Note that neither of these vectors is a multiple of the other. Hence, we may think of them as spanning a (2-dimensional) plane in \mathbf{R}^4 . Putting $s_1 = x_2$ and $s_2 = x_4$, we may express the general solution vector as

$$\mathbf{x} = s_1 \mathbf{v}_1 + s_2 \mathbf{v}_2,$$

so the solution set of the system (164) may be identified with the plane spanned by $\{\mathbf{v}_1, \mathbf{v}_2\}$.

Make sure you understand the procedure used in the above examples to express the general solution vector \mathbf{x} entirely in terms of the free variables. We shall use it quite generally.

Any system of equations with real coefficients may be interpreted as defining a locus in \mathbf{R}^n , and studying the structure—in particular, the dimensionality—of such a locus is something which will concern us later.

Example 200 Consider

$$\begin{bmatrix} 1 & 2 \\ 1 & 0 \\ -1 & 1 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \\ -7 \\ 10 \end{bmatrix}.$$

Reducing the augmented matrix yields

$$\begin{aligned} \left[\begin{array}{cc|c} 1 & 2 & 1 \\ 1 & 0 & 5 \\ -1 & 1 & -7 \\ 2 & 0 & 10 \end{array} \right] &\rightarrow \left[\begin{array}{cc|c} 1 & 2 & 1 \\ 0 & -2 & 4 \\ 0 & 3 & -6 \\ 0 & -4 & 8 \end{array} \right] \rightarrow \left[\begin{array}{cc|c} 1 & 2 & 1 \\ 0 & -2 & 4 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right] \\ &\rightarrow \left[\begin{array}{cc|c} 1 & 2 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right] \rightarrow \left[\begin{array}{cc|c} 1 & 0 & 5 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right] \end{aligned}$$

which is equivalent to

$$\begin{aligned}x_1 &= 5 \\ x_2 &= -2.\end{aligned}$$

Thus the unique solution vector is

$$\mathbf{x} = \begin{bmatrix} 5 \\ -2 \end{bmatrix}.$$

These examples and the preceding discussion lead us to certain conclusions about a system of the form

$$A\mathbf{x} = \mathbf{b}$$

where A is an $m \times n$ matrix, \mathbf{x} is an $n \times 1$ column vector of unknowns, and \mathbf{b} is an $m \times 1$ column vector that is given.

The number r of pivots of A is called the *rank* of A , and clearly it plays a crucial role. It is the same as the number of non-zero rows at the end of the Gauss-Jordan reduction since there is exactly one pivot in each non-zero row. The rank is certainly not greater than either the number of rows m or the number of columns n of A .

If $m = n$, i.e., A is a square matrix, then A is non-singular when its rank is n and it is singular when its rank is smaller than n .

More generally for an $m \times n$ matrix A , if the rank r is smaller than the number of rows m , then there are $m \times 1$ column vectors \mathbf{b} such that the system $A\mathbf{x} = \mathbf{b}$ *does not have any solutions*. The argument is basically the same as for the case of a singular square matrix. Transform A by a sequence of elementary row operations to a matrix A' with its last row consisting of zeroes, choose \mathbf{b}' so that $A'\mathbf{x} = \mathbf{b}'$ is inconsistent, and reverse the operations to find an inconsistent $A\mathbf{x} = \mathbf{b}$.

If for a given \mathbf{b} , the system $A\mathbf{x} = \mathbf{b}$ does have solutions, then the unknowns x_1, x_2, \dots, x_n may be partitioned into two sets: r bound unknowns and $n - r$ free unknowns. The bound unknowns are expressed in terms of the free unknowns. The number $n - r$ of free unknowns is sometimes called the *nullity* of the matrix A . If the nullity $n - r > 0$, i.e., $n > r$, then (if there are any solutions at all) there are infinitely many solutions.

Systems of the form

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

are called *homogeneous*. Example 199 is a homogeneous system. Gauss-Jordan reduction of a homogeneous system always succeeds since the matrix \mathbf{b}' obtained from $\mathbf{b} = 0$ is also zero. If $m = n$, i.e., the matrix is square, and A is non-singular, the *only solution* is 0, but if A is singular, i.e., $r < n$, then there are definitely non-zero solutions since there are some free unknowns which can be assigned non-zero values. This rank argument works for any m and n : if $r < n$, then there are definitely non-zero solutions for the homogeneous system $A\mathbf{x} = 0$. One special case of interest is $m < n$. Since $r \leq m$, we must have $r < n$ in that case. That leads to the following important principle: *a homogeneous system of linear algebraic equations for which there are more unknowns than equations always has some non-trivial solutions*.

Pseudo-inverses In some applications, one needs to try to find ‘inverses’ of non-square matrices. Thus, if A is a $m \times n$ matrix, one might need to find an $n \times m$ matrix A' such that

$$AA' = I \quad \text{the } m \times m \text{ identity.}$$

Such an A' would be called a *right pseudo-inverse*. Similarly, an $n \times m$ matrix A'' such that

$$A''A = I \quad \text{the } n \times n \text{ identity}$$

is called a left pseudo-inverse.

If $m > n$, i.e., A has more rows than columns, then *no right pseudo-inverse is possible*. For, suppose we could find an $n \times m$ matrix A' such that $AA' = I$ (the $m \times m$ identity matrix). Then for any $m \times 1$ column vector \mathbf{b} , $\mathbf{x} = A'\mathbf{b}$ is a solution of $A\mathbf{x} = \mathbf{b}$ since

$$A\mathbf{x} = A(A'\mathbf{b}) = (AA')\mathbf{b} = I\mathbf{b} = \mathbf{b}.$$

On the other hand, we know that since $m > n \geq r$, there is at least one \mathbf{b} such that $A\mathbf{x} = \mathbf{b}$ does not have a solution.

On the other hand, if $m < n$ and the rank of A is m (which is as large as it can get in any case), then it is always possible to find a right pseudo-inverse. To see this, let

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

and consider the matrix equation

$$AX = I.$$

It may be viewed as m separate equations of the form

$$A\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, A\mathbf{x} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \dots, A\mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

one for each column of I . Since $r = m$, each of these equations has a solution. (In fact it will generally have infinitely many solutions.)

Exercises for 10.6.

1. In each of the following cases, apply the Gauss-Jordan reduction process to find a general solution, if one exists. As in the text, the answer should express the general solution \mathbf{x} as a ‘particular solution’ (possibly zero) plus a linear combination of ‘basic solutions’ with the free unknowns (if any) as coefficients.

$$(a) \begin{bmatrix} 1 & -6 & -4 \\ 3 & -8 & -7 \\ -2 & 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -3 \\ -5 \\ 2 \end{bmatrix}.$$

$$(b) \begin{bmatrix} 1 & 2 \\ 3 & 1 \\ 4 & 3 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 1 \end{bmatrix}.$$

$$(c) \begin{bmatrix} 1 & -2 & 2 & 1 \\ 1 & -2 & 1 & 2 \\ 3 & -6 & 4 & 5 \\ 1 & -2 & 3 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 6 \\ 4 \\ 14 \\ 8 \end{bmatrix}.$$

2. Find a general solution vector of the system $A\mathbf{x} = 0$ where

$$(a) A = \begin{bmatrix} 1 & 0 & 1 & 2 \\ 2 & -1 & 1 & 0 \\ -1 & 4 & -1 & -2 \end{bmatrix} \quad (b) A = \begin{bmatrix} 1 & 3 & 4 & 0 & 2 \\ 2 & 7 & 6 & 1 & 1 \\ 4 & 13 & 14 & 1 & 3 \end{bmatrix}$$

3. What is the rank of the coefficient matrix for each of the matrices in the previous problem.
4. What is the rank of each of the following matrices?

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}, \quad \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

5. Let A be an $m \times n$ matrix with $m < n$, and let r be its rank. Which of the following is always true, sometimes true, never true?
- (a) $r \leq m < n$. (b) $m < r < n$. (c) $r = m$. (d) $r = n$. (e) $r < m$. (f) $r = 0$.
6. How do you think the rank of a product AB compares to the rank of A ? Is the former rank always \leq , \geq , or $=$ the latter rank? Try some examples, make a conjecture, and see if you can prove it. Hint: Look at the number of rows of zeroes after you reduce A completely to A' . Could further reduction transform $A'B$ to a matrix with more rows of zeroes?
7. Find a right pseudo-inverse A' for

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \end{bmatrix}.$$

Note that there are infinitely many answers to this problem. You need only find one, but if you are ambitious, you can find all of them. Is there a left pseudo-inverse for A . If there is find one, if not explain why not.

10.7 Vector Spaces

Many of the notions we encountered when studying linear differential equations have interesting analogues in the theory of solutions of systems of algebraic equations. Both these theories have remarkable similarities to the theory of vectors in two and three dimensions. We give some examples.

The general solution of a first order linear equation

$$y' + a(t)y = b(t)$$

has the form

$$y = y_p + ch$$

where y_p is one solution, h is a solution of the corresponding homogeneous equation, and c is a scalar which may assume any value. Similarly, we saw in Example 1 in the previous section that the general solution of the system of equations had the form

$$\mathbf{x} = \mathbf{x}_0 + s\mathbf{v}$$

where $\mathbf{x}_0 = \langle -3, 0, 2 \rangle$ is one solution (for $s = 0$), $\mathbf{v} = \langle -1, 1, 0 \rangle$, and s is a scalar which may assume any value. Each of these is reminiscent of the equation of a line in \mathbf{R}^3

$$\mathbf{r} = \mathbf{r}_0 + s\mathbf{v}$$

where \mathbf{r}_0 is the position vector of a point on the line, \mathbf{v} is a vector parallel to the line, and s is a scalar which may assume any value.

Furthermore, we saw that the general solution of the second order linear homogeneous differential equation

$$y'' + p(t)y' + q(t)y = 0$$

has the form

$$y = c_1y_1 + c_2y_2$$

where $\{y_1, y_2\}$ is a linearly independent pair of solutions and c_1 and c_2 are two scalars which may assume any values. Similarly, we saw in Example 2 in the previous section that the general solution of the system had the form

$$\mathbf{x} = s_1\mathbf{v}_1 + s_2\mathbf{v}_2$$

where $\mathbf{v}_1 = \langle -2, 1, 0, 0 \rangle$, $\mathbf{v}_2 = \langle -3/2, 0, -3/2, 0 \rangle$ and s_1 and s_2 are scalars which may assume any values. Note that the pair $\{\mathbf{v}_1, \mathbf{v}_2\}$ is *linearly independent* in exactly the same sense that a pair of solutions of a differential equation is linearly independent, i.e., *neither vector is a scalar multiple of the other*. Both these situations are formally similar to what happens for a plane in \mathbf{R}^3 which passes through the origin. If $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a pair of vectors in the plane, and neither is a multiple of the other, then a general vector in that plane can be expressed as a linear combination

$$\mathbf{v} = s_1\mathbf{v}_1 + s_2\mathbf{v}_2.$$

As one studies these different theories, more and more similarities arise. For example, we just remarked that the general solution of the differential equation

$$y'' + p(t)y' + q(t)y = f(t)$$

has the form

$$y = y_p + \text{a general solution of the corresponding homogeneous equation}$$

where y_p is particular solution of the inhomogeneous equation. Exactly the same rule applies to the general solution of the inhomogeneous algebraic system

$$A\mathbf{x} = \mathbf{b}.$$

Namely, suppose \mathbf{x}_p is *one particular solution* of this inhomogeneous system, and \mathbf{x} is any other solution. Then,

$$\begin{aligned} A\mathbf{x} &= \mathbf{b} \\ A\mathbf{x}_p &= \mathbf{b} \end{aligned}$$

and subtraction yields

$$A\mathbf{x} - A\mathbf{x}_p = A(\mathbf{x} - \mathbf{x}_p) = 0,$$

i.e., $\mathbf{z} = \mathbf{x} - \mathbf{x}_p$ is a solution of the homogeneous system $A\mathbf{z} = 0$. Thus

$$\mathbf{x} = \mathbf{x}_p + \mathbf{z} = \mathbf{x}_p + \text{a general solution of the homogeneous system}.$$

The important point to note is that not only is the conclusion the same, but the argument used to derive it is also essentially the same.

Whenever mathematicians notice that the same phenomena are observed in different contexts, and that the same arguments are used to study those phenomena, they look for a common way to describe what is happening. One of the major accomplishments of the late nineteenth and early twentieth centuries was the realization among mathematicians that there is a single concept, that of an *abstract vector space*, which may be used to study many diverse mathematical phenomena, including those mentioned above. They discovered that the common aspects of all such theories were based on the fact that they share certain *operations*, and that these operations, although defined differently in each individual case, *all obey common rules*. Any argument which uses only those rules will be valid in all cases.

There are two basic operations which must be present for a collection of objects to constitute a vector space. (Additional operations may also be present, but for the moment we ignore them.) First, there should be an operation of addition which obeys the usual rules you are familiar with for addition of vectors in space. Thus, addition should be an associative operation, it should obey the commutative law, and there should be an element called zero (0) which added to any other element results in the same element. Finally, every element must have a *negative* which when added to the original element yields zero.

For example, in \mathbf{R}^n , addition $\mathbf{x} + \mathbf{y}$ is done by adding corresponding entries of the two n -tuples. The zero element is the n -tuple $\mathbf{0}$ for which all entries are zero. For any \mathbf{x} in \mathbf{R}^n , the negative of \mathbf{x} is just the n -tuple $-\mathbf{x}$.

Alternatively, let \mathbf{S} denote the set of all solutions of the second order homogeneous linear differential equation

$$y'' + p(t)y' + q(t)y = 0.$$

In this case, the addition operation $y_1 + y_2$ is just the ordinary addition of functions, and the zero element is the function which is identically zero. The negative of a function is defined as expected by the rule $(-y)(t) = -y(t)$. Since the equation is homogeneous, the sum of two solutions is again a solution, the zero function is also a solution, and the negative of a solution is a solution. If such were not the case, the set \mathbf{S} would not be *closed* under the operations, so it would not be a vector space.

To have a vector space, we also need a second operation: we must be able to multiply objects by scalars. This operation must also obey the usual rules of vector algebra. Namely, it must satisfy the associative law, distributive laws, and multiplying an object by the scalar 1 should not change the object.

For example, in \mathbf{R}^n , we multiply a vector \mathbf{x} by a scalar c in the usual way.

For \mathbf{S} , the vector space of solutions of a 2nd order homogeneous linear differential equation, a function (solution) is multiplied by a scalar by the rule $(cy)(t) = cy(t)$. Again, since the differential equation is homogeneous, any scalar multiple of a solution is a solution, so the set \mathbf{S} is closed under the operation.

In courses in abstract algebra, one studies in detail the list of rules (or axioms) which govern the operations in a vector space. In particular, one derives all the usual rules of vector algebra from the properties listed above. In this course, we shall assume all that has been done, so you may safely manipulate objects in any vector space just the way you would manipulate ordinary vectors in space. Note the different levels of abstraction you have seen in this course for the concept ‘vector’. First a vector was just a quantity in the plane or in space with a magnitude and a direction. Vectors were added or multiplied by scalars using simple geometric rules. Later, we introduced the concept of a ‘vector’ in \mathbf{R}^n which was an n -tuple of real numbers. Such ‘vectors’ were added or multiplied by scalars by doing the same to their components. Finally, we are now considering ‘vectors’ which are *functions*. Such ‘vectors’ are added or multiplied by scalars by doing the same to the function values. As noted above, what is important is not the nature of the individual object we call a ‘vector’, but the properties of the operations we define on the *set* of all such ‘vectors’.

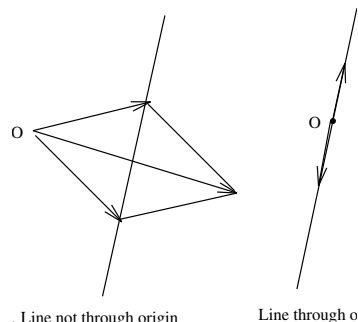
There are many other important examples of vector spaces. Often they are *function spaces*, that is, their elements are functions defined on some common domain. Here are a few examples of such vector spaces.

Let I be a real interval. The set $\mathcal{F}(I)$ of all real valued functions defined on I is vector space if we use the operations of adding functions and multiplying them by scalars. The set $\mathcal{C}(I)$ of all *continuous real valued functions* defined on I is also a vector space because the sum of two continuous functions is continuous and any scalar multiple of a continuous function is continuous. Similarly, the set $\mathcal{C}^2(I)$ of all *twice differentiable real valued functions* defined on I with *continuous second derivatives* is a vector space. As mentioned before, the set of all such differentiable functions which are solutions of a specified second order homogeneous linear differential equation is a vector space. Finally, the set of all real valued analytic functions defined on I is also a vector space.

Not every set of ‘vector’ like objects is a vector space. We have already seen several such examples. For example, the vector equation of a line in \mathbf{R}^3 has the form

$$\mathbf{r} = \mathbf{r}_0 + s\mathbf{v}$$

where \mathbf{r} is the position vector connecting the origin to a general point on the line, \mathbf{r}_0 is the position vector of one particular point on the line, and \mathbf{v} is a vector parallel to the line. If the line does not pass through the origin, the sum of two position vectors with end points on the line won’t be a vector with endpoint on the line. Hence, the set is not closed under addition. It is also not closed under multiplication by scalars.



Similarly, while the set of solutions of the homogeneous first order equation

$$y' + a(t)y = 0$$

is a vector space, the set of solution of the (inhomogeneous) first order equation

$$y' + a(t)y = b(t)$$

is not a vector space if $b(t)$ is not identically 0. This rule holds in general. The set of solutions of a homogeneous linear equation (of any kind) is a vector space, but the set of solutions of an inhomogeneous linear equation is not a vector space.

The Domain of Acceptable Scalars In the previous discussion, we assumed implicitly that all scalars were *real*. However, there are circumstances where it would be more appropriate to allow scalars which are either real or *complex*. For example, we know that it is easier to study solutions of a second order linear equation with constant coefficients if we allow *complex valued* solutions and use complex scalar constants when writing out the general solution. Thus, one must specify in any given case whether one is talking about a *real* vector space, where the set of possible scalars is restricted to \mathbf{R} , or a *complex* vector space, where the set of possible scalars is the larger domain \mathbf{C} .

One important complex vector space is the set \mathbf{C}^n of all $n \times 1$ column vectors with *complex* entries. Another is the set of all complex valued solutions of a given homogeneous linear differential equation.

One confusing point is that every complex vector space may also be considered to be a real vector space simply by agreeing to allow only real scalars in any expression in which a ‘vector’ is multiplied by a scalar. For example, the set \mathbf{C} of all complex numbers may itself be viewed as a real vector space in this way. If we do so, then, since any complex number $a + bi$ is determined by the pair (a, b) of its real and imaginary parts, as a real vector space \mathbf{C} is essentially the same as \mathbf{R}^2 . **Subspaces**

You may have noticed that in many of the examples listed previously, some vector spaces were *subsets* of others. If \mathbf{V} is a vector space, and \mathbf{W} is a non-empty subset, then we call \mathbf{W} a *subspace* of \mathbf{V} if whenever two elements of \mathbf{V} are in \mathbf{W} so is their sum and whenever an element of \mathbf{V} is in \mathbf{W} so is any scalar multiple of that element. This means that \mathbf{W} becomes a vector space in its own right under the operations it inherits from \mathbf{V} .

For $\mathbf{V} = \mathbf{R}^3$, most subspaces are what you would expect. Any line passing through the origin yields a subspace, but a line which does not pass through the origin does not. Similarly, any plane passing through the origin yields a subspace but a plane which does not pass through the origin does not. A less obvious subspace is the *zero subspace* which consists of the single vector $\mathbf{0}$. Also, in general mathematical usage, any set is considered to be a subset of itself, so \mathbf{R}^3 is also a subspace of \mathbf{R}^3 .

For any vector space \mathbf{V} , the zero subspace and the whole vector space are subspaces of \mathbf{V} .

For a less obvious example, let $\mathcal{C}^2(I)$ be the vector space of all functions, defined on the real interval I , with continuous second derivatives. Then the set of solutions of the homogeneous differential equation

$$y'' + p(t)y' + q(t)y = 0$$

is a subspace. In essence, we know this from earlier work, but let’s derive it again by a general argument. Consider the *operator*

$$L = \frac{d^2}{dt^2} + p(t)\frac{d}{dt} + q(t)$$

acting on twice differentiable functions $y(t)$, i.e., let

$$L(y) = y'' + p(t)y' + q(t)y.$$

With this notation, the differential equation may be written $L(y) = 0$.

L is what we call a *linear operator* because it obeys the following rules:

$$L(y_1 + y_2) = L(y_1) + L(y_2) \quad \text{for functions } y_1, y_2 \text{ one} \quad (166)$$

$$L(cy) = cL(y) \quad \text{for a function } y \text{ and a scalar } c \text{ two} \quad (166)$$

The fact that the set of solutions of $L(y) = 0$ is a subspace is a direct consequence of these rules. Namely, if $L(y_1) = 0$ and $L(y_2) = 0$, then rule (165) immediately gives $L(y_1 + y_2) = L(y_1) + L(y_2) = 0$. Similarly, if $L(y) = 0$, rule (166) immediately gives $L(cy) = cL(y) = 0$ for any scalar c .

This same argument would work for *any linear operator*, and we shall see that one of the most common ways to obtain a subspace is as the solution set or *null space* of a linear operator. For example, the solution set of a homogeneous system of algebraic equations

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

is also a subspace because it is the null space of an appropriate linear operator. (See the Exercises.)

Exercises for 10.7.

1. Determine if each of the following subsets of \mathbf{R}^3 is a vector subspace of \mathbf{R}^3 . If it is not a subspace, explain what fails.

(a) The set of all $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ such that $2x_1 - x_2 + 4x_3 = 0$.

(b) The set of all $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ such that $2x_1 - x_2 + 4x_3 = 3$.

(c) The set of all $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ such that $x_1^2 + x_2^2 - x_3^2 = 1$.

(d) The set of all \mathbf{x} of the form $\mathbf{x} = \begin{bmatrix} 1 + 2t \\ -3t \\ 2t \end{bmatrix}$ where t is allowed to assume any real value.

(e) The set of all \mathbf{x} of the form $\mathbf{x} = \begin{bmatrix} s + 2t \\ 2s - 3t \\ s + 2t \end{bmatrix}$ where s and t are allowed to assume any real values.

2. Which of the following sets of functions is a vector space under the operations discussed in the section. If not a vector space, what fails?
 - (a) The set of all polynomial functions of the form $f(t) = a_0 + a_1t + a_2t^2 + a_3t^3$ (of degree ≤ 3 .)
 - (b) The set of all polynomial functions with constant term 1.
 - (c) The set of all continuous, real valued functions f with domain $-1 \leq t \leq 1$ such that $f(0) = 0$.
 - (d) The set of all continuous, real valued functions f with domain $-1 \leq t \leq 1$ such that $f(0) = 1$.
 - (e) The set of all solutions of the differential equation $y'' + 4y' + 5y = \cos t$.
3. Show that the set of solutions of the first order equation $y' + a(t)y = b(t)$ is a vector space if and only if $b(t)$ is identically 0.
4. This problem is mostly just a matter of translating terminology from one context to another. Let A be an $m \times n$ matrix with real entries. Define an operator $L : \mathbf{R}^n \rightarrow \mathbf{R}^m$ which transforms n -tuples to m -tuples by the rule: $L(\mathbf{x}) = A\mathbf{x}$. (a) Show that L is a linear operator as defined in Section 7. (b) Show that the set of solutions of the homogeneous system $A\mathbf{x} = 0$ is a subspace of \mathbf{R}^n by repeating the argument given in the text that the null space of the operator L is a subspace.

10.8 Linear Independence, Bases, and Dimension

Let \mathbf{V} be a vector space. In general, \mathbf{V} will have infinitely many elements, but it is often possible to specify \mathbf{V} in terms of an appropriate finite subset. For example, we know that the vector space of solutions of a homogeneous second order linear differential equation consists of all linear combinations

$$c_1y_1 + c_2y_2$$

where $\{y_1, y_2\}$ is a linearly independent pair of solutions, and c_1, c_2 are arbitrary scalars. We say in this case that the pair $\{y_1, y_2\}$ is a *basis* for the space of solutions. We want to generalize this to ‘bases’ with more than two elements.

As before, let \mathbf{V} be any vector space and let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ be a non-empty finite subset of elements of \mathbf{V} . Such a *set* is called *linearly independent* if no element of the set can be expressed as a *linear combination* of the other elements in the set. For a set $\{\mathbf{v}_1, \mathbf{v}_2\}$ with two vectors, this subsumes the previous definition: neither vector should be a scalar multiple of the other. For a set $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ with three

elements it means that *no* relation of any of the following forms is possible:

$$\begin{aligned}\mathbf{v}_1 &= a_2\mathbf{v}_2 + a_3\mathbf{v}_3 \\ \mathbf{v}_2 &= b_1\mathbf{v}_1 + b_3\mathbf{v}_3 \\ \mathbf{v}_3 &= c_1\mathbf{v}_1 + c_2\mathbf{v}_2.\end{aligned}$$

The opposite of ‘linearly independent’ is ‘linearly dependent’.

To get a better hold on the concept, consider the (infinite) set of all *linear combinations*

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k = \sum_{i=1}^k c_i\mathbf{v}_i \quad (167)$$

where each coefficient c_i is allowed to range arbitrarily over the domain of scalars. It is not very hard to see that this infinite set is a *subspace* of \mathbf{V} . It is called the *subspace spanned by* $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$. You can think of the elements of this subspace as forming a general solution of some (homogeneous) problem. We would normally want to be sure that there aren’t any *redundant elements* in the spanning set $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$. If one \mathbf{v}_i could be expressed linearly in terms of the others, that expression for \mathbf{v}_i could be substituted in (167), and the result could be simplified by combining terms. We could thereby omit \mathbf{v}_i and express the general element in (167) as a linear combination of the other elements in the spanning set.

Example 201 Consider the set consisting of the following four vectors in \mathbf{R}^4 .

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \mathbf{v}_4 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ -1 \end{bmatrix}.$$

This set is not linearly independent since

$$\mathbf{v}_2 = \mathbf{v}_1 - \mathbf{v}_3. \quad (168)$$

Thus, any element in the subspace spanned by $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$ can be rewritten

$$\begin{aligned}c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 + c_4\mathbf{v}_4 &= c_1\mathbf{v}_1 + c_2(\mathbf{v}_1 - \mathbf{v}_3) + c_3\mathbf{v}_3 + c_4\mathbf{v}_4 \\ &= (c_1 + c_2)\mathbf{v}_1 + (c_3 - c_2)\mathbf{v}_3 + c_4\mathbf{v}_4 \\ &= c'_1\mathbf{v}_1 + c'_3\mathbf{v}_3 + c_4\mathbf{v}_4.\end{aligned}$$

On the other hand, if we delete the element \mathbf{v}_2 , the set consisting of the vectors

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \mathbf{v}_4 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ -1 \end{bmatrix}.$$

is linearly independent. To see this, just look carefully at the pattern of zeroes. For example, \mathbf{v}_1 has first component 1, and the other two have first component 0, so \mathbf{v}_1

could not be a linear combination of \mathbf{v}_2 and \mathbf{v}_3 . Similar arguments eliminate the other two possible relations. (What are those arguments?)

In the above example, we could just as well have written

$$\mathbf{v}_1 = \mathbf{v}_2 + \mathbf{v}_3$$

and eliminated \mathbf{v}_1 from the spanning set without loss. In general, there are many possible ways to delete redundant vectors from a spanning set.

A linearly independent subset $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ of a vector space \mathbf{V} which also spans \mathbf{V} is called a *basis* for \mathbf{V} . Many of the algorithms we have for solving homogeneous problems yield bases for the solution space. For example, as noted above, any linearly independent pair of solutions of a second order homogeneous linear equation is a basis for its solution space. Much of what we do later will be designed to generalize that to higher order differential equations and to systems of differential equations.

Let A be an $m \times n$ matrix, and let \mathbf{W} be the solution space of the homogeneous system

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

(To be definite, assume the matrix has real entries and that \mathbf{W} is the solution subspace of \mathbf{R}^n . However, the corresponding theory for a complex matrix with solution subspace in \mathbf{C}^n is basically the same.) The Gauss-Jordan reduction method always generates a basis for \mathbf{W} . We illustrate this with an example. (You should also go back and look at Example 2 in Section 6.)

Example 202 Consider

$$\begin{bmatrix} 1 & 1 & 0 & 3 & -1 \\ 1 & 1 & 1 & 2 & 1 \\ 2 & 2 & 1 & 5 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = 0.$$

To solve it, apply Gauss-Jordan reduction

$$\begin{aligned} \left[\begin{array}{ccccc|c} 1 & 1 & 0 & 3 & -1 & 0 \\ 1 & 1 & 1 & 2 & 1 & 0 \\ 2 & 2 & 1 & 5 & 0 & 0 \end{array} \right] &\rightarrow \left[\begin{array}{ccccc|c} 1 & 1 & 0 & 3 & -1 & 0 \\ 0 & 0 & 1 & -1 & 2 & 0 \\ 0 & 0 & 1 & -1 & 2 & 0 \end{array} \right] \\ &\rightarrow \left[\begin{array}{ccccc|c} 1 & 1 & 0 & 3 & 0 & 0 \\ 0 & 0 & 1 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right]. \end{aligned}$$

The last matrix is fully reduced with pivots in the 1,1 and 2,3 positions. The corresponding system is

$$\begin{aligned} x_1 + x_2 + 3x_4 &= 0 \\ x_3 - x_4 + 2x_5 &= 0 \end{aligned}$$

with x_1, x_3 bound and x_2, x_4 , and x_5 free. Expressing the bound variables in terms of the free variables yields

$$\begin{aligned}x_1 &= -x_2 - 3x_4 \\x_3 &= \quad + x_4 - 2x_5.\end{aligned}$$

The general solution vector, when expressed in terms of the free variables, is

$$\begin{aligned}\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} &= \begin{bmatrix} -x_2 - 3x_4 \\ x_2 \\ x_4 - 2x_5 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -x_2 \\ x_2 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -3x_4 \\ 0 \\ x_4 \\ x_4 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ -2x_5 \\ 0 \\ x_5 \end{bmatrix} \\ &= x_2 \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} -3 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} + x_5 \begin{bmatrix} 0 \\ 0 \\ -2 \\ 0 \\ 1 \end{bmatrix}.\end{aligned}$$

If we put

$$\mathbf{v}_1 = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} -3 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 0 \\ 0 \\ -2 \\ 0 \\ 1 \end{bmatrix},$$

and $c_1 = x_2$, $c_2 = x_4$, and $c_3 = x_5$, then the general solution takes the form

$$\mathbf{x} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + c_3 \mathbf{v}_3$$

where the scalars c_1, c_2, c_3 (being new names for the free variables) can assume any values. Also, the set $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is linearly independent. This is clear for the following reason. Each vector is associated with one of the free variables and has a 1 in that position where the other vectors necessarily have zeroes. Hence, none of the vectors can be linear combinations of the others. It follows that $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is a basis for the solution space.

The above example illustrates all the important aspects of the solution process for a homogeneous system

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

We state the important facts about the solution without going through the general proofs since they are just the same as what we did in the example but with a lot more confusing notation. The general solution has the form

$$\mathbf{x} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + \cdots + c_k \mathbf{v}_k$$

where $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ are *basic solutions* obtained by successively setting each free variable equal to 1 and the other free variables equal to zero. c_1, c_2, \dots, c_k are just

new names for the free variables. The set $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ is linearly independent because of the pattern of 1's and 0's at the positions of the free variables, and since it spans the solution space, it is a basis for the solution space.

There are some special cases which are a bit confusing. First, suppose there is only one basic solution \mathbf{v}_1 . Then, the set $\{\mathbf{v}_1\}$ with one element is indeed a basis. In fact, in any vector space, the set $\{\mathbf{v}\}$ consisting of a single *non-zero vector* is linearly independent. Namely, there are no other vectors in the set which it could be a linear combination of. In this case, the subspace spanned by $\{\mathbf{v}\}$ just consists of all multiples $c\mathbf{v}$ where c can be any scalar. A much more confusing case is that in which the spanning set is the *empty set*, i.e., the set with no elements. (That would arise, for example, if the zero solution were the unique solution of the homogeneous system, so there would be no free variables and no basic solutions.) This is dealt with as follows. First, the empty set is taken to be linearly independent by convention. Second, again by convention, we take every linear combination of *no vectors* to be zero. It follows that the empty set spans the zero subspace $\{0\}$, and is a basis for it. (Can you see why the above conventions imply that the set $\{0\}$ is *not* linearly independent?)

Let \mathbf{V} be a vector space. If \mathbf{V} has a basis $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ with n elements, then we say that \mathbf{V} is n -dimensional. That is, the dimension of a vector space is the number of elements in a basis.

For example, since the solution space of a 2nd order homogeneous linear differential equation has a basis with two elements, that solution space is 2-dimensional.

Not too surprisingly, the dimension of \mathbf{R}^n is n . To see this we note that the set consisting of the vectors

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \dots, \quad \mathbf{e}_n = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

is a basis. For, the set is certainly linearly independent because the pattern of 0's and 1's precludes any dependence relation. It also spans \mathbf{R}^n because any vector \mathbf{x} in \mathbf{R}^n can be written

$$\begin{aligned} \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} &= x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \dots + x_n \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \\ &= x_1 \mathbf{e}_1 + x_2 \mathbf{e}_2 + \dots + x_n \mathbf{e}_n. \end{aligned}$$

$\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ is called the *standard basis* for \mathbf{R}^n . Note that in \mathbf{R}^3 the vectors $\mathbf{e}_1, \mathbf{e}_2$, and \mathbf{e}_3 are what we previously called \mathbf{i}, \mathbf{j} , and \mathbf{k} .

In this chapter we have defined the concept dimension only for vector spaces, but the notion is considerably more general. For example, a plane in \mathbf{R}^3 should be considered two dimensional even if it doesn't pass through the origin. Also, a surface in \mathbf{R}^3 , e.g., a sphere or hyperboloid, should also be considered two dimensional. (People are often confused about curved objects because they seem to extend in extra dimensions. The point is that if you look at a small part of a surface, it normally looks like a piece of a plane, so it has the same dimension. Also, as we have seen, a surface can normally be represented parametrically with only two parameters.) Mathematicians have developed a very general theory of dimension which applies to almost any type of set. In cosmology, one envisions the entire universe as a certain type of four dimensional object. Certain bizarre sets can even have a fractional dimension, and that concept is useful in what is called 'chaos' theory. **Coordinates** Let \mathbf{V} be a vector space and suppose $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is a linearly independent subset of \mathbf{V} . Suppose \mathbf{v} is in the subspace spanned by $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$, i.e.,

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n$$

for appropriate coefficients c_1, c_2, \dots, c_n . *The coefficients in such a linear combination are unique.* For, suppose we had

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = c'_1\mathbf{v}_1 + c'_2\mathbf{v}_2 + \cdots + c'_n\mathbf{v}_n.$$

Subtract one expression from the other to obtain

$$(c_1 - c'_1)\mathbf{v}_1 + (c_2 - c'_2)\mathbf{v}_2 + \cdots + (c_n - c'_n)\mathbf{v}_n = 0.$$

We would like to conclude that all these coefficients are zero, i.e., that

$$\begin{aligned} c_1 &= c'_1 \\ c_2 &= c'_2 \\ &\vdots \\ c_n &= c'_n. \end{aligned}$$

If that were not the case, one of the coefficients would be non-zero, and we could divide by it and transpose, thus expressing one of the vectors \mathbf{v}_i as a linear combination of the others. But since, the set is linearly independent, we know that is impossible. Hence, all the coefficients are zero as required.

Suppose $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is a basis for \mathbf{V} . The above argument shows that *any* vector \mathbf{v} in \mathbf{V} may be expressed *uniquely*

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n,$$

and the coefficients c_1, c_2, \dots, c_n are called the *coordinates* of the vector \mathbf{v} with respect to the basis $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$. A convenient way to exhibit the relationship between a vector and its coordinates is as follows. Put the coefficients c_i on the

other side of the basis vectors, and write

$$\mathbf{v} = \mathbf{v}_1 c_1 + \mathbf{v}_2 c_2 + \dots + \mathbf{v}_n c_n = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_n \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}.$$

The column vector on the right is a bona-fide element of \mathbf{R}^n (or of \mathbf{C}^n in the case of complex scalars), but the ‘row vector’ on the left is not really an $n \times 1$ matrix since its entries are vectors, not scalars.

Example 203 Consider the vector space \mathbf{S} of all real solutions of the differential equation

$$y'' + k^2 y = 0.$$

The solutions $y_1 = \cos kt$ and $y_2 = \sin kt$ constitute a linearly independent pair of solutions, so that gives a basis for \mathbf{S} . On the other hand

$$y = \cos(kt + \delta)$$

is also a solution, so it should be expressible as a linear combination of the basis elements. Indeed, by trigonometry, we have

$$\begin{aligned} y &= \cos(kt + \delta) = \cos(kt) \cos \delta - \sin(kt) \sin \delta \\ &= y_1 \cos \delta + y_2 (-\sin \delta) \\ &= \begin{bmatrix} y_1 & y_2 \end{bmatrix} \begin{bmatrix} \cos \delta \\ -\sin \delta \end{bmatrix}. \end{aligned}$$

Thus

$$\begin{bmatrix} \cos \delta \\ -\sin \delta \end{bmatrix}$$

is a vector in \mathbf{R}^2 giving the coordinates of $\cos(kt + \delta)$ with respect to the basis $\{y_1, y_2\}$.

Given a basis for a vector space \mathbf{V} , one may think of the elements of the basis as unit vectors pointing along coordinate axes in \mathbf{V} . The coordinates with respect to the basis then are the coordinates relative to these axes. If one starts (as one does normally in \mathbf{R}^n) with some specific set of axes, then the axes associated with a new basis need not be mutually perpendicular, and also the unit of length may be altered, and we may even have different units of length on each axis.

Invariance of Dimension There is a subtle point involved in the definition of dimension. The dimension of \mathbf{V} is the number of elements in a basis for \mathbf{V} , but it is at least conceivable that two different bases have different numbers of elements. If that were the case, \mathbf{V} would have two different dimensions, and that does not square with our idea of how such words should be used.

In fact it can never happen that two different bases have different numbers of elements. To see this, we shall prove something slightly different. Suppose \mathbf{V} has a basis with m elements. We shall show that

any linearly independent subset of \mathbf{V} has at most m elements.

This would suffice for what we want because if we had two bases one with n and the other with m elements, either could play the role of the basis and the other the role of the linearly independent set. (Any basis is also linearly independent!) Hence, on the one hand we would have $n \leq m$ and on the other hand $m \leq n$, whence it follows that $m = n$.

Here is the proof of the above assertion about linearly independent subsets, (but you might want to skip it your first time through the subject).

Let $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ be a linearly independent subset. Each \mathbf{u}_i can be expressed uniquely in terms of the basis

$$\begin{aligned} \mathbf{u}_1 &= \sum_{j=1}^m \mathbf{v}_j p_{j1} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \dots \quad \mathbf{v}_m] \begin{bmatrix} p_{11} \\ p_{21} \\ \vdots \\ p_{m1} \end{bmatrix} \\ \mathbf{u}_2 &= \sum_{j=1}^m \mathbf{v}_j p_{j2} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \dots \quad \mathbf{v}_m] \begin{bmatrix} p_{21} \\ p_{22} \\ \vdots \\ p_{m2} \end{bmatrix} \\ &\vdots \\ \mathbf{u}_n &= \sum_{j=1}^m \mathbf{v}_j p_{jn} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \dots \quad \mathbf{v}_m] \begin{bmatrix} p_{1n} \\ p_{2n} \\ \vdots \\ p_{mn} \end{bmatrix}. \end{aligned}$$

Each of these equations represents one column of the complete matrix equation

$$[\mathbf{u}_1 \quad \mathbf{u}_2 \quad \dots \quad \mathbf{u}_n] = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \dots \quad \mathbf{v}_m] \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1n} \\ p_{21} & p_{22} & \dots & p_{2n} \\ \vdots & \vdots & \dots & \vdots \\ p_{m1} & p_{m2} & \dots & p_{mn} \end{bmatrix}.$$

Note that the matrix on the right is an $m \times n$ matrix. Consider the homogeneous system

$$\begin{bmatrix} p_{11} & p_{12} & \dots & p_{1n} \\ p_{21} & p_{22} & \dots & p_{2n} \\ \vdots & \vdots & \dots & \vdots \\ p_{m1} & p_{m2} & \dots & p_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = 0.$$

Assume, contrary to what we hope, that $n > m$. Then, we know by the theory of homogeneous linear systems, that there is a non-trivial solution to this system, i.e., one with at least one x_i not zero. Then

$$\begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_m \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = 0.$$

Thus, 0 has a non-trivial representation

$$0 = \mathbf{u}_1 x_1 + \mathbf{u}_2 x_2 + \cdots + \mathbf{u}_n x_n$$

which we know can never happen for a linearly independent set. Thus, the only way out of this contradiction is to believe that $n \leq m$ as claimed.

One consequence of this argument is the following fact. *The dimension of a subspace cannot be larger than the dimension of the whole vector space.* The reasoning is that a basis for a subspace is necessarily a linearly independent set and so it cannot have more elements than the dimension of the whole vector space.

It is important to note that two different bases of the same vector space might have no elements whatsoever in common. All we can be sure of is that they have the same size.

Infinite Dimensional Vector Spaces Not every vector space has a finite basis. We shall not prove it rigorously here, but it is fairly clear that the vector space of all continuous functions $\mathcal{C}(I)$ cannot have a finite basis $\{f_1, f_2, \dots, f_n\}$. For, if it did, then that would mean *any* continuous function f on I could be written as a finite linear combination

$$f(t) = c_1 f_1(t) + c_2 f_2(t) + \cdots + c_n f_n(t),$$

and it is not plausible that any finite set of continuous functions could capture the full range of possibilities of continuous functions in this way.

If a vector space has a finite basis, we say that it is *finite dimensional*; otherwise we say it is *infinite dimensional*.

Most interesting function spaces are infinite dimensional. Fortunately, the subspaces of these spaces which are solutions of homogeneous linear differential equations are finite dimensional, and these are what we shall spend the next chapter studying.

We won't talk much about infinite dimensional vector spaces in this course, but you will see them again in your course on Fourier series and partial differential equations, and you will also encounter such spaces when you study quantum mechanics.

Exercises for 10.8.

1. In each of the following cases, determine if the indicated set is linearly independent or not.

$$(a) \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \right\}.$$

$$(b) \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \right\}.$$

2. Determine if each of the following sets of functions is linearly independent.

$$(a) \{1, t, t^2, t^3\}.$$

$$(b) \{e^{-t}, e^{2t}, e^t\}.$$

$$(c) \{\cos t \sin t, \sin 2t, \cos 2t\}.$$

3. Let \mathbf{V} be the vector space of all polynomials of degree at most 2. (You may assume it is known that \mathbf{V} is a vector space.)

$$(a) \text{ Show that } \{1, t, t^2\} \text{ is a basis for } \mathbf{V} \text{ so that the dimension of } \mathbf{V} \text{ is 3.}$$

$$(b) \text{ The first three Legendre polynomials (solutions of Legendre's equation } (1-t^2)y'' - 2ty' + \alpha(\alpha+1)y = 0 \text{ for } \alpha = 0, 1, 2) \text{ are } P_0(t) = 1, P_1(t) = t, \text{ and } P_2(t) = \frac{1}{2}(3t^2 - 1). \text{ Show that } \{P_0, P_1, P_2\} \text{ is a linearly independent set of functions. It follows that it is a basis. Why?}$$

4. Find a basis for the solution space of the differential equation $y'' - k^2y = 0$.
5. Find a basis for the subspace of \mathbf{R}^4 consisting of solutions of the homogeneous system

$$\begin{bmatrix} 1 & -1 & 1 & -1 \\ 1 & 2 & -1 & 1 \\ 0 & 3 & -2 & 2 \end{bmatrix} \mathbf{x} = 0.$$

6. Find the dimension of the solution space of $A\mathbf{x} = 0$ in each of the following cases. (See the Exercises for Section 6.)

$$(a) A = \begin{bmatrix} 1 & 0 & 1 & 2 \\ 2 & -1 & 1 & 0 \\ -1 & 4 & -1 & -2 \end{bmatrix} \quad (b) A = \begin{bmatrix} 1 & 3 & 4 & 0 & 2 \\ 2 & 7 & 6 & 1 & 1 \\ 4 & 13 & 14 & 1 & 3 \end{bmatrix}$$

7. Show that a set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is linearly independent if and only if the equation

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = \mathbf{0}$$

has only the solution $c_1 = c_2 = \cdots = c_n = 0$.

8. Let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ be a subset of a vector space \mathbf{V} . Show that the set is linearly independent if and only if the equation

$$0 = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n$$

has only the trivial solution, i.e., all the coefficients $c_1 = c_2 = \cdots = c_n = 0$.

This characterization is very convenient to use when proving a set is linearly independent. It is often taken as the *definition* of linear independence in books on linear algebra.

9. (Optional.) It is assumed implicitly at various points in our development that interesting vector spaces do in fact have bases. In most cases, we have an explicit method for constructing a basis, but this is not always possible.

(a) Let \mathbf{V} be a finite dimensional vector space with basis $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$, and let \mathbf{W} be a subspace. Show that \mathbf{W} has a finite basis. Hint. Construct a sequence of elements in \mathbf{W} as follows. Start by choosing $\mathbf{w}_1 \neq 0$ in \mathbf{W} . Assume $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k$ have been chosen in \mathbf{W} so that $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k\}$ is a linearly independent set. Show that either this set is a basis for \mathbf{W} or it is possible to choose \mathbf{w}_{k+1} in \mathbf{W} such that $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k, \mathbf{w}_{k+1}\}$ is linearly independent. (This is a bit harder than you might think.) This process can't go on forever. Look at the discussion of invariance of dimension to see why not.

Note that the argument won't work if \mathbf{W} is the zero subspace. What is the basis in that case?

(b) Use part (a) to conclude that any subspace of \mathbf{R}^n (or \mathbf{C}^n in the complex case) has a finite basis.

10. The set of all infinite sequences

$$\mathbf{x} = (x_1, x_2, \dots, x_n, \dots)$$

forms a vector space if two sequences are added by adding corresponding entries and a sequence is multiplied by a scalar by multiplying each entry by that scalar. If the scalars are assumed to be real, it would be appropriate to denote this vector space \mathbf{R}^∞ . Let \mathbf{e}_i be the vector with $x_i = 1$ and all other entries zero.

(a) Show that the set $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ of the first n of these is a linearly independent set for each n . Thus there is no upper bound on the size of a linearly independent subset of \mathbf{R}^∞ .

(b) Does the set of all possible \mathbf{e}_i span \mathbf{R}^∞ ? Explain.

11. (Optional) We know from our previous work that $\{e^{ikt}, e^{-ikt}\}$ is a basis for the set of complex valued solutions of the differential equation $y'' + k^2y = 0$. However, $\{\cos kt, \sin kt\}$ is also a basis for that complex vector space. What are the coordinates of e^{ikt} and e^{-ikt} with respect to the second basis? What are the coordinates of $\cos kt$ and $\sin kt$ with respect to the first basis? In each case express the answers as column vectors in \mathbf{C}^2 .

10.9 Calculations in \mathbf{R}^n or \mathbf{C}^n

Let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ be a collection of vectors in \mathbf{R}^n (or \mathbf{C}^n in the case of complex scalars.) It is often useful to have a way to pick out a linearly independent *subset* which spans the same subspace, i.e., which is a basis for that subspace. The basic idea (no pun intended) is to throw away redundant vectors until that is no longer possible, but there is a systematic way to do this all at once. Since the vectors \mathbf{v}_i are elements of \mathbf{R}^n , each may be realized as a $n \times 1$ column vector. Put these vectors together to form an $n \times k$ matrix

$$A = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \dots \quad \mathbf{v}_k].$$

(This is the same notation we used in the previous section, but now since the \mathbf{v}_i are column vectors, rather than elements of some abstract vector space, we really do get a matrix.) To find a basis, apply Gaussian reduction to the matrix A , and pick out the columns of A which in the transformed reduced matrix end up with pivots.

Example 204 Let

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 2 \\ 2 \\ 4 \\ 0 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} -1 \\ 1 \\ 0 \\ -2 \end{bmatrix}, \quad \mathbf{v}_4 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}.$$

Form the matrix A with these columns and apply Gaussian reduction

$$\begin{aligned} \begin{bmatrix} 1 & 2 & -1 & 0 \\ 0 & 2 & 1 & 1 \\ 1 & 4 & 0 & 1 \\ 1 & 0 & -2 & 0 \end{bmatrix} &\rightarrow \begin{bmatrix} 1 & 2 & -1 & 1 \\ 0 & 2 & 1 & 1 \\ 0 & 2 & 1 & 1 \\ 0 & -2 & -1 & 0 \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 1 & 2 & -1 & 1 \\ 0 & 2 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 1 & 2 & -1 & 1 \\ 0 & 2 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

This completes the Gaussian reduction, and the pivots are in the first, second, and fourth columns. Hence, the vectors

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 2 \\ 2 \\ 4 \\ 0 \end{bmatrix}, \quad \mathbf{v}_4 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

form a basis for the subspace spanned by $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$.

Let's look more closely at this example to see why the subset is linearly independent and also spans the same subspace as the original set. The proof that the algorithm works in the general case is more complicated to write down but just elaborates the ideas exhibited in the example. Consider the homogeneous system $A\mathbf{x} = 0$. This may also be written

$$A\mathbf{x} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 & \mathbf{v}_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \mathbf{v}_1x_1 + \mathbf{v}_2x_2 + \mathbf{v}_3x_3 + \mathbf{v}_4x_4 = 0.$$

In the general solution, x_1, x_2 , and x_4 will be bound variables (from the pivot positions) and x_3 will be free. That means we can set x_3 equal to anything, say $x_3 = -1$ and the other variables will be determined. For this choice, the relation becomes

$$\mathbf{v}_1x_1 + \mathbf{v}_2x_2 - \mathbf{v}_3 + \mathbf{v}_4x_4 = 0$$

which may be rewritten

$$\mathbf{v}_3 = x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + x_4\mathbf{v}_4.$$

Thus, \mathbf{v}_3 is redundant and may be eliminated from the set without changing the subspace spanned by the set. On the other hand, the set $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_4\}$ is linearly independent, since if we were to apply Gaussian reduction to the matrix

$$A' = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_4 \end{bmatrix}$$

the reduced matrix would have a pivot in every column, i.e., it would have rank 3. Thus, the system

$$\begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_4 \end{bmatrix} = \mathbf{v}_1x_1 + \mathbf{v}_2x_2 + \mathbf{v}_4x_4 = 0$$

has only the trivial solution. That means that no one of the three vectors can be expressed as a linear combination of the other two. For example, if $\mathbf{v}_2 = c_1\mathbf{v}_1 + c_4\mathbf{v}_4$, we have

$$\mathbf{v}_1c_1 + \mathbf{v}_2(-1) + \mathbf{v}_4c_4 = 0.$$

It follows that the set is linearly independent. **Column Space and Row Space**

Let A be an $m \times n$ matrix. Then the columns $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ of A are vectors in

\mathbf{R}^m (or \mathbf{C}^m in the complex case), and $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ spans a subspace of \mathbf{R}^m called the *column space* of A . The column space plays a role in the theory of inhomogeneous systems $A\mathbf{x} = \mathbf{b}$ in the following way. A vector \mathbf{b} is in the column space if and only if it is expressible as a linear combination

$$\mathbf{b} = \mathbf{v}_1x_1 + \mathbf{v}_2x_2 + \cdots + \mathbf{v}_nx_n = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = A\mathbf{x}.$$

Thus, *the column space of A consists of all vectors \mathbf{b} in \mathbf{R}^m for which the system $A\mathbf{x} = \mathbf{b}$ has a solution.*

Note that the method outlined in the beginning of this section gives a basis for the column space, and the number of elements in this basis is the rank of A . (The rank is the number of pivots!) Hence, *the rank of an $m \times n$ matrix A is the dimension of its column space.*

There is a similar concept for rows; the row space of an $m \times n$ matrix A is the subspace of \mathbf{R}^n spanned by the rows of A . It is not hard to see that *the dimension of the row space of A is also the rank of A .* For, since each row operation is reversible, applying a row operation does not change the subspace spanned by the rows. Hence, the row space of the matrix A' obtained by Gauss-Jordan reduction from A is the same as the row space of A . However, the set of non-zero rows of the reduced matrix is a basis for this subspace. To see this, note first that it certainly spans (since leaving out zero rows doesn't cost us anything). Moreover, it is also a linearly independent set because each non-zero row has a 1 in a pivot position where all the other rows are zero.

The fact that both the column space and the row space have the same dimension is sometimes expressed by saying "the column rank equals the row rank".

The column space also has an abstract interpretation. Consider the linear operator $L : \mathbf{R}^n \rightarrow \mathbf{R}^m$ defined by $L(\mathbf{x}) = A\mathbf{x}$. The *image* of this operator is defined to be the set of all vectors \mathbf{b} in \mathbf{R}^m of the form $\mathbf{b} = L(\mathbf{x})$ for some \mathbf{x} in \mathbf{R}^n . Thus, the image of L is just the set of $\mathbf{b} = A\mathbf{x}$, which by the above reasoning is the column space of A , and its dimension is the rank r . On the other hand, you should recall that the dimension of the null space of L is the number of basic solutions, i.e., $n - r$. Since these add up to n , we have

$$\dim \text{Image of } L + \dim \text{Null Space of } L = \dim \text{Domain of } L.$$

A Note on the Definition of Rank The rank of A is defined as the number of pivots in the reduced matrix obtained from A by an appropriate sequence of elementary row operations. Since we can specify a standard procedure for performing such row operations, that means the rank is a well defined number. On the other hand,

it is natural to wonder what might happen if A were reduced by an alternative, perhaps less systematic, sequence of row operations. The above analysis shows that we would still get the same answer for the rank. Namely, the rank is the dimension of the column space of A , and that number depends only on the column space itself, not on any particular basis for it. (Or you could use the same argument using the row space.)

The rank is also the number of non-zero rows in the reduced matrix, so it follows that this number does not depend on the particular sequence of row operations used to reduce A to Gauss-Jordan reduced form. In fact, the entire matrix obtained at the end (as long as it is in Gauss-Jordan reduced form) depends only on the original matrix A and not on the particular sequence of row operations used to obtain it. The proof of this fact is not so easy, and we omit it here.

Exercises for 10.9.

- Find a subset of the following set of vectors which is a basis for the subspace it spans.

$$\left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \\ -3 \\ 2 \end{bmatrix}, \begin{bmatrix} 3 \\ 3 \\ 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ -3 \\ 1 \end{bmatrix} \right\}$$

$$2. \text{ Let } A = \begin{bmatrix} 1 & 0 & 2 & 1 & 1 \\ -1 & 1 & 3 & 0 & 1 \\ 1 & 1 & 7 & 2 & 3 \end{bmatrix}.$$

- Find a basis for the column space of A .
- Find a basis for the row space of A .

- Let

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ -2 \\ -1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}.$$

Find a basis for \mathbf{R}^3 by finding a third vector \mathbf{v}_3 such that $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is linearly independent. Hint. You may find an easier way to do it, but the following method should work. Use the method suggested in Section 9 to pick out a linearly independent subset from $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$.

- Let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ be a linearly independent subset of \mathbf{R}^n . Apply the method in section 9 to the set $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k, \mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$. It will necessarily yield a basis for \mathbf{R}^n . Why? Show that this basis will include $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ as a subset. That is show that none of the \mathbf{v}_i will be eliminated by the process.

Note. If \mathbf{V} is any finite dimensional vector space with basis $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ and $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ is a linearly independent subset of \mathbf{V} , then we may

form another basis for \mathbf{V} by adding *some* of the vectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ to $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$. However, since we aren't necessarily dealing with column vectors in this case, the proof is a bit more involved.

5. Show that

$$\mathbf{u}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad \text{and} \quad \mathbf{u}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

form a linearly independent pair in \mathbf{R}^2 . It follows that they form a basis for \mathbf{R}^2 . Why? Find the coordinates of \mathbf{e}_1 and \mathbf{e}_2 with respect to this new basis. Hint. You need to solve

$$\begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{e}_1 \quad \text{and} \quad \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{e}_2.$$

You can solve these simultaneously by solving

$$\begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{bmatrix} X = I$$

for an appropriate 2×2 matrix X . What does this have to do with inverses?

Chapter 11

Determinants and Eigenvalues

11.1 Homogeneous Linear Systems of Differential Equations

Let $A = A(t)$ denote an $n \times n$ matrix with entries $a_{ij}(t)$ which are functions defined and continuous on some real interval $a < t < b$. In general, these functions may be complex valued functions. Consider the $n \times n$ system of differential equations

$$\frac{d\mathbf{x}}{dt} = A(t)\mathbf{x}$$

where $\mathbf{x} = \mathbf{x}(t)$ is a vector valued function also defined on $a < t < b$ and taking values in \mathbf{C}^n . (If $A(t)$ happens to have real entries, then we may consider solutions $\mathbf{x} = \mathbf{x}(t)$ with values in \mathbf{R}^n , but even in that case it is often advantageous to consider complex valued solutions.) The most interesting case is that in which A is a *constant* matrix, and we shall devote almost all our attention to that case.

Example 205 Consider

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \mathbf{x} \quad \text{where } \mathbf{x} = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

on the interval $-\infty < t < \infty$. Note that the entries in the coefficient matrix are constants.

The set of solutions of such a homogeneous system forms a (complex) vector space. To see this, consider the operator

$$L = \frac{d}{dt} - A$$

which is defined on the vector space of all differentiable vector valued functions. It is not hard to see that L is a linear operator, and its null space is the desired set of solutions since

$$L(\mathbf{x}) = 0 \quad \text{means} \quad \frac{d\mathbf{x}}{dt} - A\mathbf{x} = 0.$$

We shall see shortly that *this vector space is n -dimensional*. Hence, solving the system amounts to finding n solutions $\mathbf{x}_1(t), \mathbf{x}_2(t), \dots, \mathbf{x}_n(t)$ which form a basis for the solution space. That means that any solution can be written uniquely

$$\mathbf{x} = c_1\mathbf{x}_1(t) + c_2\mathbf{x}_2(t) + \dots + c_n\mathbf{x}_n(t).$$

Moreover, if the solution has a specified initial value $\mathbf{x}(t_0)$, then the c_i are determined by solving

$$\mathbf{x}(t_0) = c_1\mathbf{x}_1(t_0) + c_2\mathbf{x}_2(t_0) + \dots + c_n\mathbf{x}_n(t_0).$$

(If you look closely, you will see that this is in fact a linear system of algebraic equations with unknowns c_1, c_2, \dots, c_n .)

Example 205, continued We shall try to solve

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \mathbf{x} \quad \text{given} \quad \mathbf{x}(0) = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

The first problem is to find a linearly independent pair solutions. Assume for the moment that you have a method for generating such solutions, and it tells you to try

$$\mathbf{x}_1 = \begin{bmatrix} e^{3t} \\ e^{3t} \end{bmatrix} = e^{3t} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} -e^{-t} \\ e^{-t} \end{bmatrix} = e^{-t} \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

(We shall develop such methods later in this chapter.) It is not hard to see that these are solutions:

$$\begin{aligned} \frac{d\mathbf{x}_1}{dt} &= 3e^{3t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} & \text{and} & \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \mathbf{x}_1 = e^{3t} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = e^{3t} \begin{bmatrix} 3 \\ 3 \end{bmatrix} \\ \frac{d\mathbf{x}_2}{dt} &= -e^{-t} \begin{bmatrix} -1 \\ 1 \end{bmatrix} & \text{and} & \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \mathbf{x}_2 = e^{-t} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = e^{-t} \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \end{aligned}$$

Also, $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ form a linearly independent pair. For, otherwise one would be a scalar multiple of the other, say $\mathbf{x}_1(t) = c\mathbf{x}_2(t)$ for all t . Then the same thing would be true of their first components, e.g., $e^{3t} = ce^{-t}$ for all t , and we know that to be false.

Thus, $\{\mathbf{x}_1, \mathbf{x}_2\}$ is a basis for the solution space, so any solution may be expressed uniquely

$$\mathbf{x} = c_1\mathbf{x}_1(t) + c_2\mathbf{x}_2(t) = \mathbf{x}_1(t)c_1 + \mathbf{x}_2(t)c_2 = \begin{bmatrix} \mathbf{x}_1(t) & \mathbf{x}_2(t) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}.$$

Thus putting $t = 0$ yields

$$\mathbf{x}(0) = \begin{bmatrix} \mathbf{x}_1(0) & \mathbf{x}_2(0) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

However,

$$\mathbf{x}_1(0) = e^{3(0)} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \mathbf{x}_2(0) = e^{-(0)} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix},$$

so the above system becomes

$$\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

This system is easy to solve by Gauss-Jordan reduction

$$\left[\begin{array}{cc|c} 1 & -1 & 1 \\ 1 & 1 & -2 \end{array} \right] \rightarrow \left[\begin{array}{cc|c} 1 & -1 & 1 \\ 0 & 2 & -3 \end{array} \right] \rightarrow \left[\begin{array}{cc|c} 1 & 0 & -1/2 \\ 0 & 1 & -3/2 \end{array} \right].$$

Hence, the solution is $c_1 = -1/2$, $c_2 = -3/2$ and the desired solution of the system of differential equations is

$$\mathbf{x} = -\frac{1}{2}e^{3t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \frac{3}{2}e^{-t} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}e^{3t} + \frac{3}{2}e^{-t} \\ -\frac{1}{2}e^{3t} - \frac{3}{2}e^{-t} \end{bmatrix}.$$

The general situation is quite similar. Suppose we have some method for generating n vector valued functions $\mathbf{x}_1(t), \mathbf{x}_2(t), \dots, \mathbf{x}_n(t)$ which together constitute a linearly independent set of solutions of the $n \times n$ system

$$\frac{d\mathbf{x}}{dt} = A\mathbf{x}.$$

Then, the general solution takes the form

$$\mathbf{x} = \mathbf{x}_1(t)c_1 + \mathbf{x}_2(t)c_2 + \dots + \mathbf{x}_n(t)c_n = \begin{bmatrix} \mathbf{x}_1(t) & \mathbf{x}_2(t) & \dots & \mathbf{x}_n(t) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}.$$

To match a given initial condition $\mathbf{x}(t_0)$ at $t = t_0$, we have to solve the $n \times n$ algebraic system

$$\begin{bmatrix} \mathbf{x}_1(t_0) & \mathbf{x}_2(t_0) & \dots & \mathbf{x}_n(t_0) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \mathbf{x}(t_0). \quad (169)$$

Note that the coefficient matrix is just a specific $n \times n$ matrix with scalar entries and the quantity on the right is a specific $n \times 1$ column vector. (Everything in

sight is evaluated at t_0 , so there are no variable quantities in this equation.) We now have to rely on the results of the previous chapter. Since in principle the given initial value vector $\mathbf{x}(t_0)$ could be anything whatsoever, we hope that this algebraic system can be solved for any possible $n \times 1$ column vector on the right. However, we know this is possible only in the case that the coefficient matrix is non-singular, i.e., if it has rank n . How can we be sure of this? It turns out to be a consequence of the basic uniqueness theorem for systems of differential equations.

Existence and Uniqueness for Systems We state the basic theorem for complex valued functions. There is a corresponding theorem in the real case.

Theorem 11.16 Let $A(t)$ be an $n \times n$ complex matrix with entries continuous functions defined on a real interval $a < t < b$, and suppose t_0 is a point in that interval. Let \mathbf{x}_0 be a given $n \times 1$ column vector in \mathbf{C}^n . Then there exists a unique solution $\mathbf{x} = \mathbf{x}(t)$ of the system

$$\frac{d\mathbf{x}}{dt} = A(t)\mathbf{x}$$

defined on the interval $a < t < b$ and satisfying $\mathbf{x}(t_0) = \mathbf{x}_0$.

We shall not try to prove this theorem in this course.

The uniqueness part of the theorem has the following important consequence.

Corollary 11.17 With the notation as above, suppose $\mathbf{x}_1(t), \mathbf{x}_2(t), \dots, \mathbf{x}_n(t)$ are n solutions of the system $\frac{d\mathbf{x}}{dt} = A(t)\mathbf{x}$ on the interval $a < t < b$. Let t_0 be any point in that interval. Then the set $\{\mathbf{x}_1(t), \mathbf{x}_2(t), \dots, \mathbf{x}_n(t)\}$ is a linearly independent set of functions if and only if the set $\{\mathbf{x}_1(t_0), \mathbf{x}_2(t_0), \dots, \mathbf{x}_n(t_0)\}$ is a linearly independent set of column vectors in \mathbf{C}^n .

Example 1, again The set of functions is

$$\left\{ e^{3t} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, e^{-t} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}$$

and the set of column vectors (obtained by setting $t = t_0 = 0$) is

$$\left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}.$$

Proof. We shall prove that one set is linearly *dependent* if and only if the other is. First suppose that the set of functions is dependent. That means that one of them may be expressed as a linear combination of the others. For the sake of argument suppose the notation is arranged so that

$$\mathbf{x}_1(t) = c_2\mathbf{x}_2(t) + \dots + c_n\mathbf{x}_n(t) \tag{170}$$

holds for all t in the interval. Then, it holds for $t = t_0$ and we have

$$\mathbf{x}_1(t_0) = c_2\mathbf{x}_2(t_0) + \cdots + c_n\mathbf{x}_n(t_0). \quad (171)$$

This in turn tells us that the set of column vectors at t_0 is dependent.

Suppose on the other hand that the set of column vectors at t_0 is dependent. Then we may assume that there is a relation of the form (171) with appropriate scalars c_2, \dots, c_n . But that means that the solutions $\mathbf{x}_1(t)$ and $c_2\mathbf{x}_2(t) + \cdots + c_n\mathbf{x}_n(t)$ agree at $t = t_0$. According to the uniqueness part of Theorem 11.16, this means that they agree for all t , which means that (170) is true as a relation among functions. This in turn implies that the set of functions is dependent. \square \square

The corollary gives us what we need to show that the $n \times n$ matrix

$$[\mathbf{x}_1(t_0) \quad \mathbf{x}_2(t_0) \quad \cdots \quad \mathbf{x}_n(t_0)]$$

is non-singular. Namely, if $\{\mathbf{x}_1(t), \mathbf{x}_2(t), \dots, \mathbf{x}_n(t)\}$ is a linearly independent set of solutions, then the columns of the above matrix form a linearly independent set in \mathbf{C}^n . It follows that they form a basis for the column space of the matrix which must then have rank n , and so it is non-singular. As noted above, that means that we can always solve the algebraic system of equations (169) specifying initial conditions at t_0 .

Exercises for 11.1.

1. Verify that

$$\mathbf{x}_1 = e^{0t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{x}_2 = e^{-2t} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

are solutions of the 2×2 system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{x}.$$

Find the solution satisfying $\mathbf{x}(0) = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$.

2. Show that

$$\mathbf{x}_1 = e^t \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = e^{2t} \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}, \quad \mathbf{x}_3 = e^{-2t} \begin{bmatrix} 1 \\ -2 \\ 4 \end{bmatrix}$$

are solutions of

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -4 & 4 & 1 \end{bmatrix} \mathbf{x}. \quad (172)$$

Show that they form a linearly independent set of solutions by finding the rank of $[\mathbf{x}_1(0) \quad \mathbf{x}_2(0) \quad \mathbf{x}_3(0)]$. Finally, find the solution of (172) satisfying

$$\mathbf{x}(0) = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}.$$

3. The vector functions

$$\mathbf{x}_1 = e^{2t} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{x}_2 = e^{3t} \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}, \quad \mathbf{x}_3 = e^{-t} \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

are solutions of a system of the form $d\mathbf{x}/dt = A\mathbf{x}$ where A is an appropriate 3×3 constant matrix. Do these functions constitute a basis for the solution space of that system?

11.2 Finding Linearly Independent Solutions

Existence of a Basis The basic existence and uniqueness theorem stated in the previous section ensures that a system of the form

$$\frac{d\mathbf{x}}{dt} = A(t)\mathbf{x}$$

always has n solutions $\mathbf{u}_1(t), \mathbf{u}_2(t), \dots, \mathbf{u}_n(t)$ which form a basis for the vector space of all solutions.

To see this, fix t_0 in the interval $a < t < b$, and define the i th solution $\mathbf{u}_i(t)$ to be the *unique* solution satisfying the initial condition

$$\mathbf{u}_i(t_0) = \mathbf{e}_i$$

where as before \mathbf{e}_i is the i th vector in the *standard basis* for \mathbf{C}^n , i.e., it is the i th column of the $n \times n$ identity matrix. Since $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ is an independent set, it follows from Corollary 11.2

that $\{\mathbf{u}_1(t), \mathbf{u}_2(t), \dots, \mathbf{u}_n(t)\}$ is an independent set of solutions. It also spans the subspace of solutions. To see this, let $\mathbf{x}(t)$ denote any solution. At $t = t_0$ we have

$$\begin{aligned} \mathbf{x}(t_0) = \begin{bmatrix} x_1(t_0) \\ x_2(t_0) \\ \vdots \\ x_n(t_0) \end{bmatrix} &= x_1(t_0)\mathbf{e}_1 + x_2(t_0)\mathbf{e}_2 + \cdots + x_n(t_0)\mathbf{e}_n \\ &= c_1\mathbf{u}_1(t_0) + c_2\mathbf{u}_2(t_0) + \cdots + c_n\mathbf{u}_n(t_0), \end{aligned}$$

where $c_1 = x_1(t_0), c_2 = x_2(t_0), \dots, c_n = x_n(t_0)$. Thus, by the uniqueness theorem, we have for all t in the interval

$$\mathbf{x}(t) = c_1 \mathbf{u}_1(t) + c_2 \mathbf{u}_2(t) + \cdots + c_n \mathbf{u}_n(t).$$

Case of Constant Coefficients It is reassuring to know that in principle we can always find a basis for the solution space, but that doesn't help us find it. In this section we outline a method for generating a basis for the solutions of $\frac{d\mathbf{x}}{dt} = A\mathbf{x}$ in case A is *constant*. This method will work reasonably well for 2×2 systems, but we shall have to develop the theory of $n \times n$ determinants to get it to work for general $n \times n$ systems.

Example 206 Consider the 2×2 system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 0 & -2 \\ 1 & 3 \end{bmatrix} \mathbf{x}. \quad (173)$$

I chose this because it is just the system version of the second order equation

$$y'' - 3y' + 2y = 0 \quad (174)$$

which we already know how to solve. Namely, put $x_1 = y$ and $x_2 = y'$, so

$$\begin{aligned} x_1' &= y' & &= x_2 \\ x_2' &= y'' = -2y + 3y' & &= -2x_1 + 3x_2, \end{aligned}$$

which when put in matrix form is (173). To solve the second order equation (174), proceed in the usual manner. The roots of

$$r^2 - 3r + 2 = (r - 1)(r - 2) = 0$$

are $r_1 = 1, r_2 = 2$. Hence, $y_1 = e^t$ and $y_2 = e^{2t}$ constitute a linearly independent pair of solutions. The corresponding *vector* solutions are

$$\mathbf{x}_1 = \begin{bmatrix} y_1 \\ y_1' \end{bmatrix} = \begin{bmatrix} e^t \\ e^t \end{bmatrix} = e^t \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} y_2 \\ y_2' \end{bmatrix} = \begin{bmatrix} e^{2t} \\ 2e^{2t} \end{bmatrix} = e^{2t} \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

The above example suggests that we *look for solutions of the form*

$$\mathbf{x} = e^{\lambda t} \mathbf{v} \quad (175)$$

where λ is a scalar and \mathbf{v} is a vector, both to be determined by the solution process. Note also that we want $\mathbf{v} \neq \mathbf{0}$ since otherwise the solution of the differential equation would be identically zero and hence not very interesting.

Substitute (175) in $\frac{d\mathbf{x}}{dt} = A\mathbf{x}$ to obtain

$$\frac{d\mathbf{x}}{dt} = \lambda e^{\lambda t} \mathbf{v} = A e^{\lambda t} \mathbf{v}.$$

The factor $e^{\lambda t}$ is a scalar and non-zero for all t , so we cancel it from the above equation. The resulting equation may be rewritten

$$A\mathbf{v} = \lambda\mathbf{v} \quad \text{where } \mathbf{v} \neq \mathbf{0}. \quad (176)$$

We introduce special terminology for the situation described by this equation. If (176) has a non-zero solution for a given scalar λ , then λ is called an *eigenvalue* of the matrix A , and any *non-zero* vector \mathbf{v} which works for that eigenvalue is called an *eigenvector* of A corresponding to λ . These related concepts are absolutely essential for understanding systems of differential equations, and they arise in fundamental ways in a wide variety of applications of linear algebra.

Let's analyze the problem of finding eigenvalues and eigenvectors for A a 2×2 matrix. Then, (176) may be rewritten

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} \lambda v_1 \\ \lambda v_2 \end{bmatrix}$$

or

$$\begin{aligned} a_{11}v_1 + a_{12}v_2 &= \lambda v_1 \\ a_{21}v_1 + a_{22}v_2 &= \lambda v_1, \end{aligned}$$

which, after transposing, becomes

$$\begin{aligned} (a_{11} - \lambda)v_1 + a_{12}v_2 &= 0 \\ a_{21}v_1 + (a_{22} - \lambda)v_2 &= 0. \end{aligned}$$

This is a homogeneous system which may be put in matrix form

$$\begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \mathbf{0}. \quad (177)$$

This system will have a *non-zero* solution \mathbf{v} , as required, if and only if the coefficient matrix has rank less than $n = 2$. Unless the matrix consists of zeroes, this means it must have rank one. That, in turn, amounts to saying that one of the rows is a multiple of the other, i.e., that the ratios of corresponding components are the same, or, in symbols,

$$\frac{a_{11} - \lambda}{a_{21}} = \frac{a_{12}}{a_{22} - \lambda}.$$

Cross multiplying and transposing yields the quadratic equation

$$(a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0. \quad (178)$$

Our strategy then is to solve this equation for λ to find the possible *eigenvalues* and then for each eigenvalue λ to find the non-zero solutions of (177) which are the eigenvectors corresponding to that eigenvalue. In this way, each eigenvalue and eigenvector pair will generate a solution $\mathbf{x} = e^{\lambda t}\mathbf{v}$ of the original system of differential equations.

Note that (178) may be rewritten using 2×2 determinants as

$$\det \begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} = 0. \quad (179)$$

This equation is called the *characteristic equation* of the matrix.

Example 207 Consider (as in Section 1) the system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \mathbf{x}. \quad (180)$$

Try $\mathbf{x} = e^{\lambda t} \mathbf{v}$ as above. As we saw, this comes down to solving the eigenvalue–eigenvector problem for the coefficient matrix

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}.$$

To do so, we first solve the characteristic equation

$$\begin{aligned} \det \begin{bmatrix} 1 - \lambda & 2 \\ 2 & 1 - \lambda \end{bmatrix} &= (1 - \lambda)^2 - 4 = 0 \\ \text{or } 1 - 2\lambda + \lambda^2 - 4 &= \lambda^2 - 2\lambda - 3 = (\lambda - 3)(\lambda + 1) = 0. \end{aligned}$$

The roots of this equation are $\lambda = 3$ and $\lambda = -1$. First consider $\lambda = 3$. Putting this in (177) yields

$$\begin{bmatrix} 1 - 3 & 2 \\ 2 & 1 - 3 \end{bmatrix} \mathbf{v} = \begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix} \mathbf{v} = \mathbf{0}. \quad (181)$$

Gauss-Jordan reduction yields the solution $v_1 = v_2$ with v_2 free. A general solution vector has the form

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} v_2 \\ v_2 \end{bmatrix} = v_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Put $v_2 = 1$ to obtain

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

which will form a basis for the solution space of (181). Thus, we obtain as one solution of (180)

$$\mathbf{x}_1 = e^{\lambda t} \mathbf{v}_1 = e^{3t} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Note that any other eigenvector for $\lambda = 3$ is a non-zero multiple of the basis vector \mathbf{v}_1 , so choosing another eigenvector in this case will result in a solution of the differential equation which is just a constant multiple of \mathbf{x}_1 .

To find a second solution, consider the root $\lambda = -1$. Put this in (177) to obtain

$$\begin{bmatrix} 1 - (-1) & 2 \\ 2 & 1 - 3 \end{bmatrix} \mathbf{v} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \mathbf{v} = \mathbf{0}. \quad (182)$$

Gauss-Jordan reduction yields the general solution $v_1 = -v_2$ with v_2 free. The general solution vector is

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = v_2 \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Putting $v_2 = 1$ yields a the basic eigenvector

$$\mathbf{v}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

and the corresponding solution of the differential equation

$$\mathbf{x}_2 = e^{-t} \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Note that these are the solutions we used to form a basis in Example 1 in the previous section.

The above procedure appears similar to what we did to solve a second order equation $y'' + py' + qy = 0$ with constant coefficients. This is no accident!. The quadratic equation

$$r^2 + pr + q = 0$$

is just the characteristic equation (with r replacing λ) of the 2×2 matrix you obtain when you reformulate the problem as a first order system. You should check this explicitly in Example 206. (The general case is the topic of an exercise.)

The method for $n \times n$ systems is very similar to what we did above. However, the analogue of (179), i.e., the characteristic equation, takes the form

$$\det \begin{bmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{bmatrix} = 0$$

which requires the use of $n \times n$ determinants. So far in this course we have only discussed 2×2 determinants and briefly 3×3 determinants. Hence, to develop the general theory, we need to define and study the properties of $n \times n$ determinants.

Exercises for 11.2.

1. In each of the following examples, try to find a linearly independent pair of solutions of the 2×2 system $d\mathbf{x}/dt = A\mathbf{x}$ by the method outlined in this section. It may not be possible to do so in all cases.

(a) $A = \begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix}.$

$$(b) A = \begin{bmatrix} -2 & 0 \\ 1 & -1 \end{bmatrix}.$$

$$(c) A = \begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix}.$$

$$(d) A = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

2. Show that for the second order system $y'' + py' + qy = 0$, the characteristic equation of the corresponding matrix $A = \begin{bmatrix} 0 & 1 \\ -q & -p \end{bmatrix}$ is $\lambda^2 + p\lambda + q = 0$.

This helps us to subsume the theory of second order equations under that of systems.

11.3 Definition of the Determinant

Let A be an $n \times n$ matrix.

By definition

$$\begin{aligned} \text{for } n = 1 & \quad \det [a] = a \\ \text{for } n = 2 & \quad \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a_{22} - a_{12}a_{21}. \end{aligned}$$

For $n > 2$, the definition is much more complicated. It is a sum of many terms formed as follows. Choose any entry from the first row of A ; there are n possible ways to do that. Next, choose any entry from the second row which is not in the same column as the first entry chosen; there are $n - 1$ possible ways to do that. Continue in this way until you have chosen one entry from each row in such a way that no column is repeated; there are $n!$ ways to do that. Now multiply all these entries together to form a typical term. If that were all, it would be complicated enough, but there is one further twist. The products are divided into two classes of equal size according to a rather complicated rule and then the sum is formed with the terms in one class multiplied by $+1$ and those in the other class multiplied by -1 .

Here is the definition for $n = 3$ arranged to exhibit the signs.

$$\begin{aligned} \det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = & \\ & a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ & - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}. \end{aligned}$$

The definition for $n = 4$ involves $4! = 24$ terms, and I won't bother to write it out.

A better way to develop the theory is *recursively*. That is, we assume that determinants have been defined for all $(n-1) \times (n-1)$ matrices, and then use this to define determinants for $n \times n$ matrices. Since we have a definition for 1×1 matrices, this allows us in principle to find the determinant of any $n \times n$ matrix by recursively invoking the definition. This is less explicit, but it is easier to work with.

Here is the recursive definition. Let A be an $n \times n$ matrix, and let $D_j(A)$ be the determinant of the $(n-1) \times (n-1)$ matrix obtained by *deleting* the j th row and the *first column* of A . Then, define

$$\det A = a_{11}D_1(A) - a_{21}D_2(A) + \cdots + (-1)^{j+1}a_{j1}D_j(A) + \cdots + (-1)^{n+1}a_{n1}D_n(A).$$

In words: take each entry in the first column of A , multiply it by the determinant of the $(n-1) \times (n-1)$ matrix obtained by deleting the first column and that row, and then add up these entries alternating signs as you do.

Examples

$$\begin{aligned} \det \begin{bmatrix} 2 & -1 & 3 \\ 1 & 2 & 0 \\ 0 & 3 & 6 \end{bmatrix} &= 2 \det \begin{bmatrix} 2 & 0 \\ 3 & 6 \end{bmatrix} - 1 \det \begin{bmatrix} -1 & 3 \\ 3 & 6 \end{bmatrix} + 0 \det \begin{bmatrix} -1 & 3 \\ 2 & 0 \end{bmatrix} \\ &= 2(12 - 0) - 1(-6 - 9) + 0(\dots) = 24 + 15 = 39. \end{aligned}$$

Note that we didn't bother evaluating the 2×2 determinant with coefficient 0. You should check that the earlier definition gives the same result.

$$\begin{aligned} \det \begin{bmatrix} 1 & 2 & -1 & 3 \\ 0 & 1 & 2 & 0 \\ 2 & 0 & 3 & 6 \\ 1 & 1 & 2 & 1 \end{bmatrix} &= 1 \det \begin{bmatrix} 1 & 2 & 0 \\ 0 & 3 & 6 \\ 1 & 2 & 1 \end{bmatrix} - 0 \det \begin{bmatrix} 2 & -1 & 3 \\ 0 & 3 & 6 \\ 1 & 2 & 1 \end{bmatrix} \\ &\quad + 2 \det \begin{bmatrix} 2 & -1 & 3 \\ 1 & 2 & 0 \\ 1 & 2 & 1 \end{bmatrix} - 1 \det \begin{bmatrix} 2 & -1 & 3 \\ 1 & 2 & 0 \\ 0 & 3 & 6 \end{bmatrix}. \end{aligned}$$

Each of these 3×3 determinants may be evaluated recursively. (In fact we just did the last one in the previous example.) You should work them out for yourself. The answers yield

$$\det \begin{bmatrix} 1 & 2 & -1 & 3 \\ 0 & 1 & 2 & 0 \\ 2 & 0 & 3 & 6 \\ 1 & 1 & 2 & 1 \end{bmatrix} = 1(3) - 0(\dots) + 2(5) - 1(39) = -26.$$

Although this definition allows one to compute the determinant of any $n \times n$ matrix *in principle*, the number of operations grows very quickly with n . In such calculations one usually keeps track only of the multiplications since they are usually the most time consuming operations. Here are some values of $N(n)$, the number of multiplications needed for a recursive calculation of the determinant of an $n \times n$ determinant. We also tabulate $n!$ for comparison.

n	$N(n)$	$n!$
2	2	2
3	6	6
4	28	24
5	145	120
6	876	720
\vdots	\vdots	\vdots

Thus, we clearly need a more efficient method to calculate determinants. As is often the case in linear algebra, elementary row operations provide us with such a method. This is based on the following rules relating such operations to determinants.

Rule (i): If A' is obtained from A by adding a multiple of one row of A to another, then $\det A' = \det A$.

Example 208

$$\det \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \\ 1 & 2 & 1 \end{bmatrix} = 1(1 - 6) - 2(2 - 6) + 1(6 - 3) = 6$$

$$\det \begin{bmatrix} 1 & 2 & 3 \\ 0 & -3 & -3 \\ 1 & 2 & 1 \end{bmatrix} = 1(-3 + 6) - 0(2 - 6) + 1(-6 + 9) = 6.$$

Rule (ii): if A' is obtained from A by multiplying one row by a scalar c , then $\det A' = c \det A$.

Example 209

$$\det \begin{bmatrix} 1 & 2 & 0 \\ 2 & 4 & 2 \\ 0 & 1 & 1 \end{bmatrix} = 1(4 - 2) - 2(2 - 0) + 0(\dots) = -2$$

$$2 \det \begin{bmatrix} 1 & 2 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix} = 2(1(2 - 1) - 1(2 - 0) + 0(\dots)) = 2(-1) = -2.$$

One may also state this rule as follows: *any common factor of a row of A may be 'pulled out' from its determinant.* Rule (iii): If A' is obtained from A by interchanging two rows, then $\det A' = -\det A$.

Example 210

$$\det \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} = -3 \qquad \det \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = 3.$$

The verification of these rules is a bit involved, so we relegate it to an appendix. The rules allow us to compute the determinant of any $n \times n$ matrix with specific numerical entries.

Example 211 We shall calculate the determinant of a 4×4 matrix. You should make sure you keep track of which elementary row operations have been performed at each stage.

$$\begin{aligned} \det \begin{bmatrix} 1 & 2 & -1 & 1 \\ 0 & 2 & 1 & 2 \\ 3 & 0 & 1 & 1 \\ -1 & 6 & 0 & 2 \end{bmatrix} &= \det \begin{bmatrix} 1 & 2 & -1 & 1 \\ 0 & 2 & 1 & 2 \\ 0 & -6 & 4 & -2 \\ 0 & 8 & -1 & 3 \end{bmatrix} = \det \begin{bmatrix} 1 & 2 & -1 & 1 \\ 0 & 2 & 1 & 2 \\ 0 & 0 & 7 & 4 \\ 0 & 0 & -5 & -5 \end{bmatrix} \\ &= -5 \det \begin{bmatrix} 1 & 2 & -1 & 1 \\ 0 & 2 & 1 & 2 \\ 0 & 0 & 7 & 4 \\ 0 & 0 & 1 & 1 \end{bmatrix} = +5 \det \begin{bmatrix} 1 & 2 & -1 & 1 \\ 0 & 2 & 1 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 7 & 4 \end{bmatrix} \\ &= +5 \det \begin{bmatrix} 1 & 2 & -1 & 1 \\ 0 & 2 & 1 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -3 \end{bmatrix}. \end{aligned}$$

We may now use the recursive definition to calculate the last determinant. In each case there is only one non-zero entry in the first column.

$$\begin{aligned} \det \begin{bmatrix} 1 & 2 & -1 & 1 \\ 0 & 2 & 1 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -3 \end{bmatrix} &= 1 \det \begin{bmatrix} 2 & 1 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & -3 \end{bmatrix} \\ &= 1 \cdot 2 \det \begin{bmatrix} 1 & 1 \\ 0 & -3 \end{bmatrix} = 1 \cdot 2 \cdot 1 \det [-3] \\ &= 1 \cdot 2 \cdot 1 \cdot (-3) = -6. \end{aligned}$$

Hence, the determinant of the original matrix is $5(-6) = -30$.

The last calculation is a special case of a general fact which is established in much the same way by repeating the recursive definition.

$$\det \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2n} \\ 0 & 0 & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{bmatrix} = a_{11}a_{22}a_{33} \cdots a_{nn}.$$

In words, *the determinant of an upper triangular matrix is the product of its diagonal entries*. It is important to be able to tell when the determinant of an $n \times n$ matrix

A is zero. Certainly, this will be the case if the first column consists of zeroes, and indeed it turns out that the determinant vanishes if any row or any column consists only of zeroes. More generally, if either the set of rows or the set of columns is a linearly *dependent* set, then the determinant is zero. (That will be the case if the rank $r < n$ since the rank is the dimension of both the row space and the column space.) This follows from the following important theorem.

Theorem 11.18 Let A be an $n \times n$ matrix. Then A is singular if and only if $\det A = 0$. Equivalently, A is invertible, i.e., has rank n , if and only if $\det A \neq 0$.

Proof. If A is invertible, then Gaussian reduction leads to an upper triangular matrix with non-zero entries on its diagonal, and the determinant of such a matrix is the product of its diagonal entries, which is also non-zero. No elementary row operation can make the determinant zero. For, type (i) operations don't change the determinant, type (ii) operations multiply by non-zero scalars, and type (iii) operations change its sign. Hence, $\det A \neq 0$.

If A is singular, then Gaussian reduction also leads to an upper triangular matrix, but one in which at least the last row consists of zeroes. Hence, at least one diagonal entry is zero, and so is the determinant. \square \square

Example 212

$$\det \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 3 \\ 1 & 0 & 1 \end{bmatrix} = 1(1 - 0) - 2(1 - 0) + 1(3 - 2) = 0$$

so the matrix must be singular. To confirm this, we reduce

$$\begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 3 \\ 1 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 2 \\ 0 & -1 & -1 \\ 0 & -1 & -1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 2 \\ 0 & -1 & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

which shows that the matrix is singular.

In the previous section, we encountered 2×2 matrices with symbolic non-numeric entries. For such a matrix, Gaussian reduction doesn't work very well because we don't know whether the non-numeric expressions are zero or not.

Example 213 Suppose we want to know whether or not the matrix

$$\begin{bmatrix} -\lambda & 1 & 1 & 1 \\ 1 & -\lambda & 0 & 0 \\ 1 & 0 & -\lambda & 0 \\ 1 & 0 & 0 & -\lambda \end{bmatrix}$$

is singular. We could try to calculate its rank, but since we don't know what λ is, it is not clear how to proceed. Clearly, the row reduction works differently if $\lambda = 0$ than if $\lambda \neq 0$. However, we can calculate the determinant by the recursive method.

$$\begin{aligned} \det \begin{bmatrix} -\lambda & 1 & 1 & 1 \\ 1 & -\lambda & 0 & 0 \\ 1 & 0 & -\lambda & 0 \\ 1 & 0 & 0 & -\lambda \end{bmatrix} &= (-\lambda) \det \begin{bmatrix} -\lambda & 0 & 0 \\ 0 & -\lambda & 0 \\ 0 & 0 & -\lambda \end{bmatrix} - 1 \det \begin{bmatrix} 1 & 1 & 1 \\ 0 & -\lambda & 0 \\ 0 & 0 & -\lambda \end{bmatrix} \\ &\quad + 1 \det \begin{bmatrix} 1 & 1 & 1 \\ -\lambda & 0 & 0 \\ 0 & 0 & -\lambda \end{bmatrix} - 1 \det \begin{bmatrix} 1 & 1 & 1 \\ -\lambda & 0 & 0 \\ 0 & -\lambda & 0 \end{bmatrix} \\ &= (-\lambda)(-\lambda^3) - (\lambda^2) + (-\lambda^2) - (\lambda^2) \\ &= \lambda^4 - 3\lambda^2 = \lambda^2(\lambda - \sqrt{3})(\lambda + \sqrt{3}). \end{aligned}$$

Hence, this matrix is singular just in the cases $\lambda = 0$, $\lambda = \sqrt{3}$, and $\lambda = -\sqrt{3}$.

Appendix. Some Proofs We now establish the basic rules relating determinants to elementary row operations. If you are of a skeptical turn of mind, you should study this section, since the relation between the recursive definition and rules (i), (ii), and (iii) is not at all obvious. However, if you have a trusting nature, you might want to skip this section since the proofs are quite technical and not terribly enlightening.

The idea behind the proofs is to assume that the rules—actually, modified forms of the rules—have been established for $(n-1) \times (n-1)$ determinants, and then to prove them for $n \times n$ determinants. To start it all off, the rules must be checked explicitly for 2×2 determinants. I leave that step for you in the Exercises.

We start with the hardest case, rule (iii). First we consider the special case that A' is obtained from A by switching two *adjacent* rows, the i th row and the $(i+1)$ st row. Consider the recursive definition

$$\begin{aligned} \det A' &= a'_{11}D_1(A') - \cdots + (-1)^{i+1}a'_{i1}D_i(A') \\ &\quad + (-1)^{i+2}a'_{i+1,1}D_{i+1}(A') + \cdots + (-1)^{n+1}a'_{n1}D_n(A'). \end{aligned}$$

Look at the subdeterminants occurring in this sum. For $j \neq i, i+1$, we have

$$D_j(A') = -D_j(A)$$

since deleting the first column and j th row of A and then switching two rows—neither of which was deleted—changes the sign by rule (iii) for $(n-1) \times (n-1)$ determinants. The situation for $j = i$ or $j = i+1$ is different; in fact, we have

$$D_i(A') = D_{i+1}(A) \quad \text{and} \quad D_{i+1}(A') = D_i(A).$$

The first equation follows because switching rows i and $i+1$ and then deleting row i is the same as deleting row $i+1$ without touching row i . A similar argument establishes the second equation. Using this together with $a'_{i1} = a_{i+1,1}$, $a'_{i+1,1} = a_{i1}$ yields

$$\begin{aligned} (-1)^{i+1} a'_{i1} D_i(A') &= (-1)^{i+1} a_{i+1,1} D_{i+1}(A) = -(-1)^{i+2} a_{i+1,1} D_{i+1}(A) \\ (-1)^{i+2} a'_{i+1,1} D_{i+1}(A') &= (-1)^{i+2} a_{i1} D_i(A) = -(-1)^{i+1} a_{i1} D_i(A). \end{aligned}$$

In other words, all terms in the recursive definition of $\det A'$ are negatives of the corresponding terms of $\det A$ *except* those in positions i and $i+1$ which get reversed with signs changed. Hence, the effect of switching adjacent rows is to change the sign of the sum.

Suppose instead that non-adjacent rows in positions i and j are switched, and suppose for the sake of argument that $i < j$. One way to do this is as follows. First move row i past each of the rows between row i and row j . This involves some number of switches of adjacent rows—call that number k . ($k = j - i - 1$, but it that doesn't matter in the proof.) Next, move row j past row i and then past the k rows just mentioned, all in their new positions. That requires $k+1$ switches of adjacent rows. All told, to switch rows i and j in this way requires $2k+1$ switches of adjacent rows. The net effect is to multiply the determinant by $(-1)^{2k+1} = -1$ as required.

There is one important consequence of rule (iii) which we shall use later in the proof of rule (i). Rule (iii): *If an $n \times n$ matrix has two equal rows, then $\det A = 0$.*

This is not too hard to see. Interchanging two rows changes the sign of $\det A$, but if the rows are equal, it doesn't change anything. However, the only number with the property that it isn't changed by changing its sign is the number 0. Hence, $\det A = 0$. We next verify rule (ii). Suppose A' is obtained from A by multiplying the i th row by c . Consider the recursive definition

$$\det A' = a'_{11} D_1(A') + \cdots + (-1)^{i+1} a'_{i1} D_i(A') + \cdots + (-1)^{n+1} a'_{n1} D_n(A'). \quad (183)$$

For any $j \neq i$, $D_j(A') = c D_j(A)$ since one of the rows appearing in that determinant is multiplied by c . Also, $a'_{j1} = a_{j1}$ for $j \neq i$. On the other hand, $D_i(A') = D_i(A)$ since the i th row is deleted in calculating these quantities, and, except for the i th row, A' and A agree. In addition, $a'_{i1} = c a_{i1}$ so we pick up the extra factor of c in any case. It follows that every term on the right of (183) has a factor c , so $\det A' = c \det A$. Finally, we attack the proof of rule (i). It turns out to be necessary to verify the following stronger rule.

Rule (ia): Suppose A, A' , and A'' are three $n \times n$ matrices which agree except in the i th row. Suppose moreover that the i th row of A is the sum of the i th row of A' and the i th row of A'' . Then $\det A = \det A' + \det A''$. Let's first see why rule (ia)

implies rule (i). We can add c times the j th row of A to its i row as follows. Let $B' = A$, let B'' be the matrix obtained from A by replacing its i th row by c times its j th row, and let B be the matrix obtained from A by adding c times its j th row to its i th row. Then according to rule (ia), we have

$$\det B = \det B' + \det B'' = \det A + \det B''.$$

On the other hand, by rule (ii), $\det B'' = c \det A''$ where A'' has both i th and j th rows equal to the j th row of A . Hence, by rule (iiie), $\det A'' = 0$, and $\det B = \det A$. Finally, we establish rule (1a). Assume it is known to be true for $(n-1) \times (n-1)$ determinants. We have

$$\det A = a_{i1}D_1(A) - \cdots + (-1)^{i+1}a_{i1}D_i(A) + \cdots + (-1)^{n+1}a_{n1}D_n(A). \quad (184)$$

For $j \neq i$, the sum rule (ia) may be applied to the determinants $D_i(A)$ because the appropriate submatrix has one row which breaks up as a sum as needed. Hence,

$$D_j(A) = D_j(A') + D_j(A'').$$

Also, for $j \neq i$, we have $a_{j1} = a'_{j1} = a''_{j1}$ since all the matrices agree in any row except the i th row. Hence, for $j \neq i$,

$$a_{i1}D_i(A) = a_{i1}D_i(A') + a_{i1}D_i(A'') = a'_{i1}D_i(A') + a''_{i1}D_i(A'').$$

On the other hand, $D_i(A) = D_i(A') = D_i(A'')$ because in each case the i th row was deleted. But $a_{i1} = a'_{i1} + a''_{i1}$, so

$$a_{i1}D_i(A) = a'_{i1}D_i(A) + a''_{i1}D_i(A) = a'_{i1}D_i(A') + a''_{i1}D_i(A'').$$

It follows that every term in (184) breaks up into a sum as required, and $\det A = \det A' + \det A''$.

Exercises for 11.3.

1. Find the determinants of each of the following matrices. Use whatever method seems most convenient, but seriously consider the use of elementary row operations.

(a) $\begin{bmatrix} 1 & 1 & 2 \\ 1 & 3 & 5 \\ 6 & 4 & 1 \end{bmatrix}.$

(b) $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \\ 1 & 4 & 2 & 3 \\ 4 & 3 & 2 & 1 \end{bmatrix}.$

$$(c) \begin{bmatrix} 0 & 0 & 0 & 0 & 3 \\ 1 & 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 2 \end{bmatrix}.$$

$$(d) \begin{bmatrix} 0 & x & y \\ -x & 0 & z \\ -y & -z & 0 \end{bmatrix}.$$

2. Verify the following rules for 2×2 determinants.

(i) If A' is obtained from A by adding a multiple of the first row to the second, then $\det A' = \det A$.

(ii) If A' is obtained from A by multiplying its first row by c , then $\det A' = c \det A$.

(iii) If A' is obtained from A by interchanging its two rows, then $\det A' = -\det A$.

Rules (i) and (ii) for the first row, together with rule (iii) allow us to derive rules (i) and (ii) for the second row. Explain.

3. Derive the following generalization of rule (i) for 2×2 determinants.

$$\det \begin{bmatrix} a' + a'' & b' + b'' \\ c & d \end{bmatrix} = \det \begin{bmatrix} a' & b' \\ c & d \end{bmatrix} + \det \begin{bmatrix} a'' & b'' \\ c & d \end{bmatrix}.$$

What is the corresponding rule for the second row? Why do you get it for free if you use the results of the previous problem?

11.4 Some Important Properties of Determinants

Theorem 11.19 (The Product Rule) Let A and B be $n \times n$ matrices. Then

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

Proof. First assume that A is non-singular. Then there is a sequence of row operations which reduces A to the identity

$$A \rightarrow A_1 \rightarrow A_2 \rightarrow \dots \rightarrow A_k = I.$$

Associated with each of these operations will be a multiplier c_i which will depend on the particular operation, and

$$\det A = c_1 \det A_1 = c_1 c_2 \det A_2 = \dots = c_1 c_2 \dots c_k \det A_k = c_1 c_2 \dots c_k$$

since $A_k = I$ and $\det I = 1$. Now apply exactly these row operations to the product AB

$$AB \rightarrow A_1B \rightarrow A_2B \rightarrow \dots \rightarrow A_kB = IB = B.$$

The same multipliers contribute factors at each stage, and

$$\det AB = c_1 \det A_1B = c_1 c_2 \det A_2B = \dots = \underbrace{c_1 c_2 \dots c_k}_{\det A} \det B = \det A \det B.$$

Assume instead that A is singular. Then, AB is also singular. (This follows from the fact that the rank of AB is at most the rank of A , as mentioned in the Exercises for Chapter X, Section 6. However, here is a direct proof for the record. Choose a sequence of elementary row operations for A , the end result of which is a matrix A' with at least one row of zeroes. Applying the same operations to AB yields $A'B$ which also has to have at least one row of zeroes.) It follows that both $\det AB$ and $\det A \det B$ are zero, so they are equal. \square \square

Transposes Let A be an $m \times n$ matrix. The *transpose* of A is the $n \times m$ matrix for which the columns are the rows of A . (Also, its rows are the columns of A .) It is usually denoted A^t , but other notations are possible.

Examples

$$\begin{aligned} A &= \begin{bmatrix} 2 & 0 & 1 \\ 2 & 1 & 2 \end{bmatrix} & A^t &= \begin{bmatrix} 2 & 2 \\ 0 & 1 \\ 1 & 2 \end{bmatrix} \\ A &= \begin{bmatrix} 1 & 2 & 3 \\ 0 & 2 & 3 \\ 0 & 0 & 3 \end{bmatrix} & A^t &= \begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 3 & 3 & 3 \end{bmatrix} \\ \mathbf{a} &= \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} & \mathbf{a}^t &= [a_1 \quad a_2 \quad a_3]. \end{aligned}$$

The following rule follows almost immediately from the definition. Assume A is an $m \times n$ matrix and B is an $n \times p$ matrix. Then

$$(AB)^t = B^t A^t.$$

Note that *the order on the right is reversed*. Unless the matrices are square, the shapes won't even match if the order is not reversed.

Theorem 11.20 Let A be an $n \times n$ matrix. Then

$$\det A^t = \det A.$$

Example 214

$$\det \begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} = 1(1 - 0) - 2(0 - 0) + 0(\dots) = 1$$

$$\det \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} = 1(1 - 0) - 0(\dots) + 1(0 - 0) = 1.$$

The importance of this theorem is that it allows us to go freely from statements about determinants involving rows of the matrix to corresponding statements involving columns and vice-versa.

Proof. If A is singular, then A^t is also singular and vice-versa. For, the rank may be characterized as either the dimension of the row space or the dimension of the column space, and an $n \times n$ matrix is singular if its rank is less than n . Hence, in the singular case, $\det A = 0 = \det A^t$.

Suppose then that A is non-singular. Then there is a sequence of elementary row operations

$$A \rightarrow A_1 \rightarrow A_2 \rightarrow \dots \rightarrow A_k = I.$$

Recall from Chapter X, Section 4

that each elementary row operation may be accomplished by multiplying by an appropriate elementary matrix. Let C_i denote the elementary matrix needed to perform the i th row operation. Then,

$$A \rightarrow A_1 = C_1 A \rightarrow A_2 = C_2 C_1 A \rightarrow \dots \rightarrow A_k = C_k C_{k-1} \dots C_2 C_1 A = I.$$

In other words,

$$A = (C_k \dots C_2 C_1)^{-1} = C_1^{-1} C_2^{-1} \dots C_k^{-1}.$$

To simplify the notation, let $D_i = C_i^{-1}$. The inverse D of an elementary matrix C is also an elementary matrix; its effect is the row operation which reverses the effect of C . Hence, we have shown that *any non-singular square matrix A may be expressed as a product of elementary matrices*

$$A = D_1 D_2 \dots D_k.$$

Hence, by the product rule

$$\det A = (\det D_1)(\det D_2) \dots (\det D_k).$$

On the other hand, we have by rule for the transpose of a product

$$A^t = D_k^t \dots D_2^t D_1^t,$$

so by the product rule

$$\det A^t = \det(D_k^t) \dots \det(D_2^t) \det(D_1^t).$$

Suppose we know the rule $\det D^t = \det D$ for any elementary matrix D . Then,

$$\begin{aligned} \det A^t &= \det(D_k^t) \dots \det(D_2^t) \det(D_1^t) \\ &= \det(D_k) \dots \det(D_2) \det(D_1) \\ &= (\det D_1)(\det D_2) \dots (\det D_k) = \det A. \end{aligned}$$

(We used the fact that the products on the right are products of scalars and so can be rearranged any way we like.)

It remains to establish the rule for elementary matrices. If $D = E_{ij}(c)$ is obtained from the identity matrix by adding c times its j th row to its i th row, then $D^t = E_{ji}(c)$ is a matrix of exactly the same type. In each case, $\det D = \det D^t = 1$. If $D = E_i(c)$ is obtained by multiplying the i th row of the identity matrix by c , then D^t is exactly the same matrix $E_i(c)$. Finally, if $D = E_{ij}$ is obtained from the identity matrix by interchanging its i th and j th rows, then D^t is E_{ji} which in fact is just E_{ij} again. Hence, in each case $\det D^t = \det D$ does hold. \square \square

Because of this rule, we may use *column operations* as well as row operations to calculate determinants. For, performing a column operation is the same as transposing the matrix, performing the corresponding row operation, and then transposing back. The two transpositions don't affect the determinant.

Example

$$\begin{aligned} \det \begin{bmatrix} 1 & 2 & 3 & 0 \\ 2 & 1 & 3 & 1 \\ 3 & 3 & 6 & 2 \\ 4 & 2 & 6 & 4 \end{bmatrix} &= \det \begin{bmatrix} 1 & 2 & 2 & 0 \\ 2 & 1 & 1 & 1 \\ 3 & 3 & 3 & 2 \\ 4 & 2 & 2 & 4 \end{bmatrix} && \text{operation } (-1)c1 + c3 \\ &= 0. \end{aligned}$$

The last step follows because the 2nd and 3rd columns are equal, which implies that the rank (dimension of the column space) is less than 4. (You could also subtract the third column from the second and get a column of zeroes, etc.)

Expansion in Minors or Cofactors There is a generalization of the formula used for the recursive definition. Namely, for any $n \times n$ matrix A , let $D_{ij}(A)$ be the determinant of the $(n-1) \times (n-1)$ matrix obtained by deleting the i th row and j th column of A . Then,

$$\begin{aligned} \det A &= \sum_{i=1}^n (-1)^{i+j} a_{ij} D_{ij}(A) \\ &= (-1)^{1+j} a_{1j} D_{1j}(A) + \dots + (-1)^{i+j} a_{ij} D_{ij}(A) + \dots + (-1)^{n+j} a_{nj} D_{nj}(A). \end{aligned} \tag{185}$$

The special case $j = 1$ is the recursive definition given in the previous section. The more general rule is easy to derive from the special case $j = 1$ by means of column interchanges. Namely, form a new matrix A' by moving the j th column to the first position by successively interchanging it with columns $j - 1, j - 2, \dots, 2, 1$. There are $j - 1$ interchanges, so the determinant is changed by the factor $(-1)^{j-1}$. Now apply the rule for the first column. The first column of A' is the j th column of A , and deleting it has the same effect as deleting the j th column of A . Hence, $a'_{i1} = a_{ij}$ and $D_i(A') = D_{ij}(A)$. Thus,

$$\begin{aligned}\det A &= (-1)^{j-1} \det A' = (-1)^{j-1} \sum_{i=1}^n (-1)^{1+i} a'_{i1} D_i(A') \\ &= \sum_{i=1}^n (-1)^{i+j} a_{ij} D_{ij}(A).\end{aligned}$$

Similarly, there is a corresponding rule for any *row* of a matrix

$$\begin{aligned}\det A &= \sum_{j=1}^n (-1)^{i+j} a_{ij} D_{ij}(A) \\ &= (-1)^{i+1} a_{i1} D_{i1} + \dots + (-1)^{i+j} a_{ij} D_{ij}(A) + \dots + (-1)^{i+n} a_{in} D_{in}(A).\end{aligned}\tag{186}$$

This formula is obtained from (185) by transposing, applying the corresponding column rule, and then transposing back.

Example Expand the following determinant using its second row.

$$\begin{aligned}\det \begin{bmatrix} 1 & 2 & 3 \\ 0 & 6 & 0 \\ 3 & 2 & 1 \end{bmatrix} &= (-1)^{2+3} 0(\dots) + (-1)^{2+2} 6 \det \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix} + (-1)^{2+3} 0(\dots) \\ &= 6(1 - 9) = -48.\end{aligned}$$

There is some terminology which you may see used in connection with these formulas. The determinant $D_{ij}(A)$ of the $(n-1) \times (n-1)$ matrix obtained by deleting the i th row and j th column is called the i, j -*minor* of A . The quantity $(-1)^{i+j} D_{ij}(A)$ is called the i, j -*cofactor*. Formula (185) is called expansion in minors (or cofactors) of the j th column and formula (186) is called expansion in minors (or cofactors) of the i th row. It is not necessary to remember the terminology as long as you remember the formulas and understand how they are used. **Cramer's Rule** One may use determinants to derive a formula for the solutions of a *non-singular* system of n equations in n unknowns

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

The formula is called *Cramer's rule*, and here it is. For the j th unknown x_j , take the determinant of the matrix formed by replacing the j th column of the coefficient matrix A by \mathbf{b} , and divide it by $\det A$. In symbols,

$$x_j = \frac{\det \begin{bmatrix} a_{11} & \dots & b_1 & \dots & a_{1n} \\ a_{21} & \dots & b_2 & \dots & a_{2n} \\ \vdots & \dots & \vdots & \dots & \vdots \\ a_{n1} & \dots & b_n & \dots & a_{nn} \end{bmatrix}}{\det \begin{bmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ a_{21} & \dots & a_{2j} & \dots & a_{2n} \\ \vdots & \dots & \vdots & \dots & \vdots \\ a_{n1} & \dots & a_{nj} & \dots & a_{nn} \end{bmatrix}}$$

Example Consider

$$\begin{bmatrix} 1 & 0 & 2 \\ 1 & 1 & 2 \\ 2 & 0 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \\ 3 \end{bmatrix}.$$

We have

$$\det \begin{bmatrix} 1 & 0 & 2 \\ 1 & 1 & 2 \\ 2 & 0 & 6 \end{bmatrix} = 2.$$

(Do you see a quick way to compute that?) Hence,

$$\begin{aligned} x_1 &= \frac{\det \begin{bmatrix} 1 & 0 & 2 \\ 5 & 1 & 2 \\ 3 & 0 & 6 \end{bmatrix}}{2} = \frac{0}{2} = 0 \\ x_2 &= \frac{\det \begin{bmatrix} 1 & 1 & 2 \\ 1 & 5 & 2 \\ 2 & 3 & 6 \end{bmatrix}}{2} = \frac{8}{2} = 4 \\ x_3 &= \frac{\det \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 5 \\ 2 & 0 & 3 \end{bmatrix}}{2} = \frac{1}{2}. \end{aligned}$$

You should try to do this by Gauss-Jordan reduction.

Cramer's rule is not too useful for solving specific numerical systems of equations. The only practical method for calculating the needed determinants for n large is to use row (and possibly column) operations. It is usually easier to use row operations to solve the system without resorting to determinants. However, if the system has non-numeric symbolic coefficients, Cramer's rule is sometimes useful. Also, it is often valuable as a theoretical tool.

Cramer's rule is related to expansion in minors. You can find further discussion of it and proofs in Section 5.4 and 5.5 of *Introduction to Linear Algebra* by Johnson, Riess, and Arnold. (See also Section 4.5 of *Applied Linear Algebra* by Noble and Daniel.)

Exercises for 11.4.

1. Check the validity of the product rule for the product

$$\begin{bmatrix} 1 & -2 & 6 \\ 2 & 0 & 3 \\ -3 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 2 \\ 1 & 1 & 0 \end{bmatrix}.$$

2. Find

$$\det \begin{bmatrix} 3 & 0 & 0 & 0 \\ 2 & 2 & 0 & 0 \\ 1 & 6 & 4 & 0 \\ 1 & 5 & 4 & 3 \end{bmatrix}.$$

Of course, the answer is the product of the diagonal entries. Using the properties discussed in the section, see how many different ways you can come to this conclusion.

What can you conclude in general about the determinant of a lower triangular square matrix?

3. (a) Prove that if A is an invertible $n \times n$ matrix, then $\det(A^{-1}) = \frac{1}{\det A}$.
 (b) Using part(a), show that if A is any $n \times n$ matrix and P is an invertible $n \times n$ matrix, then $\det(PAP^{-1}) = \det A$.
4. Why does Cramer's rule fail if the coefficient matrix A is singular?
5. Use Cramer's rule to solve the system

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_2 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}.$$

Also, solve it by Gauss-Jordan reduction and compare the amount of work you had to do in each case.

11.5 Eigenvalues and Eigenvectors

As in Section 2, we want to solve an $n \times n$ system

$$\frac{d\mathbf{x}}{dt} = A\mathbf{x} \quad (187)$$

by finding a basis $\{\mathbf{x}_1(t), \mathbf{x}_2(t), \dots, \mathbf{x}_n(t)\}$ of the solution space. In case A is *constant*, it was suggested that we *look for* solutions of the form

$$\mathbf{x} = e^{\lambda t} \mathbf{v}$$

where λ and $\mathbf{v} \neq 0$ are to be determined by the process. Such solutions form a linearly independent set as long as the corresponding \mathbf{v} 's form a linearly independent set. For, suppose

$$\mathbf{x}_1 = e^{\lambda_1 t} \mathbf{v}_1, \mathbf{x}_2 = e^{\lambda_2 t} \mathbf{v}_2, \dots, \mathbf{x}_k = e^{\lambda_k t} \mathbf{v}_k$$

are k such solutions. We know that the set of solutions $\{\mathbf{x}_1(t), \dots, \mathbf{x}_k(t)\}$ is linearly independent if and only if the set of vectors obtained by evaluating the functions at $t = 0$ is linearly independent. However, this set of vectors is just $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$.

We discovered in Section 2 that trying a solution of the form $\mathbf{x} = e^{\lambda t} \mathbf{v}$ leads to the eigenvalue–eigenvector problem

$$A\mathbf{v} = \lambda\mathbf{v}. \quad (188)$$

We redo some of the algebra in Section 2 as follows. Rewrite equation (188) as

$$\begin{aligned} A\mathbf{v} &= \lambda\mathbf{v} \\ A\mathbf{v} - \lambda\mathbf{v} &= 0 \\ A\mathbf{v} - \lambda I\mathbf{v} &= 0 \\ (A - \lambda I)\mathbf{v} &= 0. \end{aligned}$$

The last equation is the homogeneous $n \times n$ system with $n \times n$ coefficient matrix

$$A - \lambda I = \begin{bmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{bmatrix}.$$

It has a *non-zero* solution vector \mathbf{v} if and only if the coefficient matrix has rank less than n , i.e., if and only if it is *singular*. By Theorem 11.3,

this will be true if and only if λ satisfies the *characteristic equation*

$$\det(A - \lambda I) = \det \begin{bmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{bmatrix} = 0. \quad (189)$$

As in Section 2, the strategy for finding eigenvalues and eigenvectors is as follows. First find the roots of the characteristic equation. These are the eigenvalues. Then for each root λ , find a general solution for the system

$$(A - \lambda I)\mathbf{v} = 0. \quad (190)$$

This gives us all the eigenvectors for that eigenvalue.

Example 216 Consider the matrix

$$A = \begin{bmatrix} 1 & 4 & 3 \\ 4 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}.$$

The characteristic equation is

$$\begin{aligned} \det(A - \lambda I) &= \det \begin{bmatrix} 1 - \lambda & 4 & 3 \\ 4 & 1 - \lambda & 0 \\ 3 & 0 & 1 - \lambda \end{bmatrix} \\ &= (1 - \lambda)((1 - \lambda)^2 - 0) - 4(4(1 - \lambda) - 0) + 3(0 - 3(1 - \lambda)) \\ &= (1 - \lambda)^3 - 25(1 - \lambda) = (1 - \lambda)((1 - \lambda)^2 - 25) \\ &= (1 - \lambda)(\lambda^2 - 2\lambda - 24) = (1 - \lambda)(\lambda - 6)(\lambda + 4) = 0. \end{aligned}$$

Hence, the eigenvalues are $\lambda = 1$, $\lambda = 6$, and $\lambda = -4$. We proceed to find the eigenvectors for each of these eigenvalues, starting with the largest.

First, take $\lambda = 6$, and put it in (190) to obtain the system

$$\begin{bmatrix} 1 - 6 & 4 & 3 \\ 4 & 1 - 6 & 0 \\ 3 & 0 & 1 - 6 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 0 \quad \text{or} \quad \begin{bmatrix} -5 & 4 & 3 \\ 4 & -5 & 0 \\ 3 & 0 & -5 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 0.$$

To solve, use Gauss-Jordan reduction

$$\begin{aligned} \begin{bmatrix} -5 & 4 & 3 \\ 4 & -5 & 0 \\ 3 & 0 & -5 \end{bmatrix} &\rightarrow \begin{bmatrix} -1 & -1 & 3 \\ 4 & -5 & 0 \\ 3 & 0 & -5 \end{bmatrix} \rightarrow \begin{bmatrix} -1 & -1 & 3 \\ 0 & -9 & 12 \\ 0 & -3 & 4 \end{bmatrix} \\ &\rightarrow \begin{bmatrix} -1 & -1 & 3 \\ 0 & 0 & 0 \\ 0 & -3 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} -1 & -1 & 3 \\ 0 & 3 & -4 \\ 0 & 0 & 0 \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 1 & 1 & -3 \\ 0 & 1 & -4/3 \\ 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & -5/3 \\ 0 & 1 & -4/3 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Note that the matrix is singular, and the rank is smaller than 3. This must be the case because the condition $\det(A - \lambda I) = 0$ guarantees it. If the coefficient matrix

were non-singular, you would know that there was a mistake: either the roots of the characteristic equation are wrong or the row reduction was not done correctly.

The general solution is

$$\begin{aligned}v_1 &= (5/3)v_3 \\v_2 &= (4/3)v_3\end{aligned}$$

with v_3 free. The general solution vector is

$$\mathbf{v} = \begin{bmatrix} (5/3)v_3 \\ (4/3)v_3 \\ v_3 \end{bmatrix} = v_3 \begin{bmatrix} 5/3 \\ 4/3 \\ 1 \end{bmatrix}.$$

Hence, the solution space is 1-dimensional. A basis may be obtained by setting $v_3 = 1$ as usual, but it is a bit neater to put $v_3 = 3$ so as to avoid fractions. Thus,

$$\mathbf{v}_1 = \begin{bmatrix} 5 \\ 4 \\ 3 \end{bmatrix}$$

constitutes a basis for the solution space. Note that we have now found all eigenvectors for the eigenvalue $\lambda = 6$. They are all the non-zero vectors in the 1-dimensional solution subspace, i.e., all non-zero multiples of \mathbf{v}_1 .

Next take $\lambda = 1$ and put it in (190) to obtain the system

$$\begin{bmatrix} 0 & 4 & 3 \\ 4 & 0 & 0 \\ 3 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 0.$$

Use Gauss-Jordan reduction

$$\begin{bmatrix} 0 & 4 & 3 \\ 4 & 0 & 0 \\ 3 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 3/4 \\ 0 & 0 & 0 \end{bmatrix}.$$

The general solution is

$$\begin{aligned}v_1 &= 0 \\v_2 &= -(3/4)v_3\end{aligned}$$

with v_3 free. Thus the general solution vector is

$$\mathbf{v} = \begin{bmatrix} 0 \\ -(3/4)v_3 \\ v_3 \end{bmatrix} = v_3 \begin{bmatrix} 0 \\ -3/4 \\ 1 \end{bmatrix}.$$

Put $v_3 = 4$ to obtain a single basis vector

$$\mathbf{v}_2 = \begin{bmatrix} 0 \\ -3 \\ 4 \end{bmatrix}.$$

The set of eigenvectors for the eigenvalue $\lambda = 1$ is the set of non-zero multiples of \mathbf{v}_2 .

Finally, take $\lambda = -4$, and put this in (190) to obtain the system

$$\begin{bmatrix} 5 & 4 & 3 \\ 4 & 5 & 0 \\ 3 & 0 & 5 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 0.$$

Solve this by Gauss-Jordan reduction.

$$\begin{aligned} \begin{bmatrix} 5 & 4 & 3 \\ 4 & 5 & 0 \\ 3 & 0 & 5 \end{bmatrix} &\rightarrow \begin{bmatrix} 1 & -1 & 3 \\ 4 & 5 & 0 \\ 3 & 0 & 5 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -1 & 3 \\ 0 & 9 & -12 \\ 0 & 3 & -4 \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 1 & -1 & 3 \\ 0 & 3 & -4 \\ 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 5/3 \\ 0 & 1 & -4/3 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

The general solution is

$$\begin{aligned} v_1 &= -(5/3)v_3 \\ v_2 &= (4/3)v_3 \end{aligned}$$

with v_3 free. The general solution vector is

$$\mathbf{v} = \begin{bmatrix} -(5/3)v_3 \\ (4/3)v_3 \\ v_3 \end{bmatrix} = v_3 \begin{bmatrix} -5/3 \\ 4/3 \\ 1 \end{bmatrix}.$$

Setting $v_3 = 3$ yields the basis vector

$$\mathbf{v}_3 = \begin{bmatrix} -5 \\ 4 \\ 3 \end{bmatrix}.$$

The set of eigenvectors for the eigenvalue $\lambda = -4$ consists of all non-zero multiples of \mathbf{v}_3 .

The set $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ obtained in the previous example is linearly independent. To see this apply Gaussian reduction to the matrix with these vectors as columns:

$$\begin{bmatrix} 5 & 0 & -5 \\ 4 & -3 & 4 \\ 3 & 4 & 3 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & -3 & 8 \\ 0 & 4 & 6 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -8/3 \\ 0 & 0 & 50/3 \end{bmatrix}.$$

The reduced matrix has rank 3, so the columns of the original matrix form an independent set.

It is no accident that a set so obtained is linearly independent. The following theorem tells us that this will always be the case.

Theorem 11.21 Let A be an $n \times n$ matrix. Let $\lambda_1, \lambda_2, \dots, \lambda_k$ be different eigenvalues of A , and let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ be corresponding eigenvectors. Then

$$\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$$

is a linearly independent set.

Proof. Assume $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ is not a linearly independent set, and try to derive a contradiction. In this case, one of the vectors in the set can be expressed as a linear combination of the others. If we number the elements appropriately, we may assume that

$$\mathbf{v}_1 = c_2 \mathbf{v}_2 + \dots + c_k \mathbf{v}_k, \quad (191)$$

where $r \leq k$. (Before renumbering, leave out any vector \mathbf{v}_i on the right if it appears with coefficient $c_i = 0$.) Note that we may also assume that no vector which appears on the right is a linear combination of the others because otherwise we could express it so and after combining terms delete it from the sum. Thus we may assume the vectors which appear on the right form a linearly independent set. Multiply (191) on the left by A . We get

$$\begin{aligned} A\mathbf{v}_1 &= c_2 A\mathbf{v}_2 + \dots + c_k A\mathbf{v}_k \\ \lambda_1 \mathbf{v}_1 &= c_2 \lambda_2 \mathbf{v}_2 + \dots + c_k \lambda_k \mathbf{v}_k \end{aligned} \quad (192)$$

where in (192) we used the fact that each \mathbf{v}_i is an eigenvector with eigenvalue λ_i . Now multiply (191) by λ_1 and subtract from (192). We get

$$0 = c_2(\lambda_2 - \lambda_1)\mathbf{v}_2 + \dots + c_k(\lambda_k - \lambda_1)\mathbf{v}_k. \quad (193)$$

Not all the coefficients on the right in this equation are zero. For at least one of the $c_i \neq 0$ (since $\mathbf{v}_1 \neq 0$), and none of the quantities $\lambda_2 - \lambda_1, \dots, \lambda_k - \lambda_1$ is zero. It follows that (193) may be used to express one of the vectors $\mathbf{v}_2, \dots, \mathbf{v}_k$ as a linear combination of the others. However, this contradicts the assertion that the set of vectors appearing on the right is linearly independent. Hence, our initial assumption that the set $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ is dependent must be false, and the theorem is proved.

You should try this argument out on a set $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ of three eigenvectors to see if you understand it. □ □

Historical Aside The concepts discussed here and in Section 2 were invented by the 19th century English mathematicians Cayley and Sylvester, but they used the terms ‘characteristic vector’ and ‘characteristic value’. These were translated into German as ‘Eigenvector’ and ‘Eigenwerte’, and then partially translated back into English—largely by physicists—as ‘eigenvector’ and ‘eigenvalue’. Some English and American mathematicians tried to retain the original English terms, but they were overwhelmed by extensive use of the physicists’ language in applications. Nowadays everyone uses the German terms. The one exception is that we still call

$$\det(A - \lambda I) = 0$$

the characteristic equation and not some strange German-English name. **Application to Homogeneous Linear Systems of Differential Equations** What lessons can we learn for solving systems of differential equations from the previous discussion of eigenvalues? First, Theorem 11.21 assures us that if the $n \times n$ matrix A has distinct eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_k$ corresponding to eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$, then the functions

$$\mathbf{x}_1 = e^{\lambda_1 t} \mathbf{v}_1, \mathbf{x}_2 = e^{\lambda_2 t} \mathbf{v}_2, \dots, \mathbf{x}_k = e^{\lambda_k t} \mathbf{v}_k$$

form a linearly independent set of solutions of $\frac{d\mathbf{x}}{dt} = A\mathbf{x}$. If $k = n$ then this set will be a basis for the space of solutions of $\frac{d\mathbf{x}}{dt} = A\mathbf{x}$. (Why?)

Example 216a Consider the system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 1 & 4 & 3 \\ 4 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix} \mathbf{x}.$$

We found that $\lambda_1 = 6, \lambda_2 = 1, \lambda_3 = -4$ are eigenvalues of the coefficient matrix corresponding to eigenvectors

$$\mathbf{v}_1 = \begin{bmatrix} 5 \\ 4 \\ 3 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 0 \\ -3 \\ 4 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} -5 \\ 4 \\ 3 \end{bmatrix}.$$

Since $k = 3$ in this case, we conclude that the general solution of the system of differential equations is

$$\mathbf{x} = c_1 e^{6t} \begin{bmatrix} 5 \\ 4 \\ 3 \end{bmatrix} + c_2 e^t \begin{bmatrix} 0 \\ -3 \\ 4 \end{bmatrix} + c_3 e^{-4t} \begin{bmatrix} -5 \\ 4 \\ 3 \end{bmatrix}.$$

The above example illustrates that we should ordinarily look for a linearly independent set of eigenvectors, as large as possible, for the $n \times n$ coefficient matrix A . If we can find such a set with n elements, then we may write out a complete solution as in the example. The condition that there is a linearly independent set of n eigenvectors for A , i.e., that *there is a basis for \mathbf{R}^n (\mathbf{C}^n in the complex case) consisting of eigenvectors for A* , will certainly be verified if there are n distinct eigenvalues (Theorem 11.21). We shall see later that there are other circumstances in which it holds. On the other hand, it is easy to find examples where it fails.

Example 217 Consider the 2×2 system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 2 & 3 \\ 0 & 2 \end{bmatrix} \mathbf{x}.$$

The eigenvalues are found by solving the characteristic equation of the coefficient matrix

$$\det \begin{bmatrix} 2-\lambda & 3 \\ 0 & 2-\lambda \end{bmatrix} = (2-\lambda)^2 = 0.$$

Hence there is only one (double) root $\lambda = 2$. To find the corresponding eigenvectors, solve

$$\begin{bmatrix} 0 & 3 \\ 0 & 0 \end{bmatrix} \mathbf{v} = 0.$$

This one is easy, (but you can make it hard for yourself if you get confused about Gauss-Jordan reduction). The general solution is

$$v_2 = 0, \quad v_1 \text{ free.}$$

Hence, the general solution vector is

$$\mathbf{v} = \begin{bmatrix} v_1 \\ 0 \end{bmatrix} = v_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Hence, a basic eigenvector for $\lambda = 2$ is

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \mathbf{e}_1.$$

$\{\mathbf{v}_1\}$ is certainly not a basis for \mathbf{R}^2 .

In the sections that follow, we shall be concerned with these two related questions. Given an $n \times n$ matrix A , when can we be sure that there is a basis for \mathbf{R}^n (\mathbf{C}^n in the complex case) consisting of eigenvectors for A ? If there is no such basis, is there another way to solve the system $d\mathbf{x}/dt = A\mathbf{x}$?

Solving Polynomial Equations To find the eigenvalues of an $n \times n$ matrix, you have to solve a polynomial equation. You all know how to solve quadratic equations, but you may be stumped by cubic or higher equations, particularly if there are no obvious ways to factor. You should review what you learned in high school about this subject, but here are a few guidelines to help you.

First, it is not generally possible to find a simple solution in closed form for an algebraic equation. For most equations you might encounter in practice, you would have to use some method to approximate a solution. (Many such methods exist. One you may have learned in your calculus course is *Newton's Method*.) Unfortunately, an approximate solution of the characteristic equation isn't much good for finding the corresponding eigenvectors. After all, the system

$$(A - \lambda I)\mathbf{v} = 0$$

must have rank smaller than n for there to be non-zero solutions. If you replace the exact value of λ by an approximation, the chances are that the new system will have rank n . Hence, the textbook method we have described for finding eigenvectors

won't work. There are in fact many alternative methods for finding eigenvalues and eigenvectors approximately when exact solutions are not available. Whole books are devoted to such methods. (See *Johnson, Riess, and Arnold* or *Noble and Daniel* for some discussion of these matters.)

Fortunately, textbook exercises and examination questions almost always involve characteristic equations for which exact solutions exist, but it is not always obvious what they are. Here is one fact (a consequence of an important result called *Gauss's Lemma*) which helps us find such exact solutions when they exist. Consider an equation of the form

$$\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n = 0$$

where all the coefficients are *integers*. (The characteristic equation of a matrix always has leading coefficient 1 or -1 . In the latter case, just imagine you have multiplied through by -1 to apply the method.) Gauss's Lemma tells us that if this equation has any roots which are *rational numbers*, i.e., quotients of integers, then any such root is actually an integer, and, moreover, it must divide the constant term a_n . Hence, the first step in solving such an equation should be checking all possible factors (positive and negative) of the constant term. Once, you know a root r_1 , you can divide through by $\lambda - r_1$ to reduce to a lower degree equation. If you know the method of synthetic division, you will find checking the possible roots and the polynomial long division much simpler.

Example 218 Solve

$$\lambda^3 - 3\lambda + 2 = 0.$$

If there are any rational roots, they must be factors of the constant term 2. Hence, we must try 1, -1 , 2, -2 . Substituting $\lambda = 1$ in the equation yields 0, so it is a root. Dividing $\lambda^3 - 3\lambda + 2$ by $\lambda - 1$ yields

$$\lambda^3 - 3\lambda + 2 = (\lambda - 1)(\lambda^2 + \lambda - 2)$$

and this may be factored further to obtain

$$\lambda^3 - 3\lambda + 2 = (\lambda - 1)(\lambda - 1)(\lambda + 2) = (\lambda - 1)^2(\lambda + 2).$$

Hence, the roots are $\lambda = 1$ which is a double root and $\lambda = -2$. **Eigenvalues and**

Eigenvectors for Function Spaces The concepts of eigenvalue and eigenvector make sense for a linear operator L defined on any vector space V , i.e., λ is an eigenvalue for L with eigenvector \mathbf{v} if

$$L(\mathbf{v}) = \lambda\mathbf{v} \quad \text{with } \mathbf{v} \neq \mathbf{0}.$$

If the vector space \mathbf{V} is not finite dimensional, then the use of the characteristic equation and the other methods introduced in this section do not apply, but the concepts are still very useful, and other methods may be employed to calculate them.

In particular, eigenvalues and eigenvectors arise naturally in the function spaces which occur in solving differential equations, both ordinary and partial. Thus, if you refer back to the analysis of the vibrating drum problem in Chapter IX, Section 2, you will recall that the process of separation of variables led to equations of the form

$$\begin{aligned}\Theta'' &= \mu\Theta \\ R'' + \frac{1}{r}R' - \frac{m^2}{r^2}R &= \gamma R \quad \text{where } \mu = -m^2.\end{aligned}$$

Here I took some liberties with the form of the equations in order to emphasize the relation with eigenvalues and eigenvectors. In each case, the equation has the form $L(\psi) = \lambda\psi$ where ψ denotes a function, and L is an appropriate differential operator:

$$\begin{aligned}L &= \frac{d^2}{d\theta^2} \\ L &= \frac{d^2}{dr^2} + \frac{1}{r} \frac{d}{dr} - \frac{m^2}{r^2}.\end{aligned}$$

There is one subtle but crucial point here. The allowable functions ψ in the domains of these operators are not arbitrary but have *other conditions* imposed on them by their interpretation in the underlying physical problem. Thus, we impose the periodicity condition $\Theta(\theta + 2\pi) = \Theta(\theta)$ because of the geometric meaning of the variable θ , i.e., the domain of the operator $L = \frac{d^2}{d\theta^2}$ is restricted to such periodic functions. The eigenvalue–eigenvector condition $L(\Theta) = \mu\Theta$ amounts to a differential equation which is easy to solve—see Chapter IX, Section 2—but the periodicity condition limits the choice of the eigenvalue μ to numbers of the form $\mu = -m^2$ where m is a non-negative integer. The corresponding eigenvectors (also called appropriately eigenfunctions) are the corresponding solutions of the differential equation given by

$$\Theta(\theta) = c_1 \cos m\theta + c_2 \sin m\theta.$$

Similarly, the allowable functions $R(r)$ must satisfy the boundary condition $R(a) = 0$. Solving the eigenvalue–eigenvector problem $L(R) = \gamma R$ in this case amounts to solving Bessel’s equation and finding the eigenvalues γ comes down to finding roots of Bessel functions.

This approach is commonly used in the study of partial differential equations, and you will go into it thoroughly in your course on Fourier series and boundary value problems. It is also part of the formalism used to describe quantum mechanics. In that theory, linear operators correspond to observable quantities like position, momentum, energy, etc., and the eigenvalues of these operators are the possible results of measurements of these quantities.

Exercises for 11.5.

1. Find the eigenvalues and eigenvectors for each of the following matrices. Use the method described above for solving the characteristic equation if it has degree greater than two.

(a) $\begin{bmatrix} 5 & -3 \\ 2 & 0 \end{bmatrix}$.

(b) $\begin{bmatrix} 3 & -2 & -2 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}$.

(c) $\begin{bmatrix} 2 & -1 & -1 \\ 0 & 0 & -2 \\ 0 & 1 & 3 \end{bmatrix}$.

(d) $\begin{bmatrix} 4 & -1 & -1 \\ 0 & 2 & -1 \\ 1 & 0 & 3 \end{bmatrix}$.

2. Taking A to be each of the matrices in previous exercise, use the eigenvalue-eigenvector method to find a basis for the vector space of solutions of the system $\frac{d\mathbf{x}}{dt} = A\mathbf{x}$ if it works. (It works if in the previous exercise, you found a basis for \mathbf{R}^n consisting of eigenvectors for the matrix.)

3. Solve the initial value problem

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 1 & -1 & 1 \\ 1 & -2 & -1 \\ 0 & 1 & 2 \end{bmatrix} \mathbf{x} \quad \text{where } \mathbf{x}(0) = \begin{bmatrix} 2 \\ 0 \\ 4 \end{bmatrix}.$$

Hint: One of the eigenvalues is zero.

4. Show that zero is an eigenvalue of the square matrix A if and only if $\det A = 0$, i.e., if and only if A is singular.
5. Let A be a square matrix, and suppose λ is an eigenvalue for A with eigenvector \mathbf{v} . Show that λ^2 is an eigenvalue for A^2 with eigenvector \mathbf{v} . What about λ^n and A^n for n a positive integer?
6. Suppose A is non-singular. Show that λ is an eigenvalue of A if and only if λ^{-1} is an eigenvalue of A^{-1} . Hint. Use the same eigenvector.
7. (a) Show that $\det(A - \lambda I)$ is a quadratic polynomial in λ if A is a 2×2 matrix.
 (b) Show that $\det(A - \lambda I)$ is a cubic polynomial in λ if A is a 3×3 matrix.
 (c) What would you guess is the coefficient of λ^n in $\det(A - \lambda I)$ for A an $n \times n$ matrix?
8. (Optional) Let A be an $n \times n$ matrix with entries not involving λ . Prove in general that $\det(A - \lambda I)$ is a polynomial in λ of degree n . Hint. Assume $B(\lambda)$ is an $n \times n$ matrix such that each column has at most one term involving λ

and that term is of the form $a + b\lambda$. Show by using the recursive definition of the determinant that $\det B(\lambda)$ is a polynomial in λ of degree at most n . Now use this fact and the recursive definition of the determinant to show that $\det(A - \lambda I)$ is a polynomial of degree exactly n .

9. Solve the 2×2 system in Example 2

$$\begin{aligned}\frac{dx_1}{dt} &= 2x_1 + 3x_2 \\ \frac{dx_2}{dt} &= 2x_2\end{aligned}$$

by solving the second equation and substituting back in the first equation.

10. Consider the infinite dimensional vector space of all infinitely differentiable real valued functions $u(x)$ defined for $0 \leq x \leq a$ and satisfying $u(0) = u(a) = 0$. Let $L = \frac{d^2}{dx^2}$. Find the eigenvalues and eigenvectors of the operator L .
Hint: A non-zero solution of the differential equation $u'' + \mu u = 0$ such that $u(0) = u(a) = 0$ is an eigenvector (eigenfunction) with eigenvalue $\lambda = -\mu$. You may assume for the purposes of the problem that $\lambda < 0$, i.e., $\mu > 0$.

11.6 Complex Roots

Let A be an $n \times n$ matrix. The characteristic equation

$$\det(A - \lambda I) = 0$$

is a polynomial equation of degree n in λ . The *Fundamental Theorem of Algebra* tells us that such an equation has n *complex* roots, at least if we count repeated roots with proper multiplicity. Some or all of these roots may be real, but even if A is a real matrix, some of the roots may be non-real complex numbers. For example, the characteristic equation of

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \quad \text{is } \det \begin{bmatrix} -\lambda & -1 \\ 1 & \lambda \end{bmatrix} = \lambda^2 + 1 = 0$$

which has roots $\lambda = \pm i$. We want to see in general how the nature of these roots affects the calculation of the eigenvectors of A .

First, suppose that A has some non-real complex entries, that is, not all its entries are real.

Example 219 Consider

$$A = \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix}.$$

The characteristic equation is

$$\det \begin{bmatrix} -\lambda & i \\ -i & -\lambda \end{bmatrix} = \lambda^2 - (-i^2) = \lambda^2 - 1 = 0.$$

Thus the eigenvalues are $\lambda = 1$ and $\lambda = -1$. For $\lambda = 1$, we find the eigenvectors by solving

$$\begin{bmatrix} -1 & i \\ -i & -1 \end{bmatrix} \mathbf{v} = 0.$$

Gauss-Jordan reduction yields

$$\begin{bmatrix} -1 & i \\ -i & -1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -i \\ 0 & 0 \end{bmatrix}.$$

(Multiply the first row by i and add it to the second row; then change the signs in the first row.) Thus the general solution is

$$v_1 = iv_2, \quad v_2 \text{ free.}$$

A general solution vector is

$$\mathbf{v} = \begin{bmatrix} v_2 i \\ v_2 \end{bmatrix} = v_2 \begin{bmatrix} i \\ 1 \end{bmatrix}$$

Thus a basic eigenvector for $\lambda = 1$ is

$$\mathbf{v}_1 = \begin{bmatrix} i \\ 1 \end{bmatrix}.$$

A similar calculation shows that a basic eigenvector for the eigenvalue $\lambda = -1$ is

$$\mathbf{v}_2 = \begin{bmatrix} -i \\ 1 \end{bmatrix}.$$

The above example shows that when some of the entries are non-real complex numbers, we should expect complex eigenvectors. That is, the proper domain to consider is the complex vector space \mathbf{C}^n .

Suppose instead that A has only real entries. It may still be the case that some of the roots of the characteristic equation are not real. We have two choices. We can consider only *real* roots as possible eigenvalues. For such roots λ , we may consider only *real* solutions of the system

$$(A - \lambda I)\mathbf{v} = 0.$$

That is, we choose as our domain of attention the *real* vector space \mathbf{R}^n . In effect, we act as if we don't know about complex numbers. Clearly, we will be missing something this way. We will have a better picture of what is happening if we also

consider the non-real complex roots of the characteristic equation. Doing that will ordinarily lead to complex eigenvectors, i.e., to the complex vector space \mathbf{C}^n .

Example 220 Consider

$$A = \begin{bmatrix} 2 & 1 \\ -2 & 0 \end{bmatrix}.$$

The characteristic equation is

$$\det \begin{bmatrix} 2-\lambda & 1 \\ -2 & -\lambda \end{bmatrix} = \lambda^2 - 2\lambda + 2 = 0.$$

The roots of this equation are

$$\frac{2 \pm \sqrt{4-8}}{2} = 1 \pm i.$$

Neither of these roots are real, so considering this a purely real problem in \mathbf{R}^n will yield *no* eigenvectors.

Consider it instead as a complex problem. The eigenvalue $\lambda = 1 + i$ yields the system

$$\begin{bmatrix} 1-i & 1 \\ -2 & -1-i \end{bmatrix} \mathbf{v} = 0.$$

Gauss-Jordan reduction (done carefully to account for the complex entries) yields

$$\begin{aligned} \begin{bmatrix} 1-i & 1 \\ -2 & -1-i \end{bmatrix} &\rightarrow \begin{bmatrix} 1 & (1+i)/2 \\ 1-i & 1 \end{bmatrix} && \text{switch rows, divide by } -2 \\ &\rightarrow \begin{bmatrix} 1 & (1+i)/2 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

(The calculation of the last 2, 2-entry is $1 - (1-i)\frac{(1+i)}{2} = 1 - 1 = 0$.) The general solution is

$$v_1 = -\frac{(1+i)}{2}v_2 \quad v_2 \text{ free.}$$

The general solution vector is

$$\mathbf{v} = v_2 \begin{bmatrix} -(1+i)/2 \\ 1 \end{bmatrix}.$$

Putting $v_2 = 2$ to avoid fractions yields a basic eigenvector

$$\mathbf{v}_1 = \begin{bmatrix} -1-i \\ 2 \end{bmatrix}$$

for the eigenvalue $\lambda = 1 + i$.

A similar calculation may be used to determine the eigenvectors for the eigenvalue $1 - i$. However, there is a shortcut based on the fact that the second eigenvalue

$1 - i$ is the *complex conjugate* $\bar{\lambda}$ of the first eigenvalue $1 + i$. To see how this works requires a short digression. Suppose \mathbf{v} is an eigenvector with eigenvalue λ . This means that

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

Now take the complex conjugate of everything in sight on both sides of this equation. This yields

$$\overline{A\mathbf{v}} = \overline{\lambda\mathbf{v}}.$$

(Here, putting a ‘bar’ over a matrix means that you should take the complex conjugate of every entry in the matrix.) Since A is *real*, we have $\overline{A} = A$. Thus, we have

$$A\bar{\mathbf{v}} = \bar{\lambda}\bar{\mathbf{v}}.$$

In words, *for a real $n \times n$ matrix, the complex conjugate of an eigenvector is also an eigenvector, and the eigenvalue corresponding to the latter is the complex conjugate of the eigenvalue corresponding to the former.*

Applying this principle in Example 220 yields

$$\mathbf{v}_2 = \bar{\mathbf{v}}_1 = \begin{bmatrix} -1 + i \\ 2 \end{bmatrix}$$

as a basic eigenvector for eigenvalue $\bar{\lambda} = 1 - i$.

Application to Homogeneous Linear Systems of Differential Equations

Given a system of the form $\frac{d\mathbf{x}}{dt} = A\mathbf{x}$ where A is a real $n \times n$ matrix, we know from our previous work with second order differential equations that it may be useful to consider complex valued solutions $\mathbf{x} = \mathbf{x}(t)$.

Example 220, expanded Consider the system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 2 & 1 \\ -2 & 0 \end{bmatrix} \mathbf{x}.$$

Since the roots of the characteristic equation of the coefficient matrix are complex, if we look for solutions $\mathbf{x}(t)$ taking values in \mathbf{R}^2 , we won’t get anything by the eigenvalue-eigenvector method. Hence, it makes sense to look for solutions with values in \mathbf{C}^2 . Then, according to our previous calculations, the general solution will be

$$\begin{aligned} \mathbf{x} &= c_1 e^{\lambda t} \mathbf{v}_1 + c_2 e^{\bar{\lambda} t} \bar{\mathbf{v}}_1 \\ &= c_1 e^{(1+i)t} \begin{bmatrix} -1 - i \\ 2 \end{bmatrix} + c_2 e^{(1-i)t} \begin{bmatrix} -1 + i \\ 2 \end{bmatrix}. \end{aligned}$$

As usual, the constants c_1 and c_2 are arbitrary complex scalars.

It is often the case in applications that the complex solution is meaningful in its own right, but there are also occasions where one wants a real solution. For this,

we adopt the same strategy we used when studying second order linear differential equations: *take the real and imaginary parts of the complex solution*. This will be valid if A is real since if $\mathbf{x}(t) = \mathbf{u}(t) + i\mathbf{v}(t)$ is a solution with $\mathbf{u}(t)$ and $\mathbf{v}(t)$ real, then we have

$$\begin{aligned}\frac{d\mathbf{x}}{dt} &= A\mathbf{x} \\ \frac{d\mathbf{u}}{dt} + i\frac{d\mathbf{v}}{dt} &= A\mathbf{u} + iA\mathbf{v}\end{aligned}$$

so comparing real and imaginary parts on both sides, we obtain

$$\frac{d\mathbf{u}}{dt} = A\mathbf{u} \quad \text{and} \quad \frac{d\mathbf{v}}{dt} = A\mathbf{v}.$$

Note that we needed to know A is real in order to know that $A\mathbf{u}$ and $A\mathbf{v}$ on the right are real.

Let's apply this in Example 220. One of the basic complex solutions is

$$\begin{aligned}\mathbf{x}(t) &= e^{(1+i)t} \begin{bmatrix} -1-i \\ 2 \end{bmatrix} = e^t (\cos t + i \sin t) \begin{bmatrix} -1-i \\ 2 \end{bmatrix} \\ &= e^t \begin{bmatrix} (\cos t + i \sin t)(-1-i) \\ 2 \cos t + i2 \sin t \end{bmatrix} \\ &= e^t \begin{bmatrix} -\cos t + \sin t - i(\sin t + \cos t) \\ 2 \cos t + i2 \sin t \end{bmatrix} \\ &= e^t \begin{bmatrix} -\cos t + \sin t \\ 2 \cos t \end{bmatrix} + ie^t \begin{bmatrix} -(\sin t + \cos t) \\ 2 \sin t \end{bmatrix}.\end{aligned}$$

Thus, the real and imaginary parts are

$$\begin{aligned}\mathbf{u}(t) &= e^t \begin{bmatrix} -\cos t + \sin t \\ 2 \cos t \end{bmatrix} \\ \mathbf{v}(t) &= e^t \begin{bmatrix} -(\sin t + \cos t) \\ 2 \sin t \end{bmatrix}.\end{aligned}$$

These form a linearly independent set since putting $t = t_0 = 0$ yields

$$\begin{aligned}\mathbf{u}(0) &= \begin{bmatrix} -1 \\ 2 \end{bmatrix} \\ \mathbf{v}(0) &= \begin{bmatrix} -1 \\ 0 \end{bmatrix}\end{aligned}$$

and these form a linearly independent pair in \mathbf{R}^2 . Hence, the general *real* solution of the system is

$$\mathbf{x} = c_1 e^t \begin{bmatrix} -\cos t + \sin t \\ 2 \cos t \end{bmatrix} + c_2 e^t \begin{bmatrix} -(\sin t + \cos t) \\ 2 \sin t \end{bmatrix}$$

where c_1 and c_2 are arbitrary real scalars.

Note that if we had used the eigenvalue $\bar{\lambda} = 1 - i$ and the corresponding basic complex solution $\bar{\mathbf{x}}(t) = \mathbf{u}(t) - i\mathbf{v}(t)$ instead, we would have obtained the same thing except for the sign of one of the basic real solutions.

The analysis in the example illustrates what happens in general. If A is a *real* $n \times n$ matrix, then the roots of its characteristic equation are either real or come in *conjugate complex pairs* $\lambda, \bar{\lambda}$. For real roots μ , we can always find basic eigenvectors in \mathbf{R}^n . For non-real complex roots λ , we need to look for basic eigenvectors in \mathbf{C}^n , but we may obtain the basic eigenvectors for $\bar{\lambda}$ by taking the conjugates of the basic eigenvectors for λ . Hence, for each pair of conjugate complex roots, we need only consider one root in the pair in order to generate an independent pair of real solutions.

We already know that if the eigenvalues are distinct then the corresponding set of eigenvectors will be linearly independent. Suppose that, for each pair of conjugate complex roots, we choose one root of the pair and take the real and imaginary parts of the corresponding basic eigenvectors. If we throw in the basic real eigenvectors associated with the real roots, then the set obtained in this way is always a linearly independent subset of \mathbf{R}^n . The proof of this fact is not specially difficult, but we shall skip it. (See the Exercises for special cases.)

Exercises for 11.6.

1. In each case, find the eigenvalues and eigenvectors of the given matrix. In case the characteristic equation is cubic, use the method described in the previous section to find a real (integer) root. The roots of the remaining quadratic equation will be complex.

(a) $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$

(b) $\begin{bmatrix} 1 & 1 \\ -2 & 3 \end{bmatrix}.$

(c) $\begin{bmatrix} 0 & 0 & 4 \\ 1 & 0 & -1 \\ 0 & 1 & 4 \end{bmatrix}.$

(d) $\begin{bmatrix} -1 & 1 & 7 \\ -1 & 2 & 3 \\ 0 & -1 & 2 \end{bmatrix}.$

(e) $\begin{bmatrix} 2 & 2i \\ -3i & 1 \end{bmatrix}.$

2. For A equal to each of the matrices in the previous problem, find the general complex solution of the system $\frac{d\mathbf{x}}{dt} = A\mathbf{x}$.

3. For A equal to each of the matrices in parts (a) through (d) of the previous problem, find a general real solution of the system $\frac{d\mathbf{x}}{dt} = A\mathbf{x}$.
4. Find the solution to the initial value problem

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 0 & 0 & 8 \\ 1 & 0 & -4 \\ 0 & 1 & 2 \end{bmatrix} \mathbf{x} \quad \text{where } \mathbf{x}(0) = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}.$$

Since the system is real, the final answer should be expressed in terms of real functions.

5. Suppose \mathbf{w} is a vector in \mathbf{C}^n , and write $\mathbf{w} = \mathbf{u} + i\mathbf{v}$ where \mathbf{u} and \mathbf{v} are vectors in \mathbf{R}^n . (Then $\overline{\mathbf{w}} = \mathbf{u} - i\mathbf{v}$.)

(a) Show that if $\{\mathbf{w}, \overline{\mathbf{w}}\}$ is a linearly independent pair in \mathbf{C}^n , then $\{\mathbf{u}, \mathbf{v}\}$ is a linearly independent pair in \mathbf{R}^n .

(b) Conversely, show that if $\{\mathbf{u}, \mathbf{v}\}$ is a linearly independent pair in \mathbf{R}^n , then $\{\mathbf{w}, \overline{\mathbf{w}}\}$ is a linearly independent pair in \mathbf{C}^n .

Note that for a pair of vectors in \mathbf{R}^n , you need only consider real scalars as multipliers in a dependence relation, but for a pair of vectors in \mathbf{C}^n , you would normally need to consider complex scalars as multipliers in a dependence relation.

(c) Can you invent a generalizations of (a) and (b) for sets with more than two elements?

11.7 Repeated Roots and the Exponential of a Matrix

We have noted that the eigenvalue-eigenvector method for solving the $n \times n$ system $\frac{d\mathbf{x}}{dt} = A\mathbf{x}$ succeeds if we can find a set of eigenvectors for A which forms a basis for \mathbf{R}^n or where necessary for \mathbf{C}^n . Also, according to Theorem 11.6,

this will always be the case if there are n distinct eigenvalues. Unfortunately, we still have to figure out what to do if the characteristic equation has repeated roots.

First, *we might be lucky*, and there might be a basis of eigenvectors of A .

Example 221 Consider the system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 1 & 1 & -1 \\ -1 & 3 & -1 \\ -1 & 1 & 1 \end{bmatrix} \mathbf{x}.$$

First solve the characteristic equation

$$\begin{aligned} \det \begin{bmatrix} 1-\lambda & 1 & -1 \\ -1 & 3-\lambda & -1 \\ -1 & 1 & 1-\lambda \end{bmatrix} &= \\ (1-\lambda)((3-\lambda)(1-\lambda)+1) + (1-\lambda+1) - (-1+3-\lambda) & \\ = (1-\lambda)(3-4\lambda+\lambda^2+1) + 2-\lambda-2+\lambda & \\ = (1-\lambda)(\lambda^2-4\lambda+4) & \\ = (1-\lambda)(\lambda-2)^2 = 0. & \end{aligned}$$

Note that 2 is a repeated root. We find the eigenvectors for each of these eigenvalues.

For $\lambda = 2$ we need to solve $(A - 2I)\mathbf{v} = 0$.

$$\begin{bmatrix} -1 & 1 & -1 \\ -1 & 1 & -1 \\ -1 & 1 & -1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The general solution of the system is $v_1 = v_2 - v_3$ with v_2, v_3 free. The general solution vector for that system is

$$\mathbf{v} = \begin{bmatrix} v_2 - v_3 \\ v_2 \\ v_3 \end{bmatrix} = v_2 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + v_3 \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$

The solution space is *two* dimensional. Thus, for the eigenvalue $\lambda = 2$ we obtain *two* basic eigenvectors

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix},$$

and any eigenvector for $\lambda = 2$ is a non-trivial linear combination of these.

For $\lambda = 1$, we need to solve $(A - I)\mathbf{v} = 0$.

$$\begin{bmatrix} 0 & 1 & -1 \\ -1 & 2 & -1 \\ -1 & 1 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}.$$

The general solution of the system is $v_1 = v_3, v_2 = v_3$ with v_3 free. The general solution vector is

$$\mathbf{v} = \begin{bmatrix} v_3 \\ v_3 \\ v_3 \end{bmatrix} = v_3 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

The solution space is one dimensional, and a basic eigenvector for $\lambda = 1$ is

$$\mathbf{v}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

It is not hard to check that the set of these basic eigenvectors

$$\left\{ \mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}$$

is linearly independent, so it is a basis for \mathbf{R}^3 .

We may now write out the general solution of the system of differential equations

$$\mathbf{x} = c_1 e^{2t} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + c_2 e^{2t} \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} + c_3 e^t \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Of course, we may not always be so lucky when we have repeated roots of the characteristic equation. (See for example Example 2 in Section 5.) Hence, we need some other method. It turns out that there is a generalization of the eigenvalue-eigenvector method which always works, but it requires a digression. **The Expo-**

ponential of a Matrix Let A be a constant $n \times n$ matrix. We could try to solve $d\mathbf{x}/dt = A\mathbf{x}$ by the following nonsensical calculations

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= A\mathbf{x} \\ \frac{d\mathbf{x}}{\mathbf{x}} &= A dt \\ \ln \mathbf{x} &= At + c \\ \mathbf{x} &= e^{At} e^c = e^{At} C. \end{aligned}$$

Practically every line in the above calculation contains some undefined quantity. For example, what in the world is $\frac{d\mathbf{x}}{\mathbf{x}}$? (\mathbf{x} is an $n \times 1$ column vector, so it isn't an invertible matrix.) Strangely enough something like this actually works, but one must first make some proper definitions. We start with the definition of ' e^{At} '.

Let B be any $n \times n$ matrix. We define

$$e^B = I + B + \frac{1}{2}B^2 + \frac{1}{3!}B^3 + \cdots + \frac{1}{j!}B^j + \cdots$$

A little explanation is necessary. Each term on the right is an $n \times n$ matrix. If there were only a finite number of such terms, there would be no problem, and the sum would also be an $n \times n$ matrix. In general, however, there are infinitely many terms, and we have to worry about whether it makes sense to add them up.

Example 222 Let

$$B = t \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Then

$$\begin{aligned} B^2 &= t^2 \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \\ B^3 &= t^3 \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \\ B^4 &= t^4 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ B^5 &= t^5 \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \\ &\vdots \end{aligned}$$

Hence,

$$\begin{aligned} e^B &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + t \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + \frac{1}{2}t^2 \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} + \frac{1}{3!}t^3 \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} + \cdots \\ &= \begin{bmatrix} 1 - \frac{t^2}{2} + \frac{t^4}{4!} - \cdots & t - \frac{t^3}{3!} + \frac{t^5}{5!} - \cdots \\ -t + \frac{t^3}{3!} - \frac{t^5}{5!} + \cdots & 1 - \frac{t^2}{2} + \frac{t^4}{4!} - \cdots \end{bmatrix} \\ &= \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}. \end{aligned}$$

As in the example, a series of $n \times n$ matrices yields a separate series for each of the n^2 possible entries. We shall say that such a series of matrices converges if the series it yields for each entry converges. With this rule, it is possible to show that the series defining e^B converges for any $n \times n$ matrix B , but the proof is a bit involved. Fortunately, as we shall see presently, we can usually avoid worrying about convergence by a trick. In what follows we shall generally ignore such matters and act as if the series were finite sums.

The exponential function for matrices obeys the usual rules you expect an exponential function to have, but sometimes you have to be careful.

1. If 0 denotes the $n \times n$ zero matrix, then $e^0 = I$.
2. The law of exponents holds if the matrices commute, i.e., if B and C are $n \times n$ matrices such that $BC = CB$, then $e^{B+C} = e^B e^C$.
3. If A is an $n \times n$ constant matrix, then $\frac{d}{dt}e^{At} = Ae^{At} = e^{At}A$. (It is worth writing this in both orders because products of matrices don't automatically commute.)

Here are the proofs of these facts. (1) $e^0 = I + 0 + \frac{1}{2}0^2 + \cdots = I$. (2) See the

Exercises. (3) Here we act as if the sum were finite (although the argument would work in general if we knew enough about convergence of series of matrices.)

$$\begin{aligned}
 \frac{d}{dt}e^{At} &= \frac{d}{dt} \left(I + tA + \frac{1}{2}t^2A^2 + \frac{1}{3!}t^3A^3 + \cdots + \frac{1}{j!}t^jA^j + \cdots \right) \\
 &= 0 + A + \frac{1}{2}(2t)A^2 + \frac{1}{3!}(3t^2)A^3 + \cdots + \frac{1}{j!}(jt^{j-1})A^j + \cdots \\
 &= A + tA^2 + \frac{1}{2}t^2A^3 + \cdots + \frac{1}{(j-1)!}t^{j-1}A^j + \cdots \\
 &= A(I + tA + \frac{1}{2}t^2A^2 + \cdots + \frac{1}{(j-1)!}t^{j-1}A^{j-1} + \cdots) \\
 &= Ae^{At}.
 \end{aligned}$$

Note that in the next to last step A could just as well have been factored out on the right, so it doesn't matter which side you put it on. Rule (3) gives us a formal

way to solve the system $\frac{d\mathbf{x}}{dt} = A\mathbf{x}$ when A is a constant $n \times n$ matrix. Namely, if \mathbf{v} is any constant $n \times 1$ column vector, then $\mathbf{x} = e^{At}\mathbf{v}$ is a solution. For,

$$\frac{d\mathbf{x}}{dt} = \frac{d}{dt}e^{At}\mathbf{v} = Ae^{At}\mathbf{v} = A\mathbf{x}.$$

Suppose $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is any basis for \mathbf{R}^n (or in the complex case for \mathbf{C}^n). That gives us n solutions

$$\mathbf{x}_1 = e^{At}\mathbf{v}_1, \quad \mathbf{x}_2 = e^{At}\mathbf{v}_2, \quad \dots, \quad \mathbf{x}_n = e^{At}\mathbf{v}_n.$$

Moreover, these solutions form a linearly independent set of solutions (hence a basis for the vector space of all solutions) since when we evaluate at $t = 0$, we get

$$\mathbf{x}_1(0) = e^0\mathbf{v}_1 = \mathbf{v}_1, \dots, \mathbf{x}_n(0) = e^0\mathbf{v}_n = \mathbf{v}_n.$$

By assumption, these form a linearly independent set of vectors in \mathbf{R}^n (or \mathbf{C}^n in the complex case).

The simplest choices for basis vectors are the standard basis vectors

$$\mathbf{v}_1 = \mathbf{e}_1, \mathbf{v}_2 = \mathbf{e}_2, \dots, \mathbf{v}_n = \mathbf{e}_n,$$

(which you should recall are just the columns of the identity matrix). In this case, we have $\mathbf{x}_i(t) = e^{At}\mathbf{e}_i$, which is the i th column of the $n \times n$ matrix e^{At} . Thus, *the columns of e^{At} always form a basis for the vector space of all solutions.*

Example 222, revisited For the system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \mathbf{x}$$

we have

$$e^{At} = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}$$

so

$$\begin{bmatrix} \cos t \\ -\sin t \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \sin t \\ \cos t \end{bmatrix}$$

form a basis for the solution space of the system.

There is one serious problem with the above analysis. Adding up the series for e^{At} is usually not very easy. Hence, the facts mentioned in the previous paragraphs are usually not very helpful if you want to write out an explicit solution. To get around this problem, we rely on the observation that a proper choice of \mathbf{v} can make the series

$$\begin{aligned} e^{At}\mathbf{v} &= (I + tA\frac{1}{2}t^2A^2 + \frac{1}{3!}t^3A^3 + \dots)\mathbf{v} \\ &= \mathbf{v} + t(A\mathbf{v}) + \frac{1}{2}t^2(A^2\mathbf{v}) + \frac{1}{3!}t^3(A^3\mathbf{v}) + \dots \end{aligned}$$

easier to calculate. The object then is to pick $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ with this strategy in mind.

For example, suppose \mathbf{v} is an eigenvector for A with eigenvalue λ . Then

$$\begin{aligned} A\mathbf{v} &= \lambda\mathbf{v} \\ A^2\mathbf{v} &= A(A\mathbf{v}) = A(\lambda\mathbf{v}) = \lambda(A\mathbf{v}) = \lambda^2\mathbf{v} \\ A^3\mathbf{v} &= \dots = \lambda^3\mathbf{v} \\ &\vdots \end{aligned}$$

In fact, if \mathbf{v} is an eigenvector with eigenvalue λ , it follows that $A^j\mathbf{v} = \lambda^j\mathbf{v}$ for any $j = 0, 1, 2, \dots$. Thus,

$$\begin{aligned} e^{At}\mathbf{v} &= \mathbf{v} + t(A\mathbf{v}) + \frac{1}{2}t^2(A^2\mathbf{v}) + \frac{1}{3!}t^3(A^3\mathbf{v}) + \dots \\ &= \mathbf{v} + t(\lambda\mathbf{v}) + \frac{1}{2}t^2(\lambda^2\mathbf{v}) + \frac{1}{3!}t^3(\lambda^3\mathbf{v}) + \dots \\ &= (1 + t\lambda + \frac{1}{2}t^2\lambda^2 + \frac{1}{3!}t^3\lambda^3 + \dots)\mathbf{v} \\ &= e^{\lambda t}\mathbf{v}. \end{aligned}$$

Thus, if \mathbf{v} is an eigenvector with eigenvalue λ , the series essentially reduces to the scalar series for $e^{\lambda t}$, and

$$\mathbf{x} = e^{At}\mathbf{v} = e^{\lambda t}\mathbf{v}$$

is exactly the solution obtained by the eigenvalue-eigenvector method.

Even where we don't have enough eigenvectors, we may exploit this strategy as follows. Let λ be an eigenvalue, and write

$$A = \lambda I + (A - \lambda I).$$

Then, since A and $A - \lambda I$ commute, the law of exponents tells us that

$$e^{At} = e^{\lambda It} e^{(A-\lambda I)t}.$$

However, as above,

$$\begin{aligned} e^{\lambda It} &= I + \lambda tI + \frac{1}{2}(\lambda t)^2 I^2 + \frac{1}{3!}(\lambda t)^3 I^3 + \dots \\ &= e^{\lambda t} I. \end{aligned}$$

Hence,

$$e^{At} = e^{\lambda t} I e^{(A-\lambda I)t} = e^{\lambda t} e^{(A-\lambda I)t}$$

which means that calculating $e^{At}\mathbf{v}$ can be reduced to calculating the *scalar multiplier* $e^{\lambda t}$ and the quantity $e^{(A-\lambda I)t}\mathbf{v}$. However,

$$\begin{aligned} e^{(A-\lambda I)t}\mathbf{v} &= (I + t(A - \lambda I) + \frac{1}{2}t^2(A - \lambda I)^2 + \dots + \frac{1}{j!}t^j(A - \lambda I)^j + \dots)\mathbf{v} \\ &= \mathbf{v} + t(A - \lambda I)\mathbf{v} + \frac{1}{2}t^2(A - \lambda I)^2\mathbf{v} + \dots + \frac{1}{j!}t^j(A - \lambda I)^j\mathbf{v} + \dots, \end{aligned} \tag{194}$$

so it makes sense to try to choose \mathbf{v} so that $(A - \lambda I)^j\mathbf{v}$ vanishes for all j beyond a certain point. Then the series (194) will reduce to a finite sum. In the next section, we shall explore a systematic method to do this.

Exercises for 11.7.

1. (a) Find a basis for \mathbf{R}^3 consisting of eigenvectors for

$$A = \begin{bmatrix} 1 & 2 & -4 \\ 2 & -2 & -2 \\ -4 & -2 & 1 \end{bmatrix}.$$

(b) Find a general solution of the system $\mathbf{x}' = A\mathbf{x}$ for this A .

(c) Find a solution of the system in (b) satisfying $\mathbf{x}(0) = \mathbf{e}_2$.

2. (a) Find a basis for \mathbf{R}^3 consisting of eigenvectors for

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

(b) Find a general solution of the system $\mathbf{x}' = A\mathbf{x}$ for this A .

3. (a) Let $A = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}$. Show that

$$e^{At} = \begin{bmatrix} e^{\lambda t} & 0 \\ 0 & e^{\mu t} \end{bmatrix}.$$

- (b) Let $A = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$. Such a matrix is called a *diagonal matrix*.

What can you say about e^{At} ?

4. (a) Let $A = \begin{bmatrix} \lambda & 0 \\ 1 & \lambda \end{bmatrix}$. Calculate e^{At} . Hint: use $A = \lambda I + (A - \lambda I)$.

- (b) Let $A = \begin{bmatrix} \lambda & 0 & 0 \\ 1 & \lambda & 0 \\ 0 & 1 & \lambda \end{bmatrix}$. Calculate e^{At} .

- (c) Let A be an $n \times n$ matrix of the form $\begin{bmatrix} \lambda & 0 & \dots & 0 & 0 \\ 1 & \lambda & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & \lambda \end{bmatrix}$. What is

the smallest integer k satisfying $(A - \lambda I)^k = 0$? What can you say about $e^{At} = e^{\lambda t} e^{(A - \lambda I)t}$?

5. Let A be an $n \times n$ matrix, and let P be a non-singular $n \times n$ matrix. Show that

$$Pe^{At}P^{-1} = e^{PAP^{-1}t}.$$

6. Let B and C be two $n \times n$ matrices such that $BC = CB$. Prove that

$$e^{B+C} = e^B e^C.$$

Hint: You may assume that the binomial theorem applies to commuting matrices, i.e.,

$$(B + C)^n = \sum_{i+j=n} \frac{n!}{i!j!} B^i C^j.$$

7. Let

$$B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad C = \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix}.$$

- (a) Show that $BC \neq CB$.

- (b) Show that

$$e^B = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad e^C = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}.$$

- (c) Show that $e^B e^C \neq e^{B+C}$. Hint: $B + C = J$, where e^{tJ} was calculated in the text.

11.8 Generalized Eigenvectors

Let A be an $n \times n$ matrix, and let λ be an eigenvalue for A . Suppose moreover that λ has multiplicity m as a root of the characteristic equation of A . Call any solution of the system

$$(A - \lambda I)\mathbf{v} = 0.$$

a *level 1 generalized eigenvector*. These include all the usual eigenvectors plus the zero vector. Similarly, consider the system

$$(A - \lambda I)^2 \mathbf{v} = 0,$$

and call any solution of that system a *level 2 generalized eigenvector*. Continuing in this way, call any solution of the system

$$(A - \lambda I)^j \mathbf{v} = 0$$

a *level j generalized eigenvector*. We will also sometimes just use the term *generalized eigenvector* without explicitly stating the level.

If \mathbf{v} is a level j generalized eigenvector, then

$$(A - \lambda I)^{j+1} \mathbf{v} = (A - \lambda I)(A - \lambda I)^j \mathbf{v} = 0$$

so it is also a level $j + 1$ generalized eigenvector, and similarly for all higher levels. Thus, one may envision first finding the level 1 generalized eigenvectors (i.e., the ordinary eigenvectors), then finding the level 2 vectors, which may constitute a larger set, then finding the level 3 vectors, which may constitute a still larger set, etc. That we need not continue this process indefinitely is guaranteed by the following theorem.

Theorem 11.22 Let A be an $n \times n$ matrix, and suppose λ is an eigenvalue for A with multiplicity m .

- (a) The solution space of $(A - \lambda I)^j \mathbf{v} = 0$ for $j > m$ is identical with the solution space of $(A - \lambda I)^m \mathbf{v} = 0$.
- (b) The solution space of $(A - \lambda I)^m \mathbf{v} = 0$ has dimension m .

Part (a) tells us that we need go no further than level m in order to obtain *all* generalized eigenvectors of any level whatsoever. Part (b) tells us that in some sense

there are ‘sufficiently many’ generalized eigenvectors. This will be important when we need to find a basis consisting of such vectors.

We shall not attempt to prove this theorem here. The proof is quite deep and closely related to the theory of the so-called *Jordan Canonical Form*. You will probably encounter this theory if you take a more advanced course in linear algebra.

Example 223 Let

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 1 & 3 \end{bmatrix}.$$

The eigenvalues of A are obtained by solving

$$\begin{aligned} \det \begin{bmatrix} 1-\lambda & 2 & 0 \\ 2 & 1-\lambda & 0 \\ 0 & 1 & 3-\lambda \end{bmatrix} &= (1-\lambda)((1-\lambda)(3-\lambda) - 0) - 2(2(3-\lambda) - 0) + 0 \\ &= (3-\lambda)((1-\lambda)^2 - 4) \\ &= (3-\lambda)(\lambda^2 - 2\lambda - 3) = -(\lambda - 3)^2(\lambda + 1). \end{aligned}$$

Hence, $\lambda = 3$ is a root of multiplicity 2 and $\lambda = -1$ is a root of multiplicity 1.

Let’s find the generalized eigenvectors for each eigenvalue.

For $\lambda = 3$, we need only go to level 2 and solve $(A - 3I)^2 \mathbf{v} = 0$. We have

$$(A - 3I)^2 = \begin{bmatrix} -2 & 2 & 0 \\ 2 & -2 & 0 \\ 0 & 1 & 0 \end{bmatrix}^2 = \begin{bmatrix} 8 & -8 & 0 \\ -8 & 8 & 0 \\ 2 & -2 & 0 \end{bmatrix},$$

and Gauss-Jordan reduction yields

$$\begin{bmatrix} 8 & -8 & 0 \\ -8 & 8 & 0 \\ 2 & -2 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The general solution is $v_1 = v_2$ with v_2, v_3 free. The general solution vector is

$$\mathbf{v} = \begin{bmatrix} v_2 \\ v_2 \\ v_2 \end{bmatrix} = v_2 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + v_3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Hence, a basis for the subspace of generalized eigenvectors for the eigenvalue $\lambda = 3$ is

$$\left\{ \mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

Theorem 11.22 tells us that we don’t need to go further than level 2 for the eigenvalue $\lambda = 3$, but it is reassuring to check that explicitly in this case. Look for level 3

vectors by solving $(A - 3I)^3 \mathbf{v} = 0$ ($j = m + 1 = 3$). We have

$$(A - 3I)^3 \mathbf{v} = \begin{bmatrix} -32 & 32 & 0 \\ 32 & -32 & 0 \\ -8 & 8 & 0 \end{bmatrix},$$

and it is clear that solving this system gives us exactly the same solutions as solving $(A - 2I)^2 \mathbf{v} = 0$.

For $\lambda = -1$, the multiplicity is 1, and we need to solve $(A - (-1)I)^1 \mathbf{v} = 0$. Hence, finding the generalized eigenvectors for $\lambda = -1$ just amounts to finding the usual eigenvectors.

$$\begin{bmatrix} 2 & 2 & 0 \\ 2 & 2 & 0 \\ 0 & 1 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & -4 \\ 0 & 1 & 4 \\ 0 & 0 & 0 \end{bmatrix}.$$

The general solution is $v_1 = 4v_3, v_2 = -4v_3$, with v_3 free. The general solution vector is

$$\mathbf{v} = \begin{bmatrix} 4v_3 \\ -4v_3 \\ v_3 \end{bmatrix} = v_3 \begin{bmatrix} 4 \\ -4 \\ 1 \end{bmatrix}.$$

Thus,

$$\mathbf{v}_3 = \begin{bmatrix} 4 \\ -4 \\ 1 \end{bmatrix}$$

forms a basis for the subspace of (generalized) eigenvectors for $\lambda = -1$.

Put these basic generalized eigenvectors together in a set

$$\left\{ \mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} 4 \\ -4 \\ 1 \end{bmatrix} \right\},$$

It is not hard to check by the usual means that we get a basis for \mathbf{R}^3 .

What we observed in this example always happens. If we find a basis for the subspace of generalized eigenvectors for each eigenvalue and put them together in a set, the result is always a linearly independent set. (See the first appendix to this section if you are interested in a proof.) If we are working in \mathbf{C}^n and using complex scalars, then the Fundamental Theorem of Algebra tells us that the multiplicities of the roots of the characteristic equation add up to n , the degree of the equation. Thus, the linearly independent set of basic generalized eigenvectors has the right number of elements for a basis, so it is a basis. If we are working in \mathbf{R}^n using real scalars, we will also get a basis in this way provided all the (potentially complex) roots of the characteristic equation are real. However, if there are any non-real complex roots, we will miss the corresponding generalized eigenvectors by sticking strictly to \mathbf{R}^n .

Diagonalizable Matrices In the simplest case, that in which all generalized eigenvectors are level one, the matrix A is said to be *diagonalizable*. In this case, there is a basis for \mathbf{C}^n consisting of eigenvectors for A . (If the eigenvalues and eigenvectors are all real, we could replace \mathbf{C}^n by \mathbf{R}^n in this statement.) This is certainly the easiest case to deal with, so it is not surprising that we give it a special name. However, why we use the term ‘diagonalizable’ requires an explanation.

Suppose $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is a basis for \mathbf{C}^n consisting of eigenvectors for A . Then we have

$$A\mathbf{v}_i = \lambda_i \mathbf{v}_i = \mathbf{v}_i \lambda_i, \quad i = 1, 2, \dots, n$$

where λ_i is the eigenvalue associated with \mathbf{v}_i . (These eigenvalues need not be distinct.) We may write this as a single matrix equation

$$\begin{aligned} A \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_n \end{bmatrix} &= \begin{bmatrix} \mathbf{v}_1 \lambda_1 & \mathbf{v}_2 \lambda_2 & \dots & \mathbf{v}_n \lambda_n \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}. \end{aligned}$$

If we put

$$P = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_n \end{bmatrix}$$

then this becomes

$$AP = PD$$

where D is a diagonal matrix with the eigenvalues of A appearing on the diagonal. P is an invertible matrix since its columns form a basis for \mathbf{C}^n , so the last equation can be written in turn

$$P^{-1}AP = D \quad \text{where } D \text{ is a diagonal matrix.}$$

Example In Example 207 of Section 11.2, we considered the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}.$$

The eigenvalues are $\lambda = 3$ and $\lambda = -1$. Corresponding eigenvectors are

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

and these form a basis for \mathbf{R}^2 so A is diagonalizable. Take

$$P = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

The theory predicts that $P^{-1}AP$ should be diagonal. Indeed,

$$\begin{aligned} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} &= \frac{1}{2} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} 3 & 3 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} 6 & 0 \\ 0 & -2 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}. \end{aligned}$$

Note that the eigenvalues appear on the diagonal as predicted.

Application to Homogeneous Linear Systems of Differential Equations

Let A be an $n \times n$ matrix and let \mathbf{v} be a generalized eigenvector for the eigenvalue λ . In this case, $e^{At}\mathbf{v}$ is specially easy to calculate. Let m be the multiplicity of λ . Then we know that $(A - \lambda I)^j \mathbf{v} = 0$ for $j \geq m$ (and perhaps also for some lesser powers). Then, as in the previous section,

$$e^{At} = e^{\lambda t} e^{(A - \lambda I)t},$$

but

$$e^{(A - \lambda I)t} \mathbf{v} = \mathbf{v} + t(A - \lambda I)\mathbf{v} + \frac{1}{2}t^2(A - \lambda I)^2\mathbf{v} + \cdots + \frac{1}{(m-1)!}t^{m-1}(A - \lambda I)^{m-1}\mathbf{v}$$

since *all other terms in the series vanish*. Hence,

$$e^{At}\mathbf{v} = e^{\lambda t}(\mathbf{v} + t(A - \lambda I)\mathbf{v} + \frac{1}{2}t^2(A - \lambda I)^2\mathbf{v} + \cdots + \frac{1}{(m-1)!}t^{m-1}(A - \lambda I)^{m-1}\mathbf{v}). \quad (195)$$

This gives us a method for solving a homogeneous system $\frac{d\mathbf{x}}{dt} = A\mathbf{x}$. First find a basis $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ consisting of generalized eigenvectors. (This may require working in \mathbf{C}^n rather than \mathbf{R}^n if some of the eigenvalues are non-real complex numbers.) Then, the solutions $\mathbf{x}_i = e^{At}\mathbf{v}_i$ may be calculated by formula (195), and together form a basis for the solution space.

Example 223a Consider the system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 1 & 3 \end{bmatrix} \mathbf{x}.$$

Then, as we determined above,

$$\left\{ \mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} 4 \\ -4 \\ 1 \end{bmatrix} \right\}$$

is a basis of \mathbf{R}^3 consisting of generalized eigenvectors of the coefficient matrix. The first two correspond to the eigenvalue $\lambda = 3$ with multiplicity 2, and \mathbf{v}_3 corresponds to the eigenvalue $\lambda = 1$ with multiplicity 1. For $\lambda = 3$, $m = 2$, so we only need terms up to the *first* power of $(A - 3I)$ in computing $e^{(A-3I)t}\mathbf{v}$ for $\mathbf{v} = \mathbf{v}_1$ or $\mathbf{v} = \mathbf{v}_2$.

$$(A - 3I)\mathbf{v}_1 = \begin{bmatrix} -2 & 2 & 0 \\ 2 & -2 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Thus

$$\mathbf{x}_1 = e^{At}\mathbf{v} = e^{3t} \left(\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right) = e^{3t} \begin{bmatrix} 1 \\ 1 \\ t \end{bmatrix}.$$

Similarly,

$$(A - 3I)\mathbf{v}_2 = \begin{bmatrix} -2 & 2 & 0 \\ 2 & -2 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

so \mathbf{v}_2 turns out to be an eigenvector. Thus,

$$\mathbf{x}_2 = e^{At}\mathbf{v}_2 = e^{3t}\mathbf{v}_2 = e^{3t} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

For the eigenvalue $\lambda = -1$, \mathbf{v}_3 is also an eigenvector, so the method also just gives the expected solution

$$\mathbf{x}_3 = e^{At}\mathbf{v}_3 = e^{-t}\mathbf{v}_3 = e^{-t} \begin{bmatrix} 4 \\ -4 \\ 1 \end{bmatrix}.$$

It follows that the general solution of the system is

$$\mathbf{x} = c_1 e^{3t} \begin{bmatrix} 1 \\ 1 \\ t \end{bmatrix} + c_2 e^{3t} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + c_3 e^{-t} \begin{bmatrix} 4 \\ -4 \\ 1 \end{bmatrix}.$$

The fact that \mathbf{v}_2 and \mathbf{v}_3 are eigenvectors simplifies the calculation of the solutions \mathbf{x}_2 and \mathbf{x}_3 . *This sort of simplification often happens, so you should be on the lookout for it.* \mathbf{v}_3 is an eigenvector because the associated eigenvalue has multiplicity one, but it is a bit mysterious why \mathbf{v}_2 should be an eigenvector. However, this may be clarified somewhat if you note that \mathbf{v}_2 turns up (in the process of finding $\mathbf{x}_1 = e^{At}\mathbf{v}_1$) as $\mathbf{v}_2 = (A - 3I)\mathbf{v}_1$. Thus, it is an eigenvector because $(A - 3I)\mathbf{v}_2 = (A - 3I)(A - 3I)\mathbf{v}_1 = (A - 3I)^2\mathbf{v}_1 = 0$.

Example 224 Consider the linear system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} -3 & -1 \\ 1 & -1 \end{bmatrix} \mathbf{x}.$$

The characteristic equation is

$$(3 + \lambda)(1 + \lambda) + 1 = \lambda^2 + 4\lambda + 4 = (\lambda + 2)^2 = 0.$$

Since $\lambda = -2$ is the only eigenvalue and its multiplicity is 2, it follows that *every element* of \mathbf{R}^2 is a generalized eigenvector for that eigenvalue. Hence, $\{\mathbf{e}_1, \mathbf{e}_2\}$ is a basis consisting of generalized eigenvectors. The first vector \mathbf{e}_1 leads to the basic solution

$$\begin{aligned} \mathbf{x}_1 &= e^{At} \mathbf{e}_1 = e^{-2t} e^{t(A+2I)} \mathbf{e}_1 = e^{-2t} (I + t(A+2I)) \mathbf{e}_1 \\ &= e^{-2t} (\mathbf{e}_1 + t(A+2I)\mathbf{e}_1) = e^{-2t} \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right) \\ &= e^{-2t} \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right) \\ &= e^{-2t} \begin{bmatrix} 1-t \\ t \end{bmatrix}. \end{aligned}$$

A similar calculation gives a second independent solution

$$e^{At} \mathbf{e}_2 = e^{-2t} \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix} + t \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right) = e^{-2t} \begin{bmatrix} -t \\ 1+t \end{bmatrix}.$$

However, we may simplify things somewhat as follows. Let

$$\mathbf{v}_2 = (A+2I)\mathbf{e}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Then, it is not hard to see that

$$\left\{ \mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}$$

is an independent pair of vectors. Hence, it must also necessarily be a basis for \mathbf{R}^2 . Using the basis $\{\mathbf{e}_1, \mathbf{v}_2\}$ rather than $\{\mathbf{e}_1, \mathbf{e}_2\}$ seems superficially to make things harder, but we gain something by using it since

$$(A+2I)\mathbf{v}_2 = (A+2I)(A+2I)\mathbf{e}_1 = (A+2I)^2 \mathbf{e}_1 = 0.$$

That is, \mathbf{v}_2 is an eigenvector for $\lambda = -2$. Hence, we may use the *simpler* second solution

$$\mathbf{x}_2 = e^{-2t} \mathbf{v}_2 = e^{-2t} \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Thus,

$$\begin{aligned}\mathbf{x}_1 &= e^{-2t}(\mathbf{e}_1 + t\mathbf{v}_2) = [\mathbf{e}_1 \quad \mathbf{v}_2] \begin{bmatrix} e^{-2t} \\ te^{-2t} \end{bmatrix} \\ \mathbf{x}_2 &= e^{-2t}\mathbf{v}_2 = [\mathbf{e}_1 \quad \mathbf{v}_2] \begin{bmatrix} 0 \\ e^{-2t} \end{bmatrix}\end{aligned}$$

form a basis for the solution space of the linear system of differential equations.

Example 225 Consider the system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ 0 & 1 & 2 \end{bmatrix} \mathbf{x}.$$

It is apparent that the characteristic equation is $-(\lambda - 2)^3 = 0$, so $\lambda = 2$ is the only root and has multiplicity 3. Also,

$$(A - 2I)^2 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}^2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

and necessarily $(A - 2I)^3 = 0$. Hence, *every* vector is a generalized eigenvector for A and

$$\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$$

is a perfectly good basis consisting of generalized eigenvectors.

The solutions are determined as before.

$$\begin{aligned}\mathbf{x}_1 &= e^{2t} \left(\mathbf{e}_1 + t \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \mathbf{e}_1 + \frac{1}{2}t^2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \mathbf{e}_1 \right) \\ &= e^{2t} \left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \frac{1}{2}t^2 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right) \\ &= e^{2t} \begin{bmatrix} 1 \\ t \\ t^2/2 \end{bmatrix}\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbf{x}_2 &= e^{2t} \left(\mathbf{e}_2 + t \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \mathbf{e}_2 + \frac{1}{2}t^2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \mathbf{e}_2 \right) \\ &= e^{2t} \left(\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + \frac{1}{2}t^2 \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \right) \\ &= e^{2t} \begin{bmatrix} 0 \\ 1 \\ t \end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}
 \mathbf{x}_3 &= e^{2t} \left(\mathbf{e}_3 + t \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \mathbf{e}_3 + \frac{1}{2} t^2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \mathbf{e}_3 \right) \\
 &= e^{2t} \left(\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + t \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \frac{1}{2} t^2 \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \right) \\
 &= e^{2t} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.
 \end{aligned}$$

Note that \mathbf{x}_2 only needs terms of degree 1 in t and \mathbf{x}_3 doesn't even need those since \mathbf{e}_3 is actually an eigenvector. This is not surprising since

$$\mathbf{e}_2 = (A - 2I)\mathbf{e}_1 \quad \text{so} \quad (A - 2I)^2 \mathbf{e}_2 = (A - 2I)^3 \mathbf{e}_1 = 0$$

and

$$\mathbf{e}_3 = (A - 2I)\mathbf{e}_2 \quad \text{so} \quad (A - 2I)\mathbf{e}_3 = (A - 2I)^2 \mathbf{e}_2 = 0.$$

The general solution is

$$\mathbf{x} = c_1 e^{2t} \begin{bmatrix} 1 \\ t \\ t^2/2 \end{bmatrix} + c_2 e^{2t} \begin{bmatrix} 0 \\ 1 \\ t \end{bmatrix} + c_3 e^{2t} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

It should be noted that if the matrix A is diagonalizable, then for each eigenvalue λ_i , we need only deal with genuine eigenvectors \mathbf{v}_i , and the series expansion $e^{(A-\lambda_i I)t} \mathbf{v}_i$ has only one term and reduces to $I \mathbf{v}_i = \mathbf{v}_i$. Thus, if A is diagonalizable, the generalized eigenvector method just reduces to the ordinary eigenvector method discussed earlier. **Appendix 1. Proof of Linear Independence of the Set of**

Basic Generalized Eigenvectors You may want to skip this proof.

Let $\lambda_1, \lambda_2, \dots, \lambda_k$ be distinct eigenvalues of the $n \times n$ matrix A . Suppose that, for each eigenvalue λ_i , we have chosen a basis for the subspace of solutions of the system $(A - \lambda_i I)^{m_i} \mathbf{v} = 0$, where m_i is the multiplicity of λ_i . Put these all together in a set S . We shall prove that S is a linearly independent set.

If not there is a dependence relation. Rearrange this dependence relation so that on the left we have a non-trivial linear combination of the basic generalized eigenvectors belonging to one of the eigenvalues, say it is λ_1 , and on the other side we have a linear combination of basic generalized eigenvectors for the other eigenvalues. Suppose this has the form

$$\mathbf{v}_1 = \sum_{i=2}^k \mathbf{u}_i \tag{196}$$

where $\mathbf{v}_1 \neq 0$ is a generalized eigenvector for λ_1 and each \mathbf{u}_i is a generalized eigenvector for λ_i for $i = 2, \dots, k$.

Since \mathbf{v}_1 is a generalized eigenvector for λ_1 , we have $(A - \lambda_1 I)^r \mathbf{v}_1 = 0$ for some $r > 0$. Assume r is chosen to be the *least* positive power for which that is true. Then $\mathbf{v}'_1 = (A - \lambda_1 I)^{r-1} \mathbf{v}_1 \neq 0$ and $(A - \lambda_1 I) \mathbf{v}'_1 = (A - \lambda_1 I)^r \mathbf{v}_1 = 0$, i.e., \mathbf{v}'_1 is an eigenvector with eigenvalue λ_1 . (Note that $r = 1$ is possible, so this requires that $\mathbf{v}_1 \neq 0$, which is true by assumption.) Multiply both sides of equation (196) by $(A - \lambda_1 I)^{r-1}$. On the left we get \mathbf{v}'_1 . On the right, each term $\mathbf{u}'_i = (A - \lambda_1 I)^{r-1} \mathbf{u}_i$ is still a generalized eigenvector for λ_i . For,

$$(A - \lambda_i I)^{m_i} (A - \lambda_1 I)^{r-1} \mathbf{u}_i = (A - \lambda_1 I)^{r-1} (A - \lambda_i I)^{m_i} \mathbf{u}_i = 0.$$

(This used the rather obvious fact that polynomial expressions in the matrix A commute with one another.) The upshot of this argument is that we may assume that \mathbf{v}_1 in (196) is an actual eigenvector for λ_1 . (Just replace \mathbf{v}_1 by \mathbf{v}'_1 and each \mathbf{u}_i by the corresponding \mathbf{u}'_i .)

Now multiply equation (196) by the product

$$(A - \lambda_2 I)^{m_2} (A - \lambda_3 I)^{m_3} \dots (A - \lambda_k I)^{m_k}.$$

Note as above that the factors in the product commute with one another so the order in which the terms are written is not important. Each \mathbf{u}_i on the right is a generalized eigenvector for λ_i , so $(A - \lambda_i I)^{m_i} \mathbf{u}_i = 0$. That means that the effect on the right of multiplying by the product is 0. Consider the effect on the left. \mathbf{v}_1 is an eigenvector for λ_1 , so

$$\begin{aligned} (A - \lambda_i I) \mathbf{v}_1 &= A \mathbf{v}_1 - \lambda_i \mathbf{v}_1 = \lambda_1 \mathbf{v}_1 - \lambda_i \mathbf{v}_1 = (\lambda_1 - \lambda_i) \mathbf{v}_1 \\ (A - \lambda_i I)^2 \mathbf{v}_1 &= (A - \lambda_i I)(A - \lambda_i I) \mathbf{v}_1 = (A - \lambda_i I)(\lambda_1 - \lambda_i) \mathbf{v}_1 \\ &= (\lambda_1 - \lambda_i)(A - \lambda_i I) \mathbf{v}_1 = (\lambda_1 - \lambda_i)^2 \mathbf{v}_1 \\ &\vdots \\ (A - \lambda_i)^{m_i} \mathbf{v}_1 &= (\lambda_1 - \lambda_i)^{m_i} \mathbf{v}_1. \end{aligned}$$

Thus the effect of the product on the left is

$$(\lambda_1 - \lambda_2)^{m_2} \dots (\lambda_1 - \lambda_k)^{m_k} \mathbf{v}_1$$

which is non-zero since the scalar multiplier is non-zero. This contradicts the fact that the effect on the right is zero, so we have a contradiction from the assumption that there is a dependence relation among the basic generalized eigenvectors.

Appendix 2. Cyclic Vectors and the Jordan Form You may want to come back and read this if you take a more advanced course in linear algebra.

It is a bit difficult to illustrate everything that can happen when one computes solutions $e^{At} \mathbf{v} = e^{\lambda t} e^{(A - \lambda I)t} \mathbf{v}$ as \mathbf{v} ranges over a basis of generalized eigenvectors, particularly since the degree is usually fairly small. However, there is one

phenomenon we encountered in some of the examples. It might happen that the system $(A - \lambda I)^2 \mathbf{v} = 0$ has some basic solutions of the form

$$\mathbf{v}_1, \mathbf{v}_2 = (A - \lambda I)\mathbf{v}_1$$

so \mathbf{v}_2 is an eigenvector. Similarly, it might happen that $(A - \lambda I)^3 \mathbf{v} = 0$ has some basic solutions of the form

$$\mathbf{v}_1, \mathbf{v}_2 = (A - \lambda I)\mathbf{v}_1, \mathbf{v}_3 = (A - \lambda I)\mathbf{v}_2 = (A - \lambda I)^2 \mathbf{v}_1.$$

More generally, we might be able to find a generalized eigenvector \mathbf{v} which satisfies $(A - \lambda I)^r \mathbf{v} = 0$, where the set formed from

$$\mathbf{v}, (A - \lambda I)\mathbf{v}, (A - \lambda I)^2 \mathbf{v}, \dots, (A - \lambda I)^{r-1} \mathbf{v} \quad (197)$$

is linearly independent. In this case, \mathbf{v} is called a *cyclic vector* for A of order r . Note that a cyclic vector of *order 1* is an *eigenvector*.

The theory of the *Jordan Canonical Form* asserts that it is *always* possible to find a basis of generalized eigenvectors formed from cyclic vectors as in (197). The advantage of using a basis derived from cyclic vectors is that the number of terms needed in the expansion of $e^{(A-\lambda I)t}$ is kept to a minimum.

It sometimes happens in solving $(A - \lambda I)^m \mathbf{v} = 0$ (where m is the multiplicity of the eigenvalue) that the solution method gives a ‘cyclic’ basis, but as we saw in examples it is not always the case. The best case is that in which one of the basic generalized eigenvectors for λ is cyclic of order m , i.e., it does not satisfy a lower order system $(A - \lambda I)^j \mathbf{v} = 0$ with $j < m$. However, it is quite possible that all cyclic vectors for a given eigenvalue have order smaller than m . In that case it is necessary to use two or more cyclic vectors to generate a basis. For example, for an eigenvalue λ of multiplicity $m = 3$, we could have a basis

$$\{\mathbf{v}_1, \mathbf{v}_2 = (A - \lambda I)\mathbf{v}_1, \mathbf{v}_3\}$$

where \mathbf{v}_1 is a cyclic vector of order 2 and \mathbf{v}_3 is a cyclic vector of order 1, i.e., it is an eigenvector. An even more extreme case would be a basis of eigenvectors, i.e., each basis vector would be a cyclic vector of order 1. In general, it may be quite difficult to find a basis derived from cyclic vectors.

Exercises for 11.8.

1. In each case, find a basis for \mathbf{R}^n consisting of generalized eigenvectors for the given matrix.

$$(a) \begin{bmatrix} 2 & 1 \\ -1 & 4 \end{bmatrix} \quad (b) \begin{bmatrix} 2 & -1 & 1 \\ 3 & 1 & -2 \\ 3 & -1 & 0 \end{bmatrix} \quad (c) \begin{bmatrix} 1 & -1 & 1 \\ 2 & -2 & 1 \\ 1 & -1 & 0 \end{bmatrix}$$

$$(d) \begin{bmatrix} 4 & -1 & 1 \\ 1 & 3 & 0 \\ 0 & 1 & 2 \end{bmatrix} \quad (e) \begin{bmatrix} -2 & -1 & 0 \\ 1 & 0 & 0 \\ -2 & -2 & -1 \end{bmatrix}$$

2. For A equal to each of the matrices in the previous problem, find a general solution of the system $\frac{d\mathbf{x}}{dt} = A\mathbf{x}$.
3. (a) Find a basis for \mathbf{R}^3 consisting of eigenvectors for

$$A = \begin{bmatrix} 1 & 2 & -4 \\ 2 & -2 & -2 \\ -4 & -2 & 1 \end{bmatrix}.$$

- (b) Let P be the matrix with columns the basis vectors in part (a). Calculate $P^{-1}AP$ and check that it is diagonal with the diagonal entries the eigenvalues you found.
4. Which of the following matrices are diagonalizable? Try to discover the answer without doing any significant amount of computation.

$$(a) \begin{bmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \quad (b) \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \quad (c) \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 3 \end{bmatrix}$$

5. Solve the initial value problem

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 3 & -1 & 1 & -1 \\ 1 & 1 & 2 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \mathbf{x} \quad \text{where } \mathbf{x}(0) = \begin{bmatrix} 1 \\ 0 \\ 3 \\ 2 \end{bmatrix}.$$

Hint. The only eigenvalue is $\lambda = 2$.

6. Theorem 11.7 asserts that the dimension of the solution spaces for $(A - \lambda I)^j \mathbf{v} = 0$, where λ is an eigenvalue for A of multiplicity m , increases to the maximum value m for $j = m$ and then stays constant thereafter. Without using the Theorem, show that the dimensions of these subspaces must increase to some value m' and stabilize thereafter. (In fact, $m' \leq m$, but you don't have to show that for this problem.)

Chapter 12

More about Linear Systems

12.1 The Fundamental Solution Matrix

Let A be an $n \times n$ matrix and consider the problem of solving

$$\frac{dX}{dt} = AX \quad (198)$$

where $X = X(t)$ is an $n \times n$ *matrix valued* function of the real variable t . If $X = [\mathbf{x}_1(t) \ \mathbf{x}_2(t) \ \dots \ \mathbf{x}_n(t)]$, then (198) amounts to the simultaneous consideration of n equations, one for each column of X ,

$$\frac{d\mathbf{x}_1}{dt} = A\mathbf{x}_1, \quad \frac{d\mathbf{x}_2}{dt} = A\mathbf{x}_2, \quad \dots, \quad \frac{d\mathbf{x}_n}{dt} = A\mathbf{x}_n.$$

Much of our previous discussion of systems still applies. In particular, if the entries of A are continuous functions on an interval $a < t < b$, then there is a unique solution of (198) defined on that interval which assumes a specified initial value $X(t_0)$ at some point t_0 in the interval.

This formalism gives us a way to discuss a basis $\{\mathbf{x}_1(t), \mathbf{x}_2(t), \dots, \mathbf{x}_n(t)\}$ for the vector space of solutions of the homogeneous linear system

$$\frac{d\mathbf{x}}{dt} = A\mathbf{x} \quad (199)$$

in one compact notational package. Clearly, if we want the columns of X to form such a basis, we need to assume that they constitute a linearly independent set. An $X = X(t)$ with these properties is called a *fundamental solution matrix* for the system (199). Finding a fundamental solution matrix is equivalent to finding a basis for the solution space of (199).

A fundamental solution matrix may also be used to express the general solution

$$\mathbf{x} = \mathbf{x}_1(t)c_1 + \mathbf{x}_2(t)c_2 + \cdots + \mathbf{x}_n(t)c_n = \begin{bmatrix} \mathbf{x}_1(t) & \mathbf{x}_2(t) & \cdots & \mathbf{x}_n(t) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}$$

or

$$\mathbf{x} = X(t)\mathbf{c} \quad \text{where } \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}.$$

Example 226 As in Example 1a of Section 8 of Chapter XI,

consider the system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 1 & 3 \end{bmatrix} \mathbf{x}.$$

Form the fundamental solution matrix by putting together the basic solutions in a 3×3 matrix

$$\begin{aligned} X(t) &= \begin{bmatrix} e^{3t} \begin{bmatrix} 1 \\ 1 \\ t \end{bmatrix} & e^{3t} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} & e^{-t} \begin{bmatrix} 4 \\ -4 \\ 1 \end{bmatrix} \end{bmatrix} \\ &= \begin{bmatrix} e^{3t} & 0 & 4e^{-t} \\ e^{3t} & 0 & -4e^{-t} \\ te^{3t} & e^{3t} & e^{-t} \end{bmatrix}. \end{aligned}$$

Suppose we want a solution $\mathbf{x}(t)$ satisfying

$$\mathbf{x}(0) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

This amounts to solving $X(0)\mathbf{c} = \mathbf{x}(0)$ or

$$\begin{bmatrix} 1 & 0 & 4 \\ 1 & 0 & -4 \\ 0 & 1 & 1 \end{bmatrix} \mathbf{c} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

for \mathbf{c} . We leave the details of the solution to you. The solution is $c_1 = 1/2$, $c_2 = -1/8$, and $c_3 = 1/8$. The desired solution is

$$\mathbf{x} = \frac{1}{2}e^{3t} \begin{bmatrix} 1 \\ 1 \\ t \end{bmatrix} - \frac{1}{8}e^{3t} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + \frac{1}{8}e^{-t} \begin{bmatrix} 4 \\ -4 \\ 1 \end{bmatrix}.$$

If A is a constant $n \times n$ matrix, then

$$X = e^{At}.$$

is always a fundamental solution matrix. Indeed, one way to characterize the exponential e^{At} is as the unique solution of $\frac{dX}{dt} = AX$ satisfying $X(0) = I$. However, the exponential matrix is defined as the sum of a series of matrices which is not usually easy to compute. Instead, it is usually easier to find a fundamental solution matrix $X(t)$ by some other method, and then use $X(t)$ to find e^{At} .

Theorem 12.23 Let A be an $n \times n$ matrix. If $X(t)$ is a fundamental solution matrix for the system $\frac{d}{dt}\mathbf{x} = A\mathbf{x}$, then for any initial value t_0 ,

$$X(t) = e^{A(t-t_0)}X(t_0).$$

In particular, for $t_0 = 0$,

$$X(t) = e^{At}X(0) \quad \text{or} \quad e^{At} = X(t)X(0)^{-1}. \quad (200)$$

Proof. By assumption, $Y = X(t)$ satisfies the matrix equation $\frac{dY}{dt} = AY$. However, $Y = e^{A(t-t_0)}X(t_0)$ also satisfies that equation since

$$\frac{dY}{dt} = \frac{d}{dt}e^{A(t-t_0)}X(t_0) = Ae^{A(t-t_0)}X(t_0) = AY.$$

Moreover, at $t = t_0$, these two functions agree since

$$e^{A(t_0-t_0)}X(t_0) = IX(t_0) = X(t_0).$$

Hence, by the uniqueness theorem, $X(t) = e^{A(t-t_0)}X(t_0)$ for all t . \square

Example 226, continued Let

$$X = \begin{bmatrix} e^{3t} & 0 & 4e^{-t} \\ e^{3t} & 0 & -4e^{-t} \\ te^{3t} & e^{3t} & e^{-t} \end{bmatrix}$$

be the fundamental solution matrix obtained above. We may calculate by the usual method

$$X(0)^{-1} = \begin{bmatrix} 1 & 0 & 4 \\ 1 & 0 & -4 \\ 0 & 1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{8} & \frac{1}{8} & 1 \\ \frac{1}{8} & -\frac{1}{8} & 0 \end{bmatrix}$$

so

$$\begin{aligned} e^{At} &= \begin{bmatrix} e^{3t} & 0 & 4e^{-t} \\ e^{3t} & 0 & -4e^{-t} \\ te^{3t} & e^{3t} & e^{-t} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{8} & \frac{1}{8} & 1 \\ \frac{1}{8} & -\frac{1}{8} & 0 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2}e^{3t} + \frac{1}{2}e^{-t} & \frac{1}{2}e^{3t} - \frac{1}{2}e^{-t} & 0 \\ \frac{1}{2}e^{3t} - \frac{1}{2}e^{-t} & \frac{1}{2}e^{3t} + \frac{1}{2}e^{-t} & 0 \\ \frac{1}{2}te^{3t} - \frac{1}{8}e^{3t} + \frac{1}{8}e^{-t} & \frac{1}{2}te^{3t} + \frac{1}{8}e^{3t} - \frac{1}{8}e^{-t} & e^{3t} \end{bmatrix}. \end{aligned}$$

Example 227 Consider the system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \mathbf{x}.$$

The characteristic equation of the coefficient matrix is $\lambda^2 + 1 = 0$ which has roots $\lambda = \pm i$. That means that we start by finding complex valued solutions. The eigenvalues are distinct in this case, so the eigenvalue-eigenvector method will suffice; we don't need to use generalized eigenvectors.

For $\lambda = i$, we need to solve $(A - iI)\mathbf{v} = 0$. We have

$$\begin{bmatrix} -i & 1 \\ -1 & -i \end{bmatrix} \rightarrow \begin{bmatrix} 1 & i \\ 0 & 0 \end{bmatrix}$$

so the general solution is $v_1 = -iv_2$ with v_2 free. The general solution vector is

$$\mathbf{v} = v_2 \begin{bmatrix} -i \\ 1 \end{bmatrix},$$

so

$$\mathbf{v}_1 = \begin{bmatrix} -i \\ 1 \end{bmatrix}$$

is a basic eigenvector. The corresponding solution of the differential equation is

$$\begin{aligned} \mathbf{x}_1 &= e^{it} \mathbf{v}_1 = \begin{bmatrix} -ie^{it} \\ e^{it} \end{bmatrix} \\ &= \begin{bmatrix} \sin t - i \cos t \\ \cos t + i \sin t \end{bmatrix}. \end{aligned}$$

To find independent real solutions, take the real and imaginary parts of this complex solution. They are

$$\mathbf{u} = \begin{bmatrix} \sin t \\ \cos t \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} -\cos t \\ \sin t \end{bmatrix}.$$

Hence, a fundamental solution matrix for this system is

$$X(t) = \begin{bmatrix} \sin t & -\cos t \\ \cos t & \sin t \end{bmatrix}.$$

Thus,

$$\begin{aligned} e^{\begin{bmatrix} 0 & t \\ -t & 0 \end{bmatrix}} &= \begin{bmatrix} \sin t & -\cos t \\ \cos t & \sin t \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \sin t & -\cos t \\ \cos t & \sin t \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}. \end{aligned}$$

We computed this earlier in Chapter XI, Section 7, Example 2 by adding up the series.

Wronskians

Let $\{\mathbf{x}_1(t), \mathbf{x}_2(t), \dots, \mathbf{x}_n(t)\}$ be a set of solutions of the system $d\mathbf{x}/dt = A(t)\mathbf{x}$ where $A(t)$ is an $n \times n$ matrix which is not necessarily constant. Let

$$X(t) = \begin{bmatrix} \mathbf{x}_1(t) & \mathbf{x}_2(t) & \dots & \mathbf{x}_n(t) \end{bmatrix}.$$

The quantity $W(t) = \det X(t)$ is called the *Wronskian*, and it generalizes the Wronskian for second order linear equations in one variable. It is not hard to see that $W(t)$ never vanishes if $X(t)$ is a fundamental solution matrix. For, if it vanished for a given $t = t_0$, the columns of $X(t_0)$ would form a dependent set of vectors, and this in turn would imply that the columns of $X(t)$ form a dependent set of functions of t .

It is possible to show that the Wronskian satisfies the first order differential equation

$$\frac{dW}{dt} = a(t)W,$$

where

$$a(t) = \sum_{i=1}^n a_{ii}(t)$$

is the sum of the diagonal entries of $A(t)$. (The sum of the diagonal entries of an $n \times n$ matrix A is called the *trace* of the matrix.)

We shall not use the Wronskian, but you may encounter it if you do further work with linear systems of differential equations.

Exercises for 12.1.

1. For each of the given matrices, solve the system $\mathbf{x}' = A\mathbf{x}$ and find a fundamental solution matrix.

(a) $\begin{bmatrix} 0 & -1 & -1 \\ 1 & 2 & 1 \\ -3 & 1 & -2 \end{bmatrix}.$

(b) $\begin{bmatrix} -3 & 4 \\ -1 & 1 \end{bmatrix}.$

(c) $\begin{bmatrix} 0 & 1 & 1 \\ 2 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix}.$

2. Using (200), calculate e^{At} for each of the matrices in the previous problem.

3. Calculate the Wronskian for each of the systems in the previous problem.
4. Suppose $X(t)$ is a fundamental solution matrix for the $n \times n$ system $\mathbf{x}' = A\mathbf{x}$. Show that if C is any invertible constant $n \times n$ matrix, then $X(t)C$ is also a fundamental solution matrix.

12.2 Inhomogeneous Systems

Having thoroughly explored methods for solving homogeneous systems, we now consider inhomogeneous systems

$$\frac{d\mathbf{x}}{dt} = A(t)\mathbf{x} + \mathbf{f}(t)$$

where $A(t)$ is a given $n \times n$ matrix, $\mathbf{f}(t)$ is a given vector function, and $\mathbf{x} = \mathbf{x}(t)$ as before is an vector solution to be found.

The analysis is similar to that we went through for second order inhomogeneous linear equations $y'' + p(t)y' + q(t)y = f(t)$. We proceed by first finding the general solution of the homogeneous equation

$$\frac{d\mathbf{x}}{dt} = A\mathbf{x}$$

to which is added a *particular* solution of the inhomogeneous equation. (Indeed, second order inhomogeneous linear equations may be reformulated as 2×2 inhomogeneous systems in the usual way.)

To find a particular solution of the inhomogeneous equation, we appeal to methods that worked for second order equations.

The simplest method, if it works, is guessing.

Example 228 Consider the system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \quad (201)$$

In Example 2 in the previous section, we found a general solution of the corresponding homogeneous system. Using the fundamental solution matrix e^{At} , it may be expressed

$$\mathbf{x}_h = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}.$$

To find a particular solution of (201), try a constant solution

$$\mathbf{x}_p = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

where a_1 and a_2 are to be determined. Putting this in the differential equation, we have

$$0 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

or

$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}.$$

This is a 2×2 algebraic system which is not very difficult to solve. The solution is $a_1 = 0, a_2 = -1$. Hence, the general solution of the inhomogeneous equation is

$$\begin{aligned} \mathbf{x} &= \begin{bmatrix} 0 \\ -1 \end{bmatrix} + \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ -1 \end{bmatrix} + c_1 \begin{bmatrix} \cos t \\ -\sin t \end{bmatrix} + c_2 \begin{bmatrix} \sin t \\ \cos t \end{bmatrix}. \end{aligned}$$

There is another method for finding a particular solution which is based on the method of ‘variations of parameters’. Unfortunately, it usually leads to extremely complex calculations, even when the answer is relatively simple. However, it is useful in many theoretical discussions. To apply the system version of *variation of parameters* we look for a particular solution of the form

$$\mathbf{x} = X(t)\mathbf{u}(t)$$

where $X(t)$ is a fundamental solution matrix of the corresponding homogeneous system and $\mathbf{u}(t)$ is a vector valued function to be determined. Substituting in the inhomogeneous equation yields

$$\begin{aligned} \frac{d}{dt}(X\mathbf{u}) &= AX\mathbf{u} + \mathbf{f} \\ \frac{dX}{dt}\mathbf{u} + X\frac{d\mathbf{u}}{dt} &= AX\mathbf{u} + \mathbf{f} \end{aligned}$$

However, $\frac{dX}{dt} = AX$ so the first term on each side may be canceled, and we get

$$\begin{aligned} X\frac{d\mathbf{u}}{dt} &= \mathbf{f} \\ \frac{d\mathbf{u}}{dt} &= X^{-1}\mathbf{f}. \end{aligned}$$

We may now determine \mathbf{u} by integrating both sides with respect to t . This could be done using indefinite integrals, but it is usually done with a dummy variable as

follows. Let t_0 be an initial value of t .

$$\begin{aligned}\frac{d\mathbf{u}}{ds} &= X(s)^{-1}\mathbf{f}(s) \\ \mathbf{u}(t)|_{t_0}^t &= \int_{t_0}^t X(s)^{-1}\mathbf{f}(s) ds \\ \mathbf{u}(t) - \mathbf{u}(t_0) &= \int_{t_0}^t X(s)^{-1}\mathbf{f}(s) ds \\ \mathbf{u}(t) &= \mathbf{u}(t_0) + \int_{t_0}^t X(s)^{-1}\mathbf{f}(s) ds.\end{aligned}$$

If we multiply this by $X(t)$ to obtain \mathbf{x} , we get the particular solution

$$\mathbf{x} = X(t)\mathbf{u}(t) = X(t)\mathbf{u}(t_0) + X(t) \int_{t_0}^t X(s)^{-1}\mathbf{f}(s) ds.$$

Since we only need one particular solution, we certainly should be free to choose $\mathbf{u}(t_0)$ any way we want. If we set it equal to 0, then the second term gives us the desired particular solution

$$\mathbf{x}_p = X(t) \int_{t_0}^t X(s)^{-1}\mathbf{f}(s) ds. \quad (202)$$

On the other hand, we may also write

$$\mathbf{u}(t_0) = \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}.$$

where \mathbf{c} is a vector of arbitrary constants. Then the above equation becomes

$$\mathbf{x} = X(t)\mathbf{c} + X(t) \int_{t_0}^t X(s)^{-1}\mathbf{f}(s) ds \quad (203)$$

which is the *general solution* of the inhomogeneous equation.

Example 228, again Use the same fundamental solution

$$X(t) = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}$$

as before, and take $t_0 = 0$. Then

$$\begin{aligned}X(s)^{-1} &= \begin{bmatrix} \cos s & \sin s \\ -\sin s & \cos s \end{bmatrix}^{-1} = \frac{1}{\cos^2 s + \sin^2 s} \begin{bmatrix} \cos s & -\sin s \\ \sin s & \cos s \end{bmatrix} \\ &= \begin{bmatrix} \cos s & -\sin s \\ \sin s & \cos s \end{bmatrix}.\end{aligned}$$

Thus,

$$\mathbf{x}_p = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \int_0^t \begin{bmatrix} \cos s & -\sin s \\ \sin s & \cos s \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} ds.$$

The integral is

$$\begin{aligned} \int_0^t \begin{bmatrix} \cos s \\ \sin s \end{bmatrix} ds &= \begin{bmatrix} \sin s \\ -\cos s \end{bmatrix} \Big|_0^t \\ &= \begin{bmatrix} \sin t \\ -\cos t \end{bmatrix} - \begin{bmatrix} 0 \\ -1 \end{bmatrix}. \end{aligned}$$

Multiplying this by $X(t)$ yields

$$\begin{aligned} \mathbf{x}_p &= \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \left(\begin{bmatrix} \sin t \\ -\cos t \end{bmatrix} - \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right) \\ &= \begin{bmatrix} \cos t \sin t - \sin t \cos t \\ -\sin^2 t - \cos^2 t \end{bmatrix} + \begin{bmatrix} -\sin t \\ -\cos t \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ -1 \end{bmatrix} - \begin{bmatrix} \sin t \\ \cos t \end{bmatrix}. \end{aligned}$$

The second term is a solution of the homogeneous equation. (In fact, except for the sign, it is the second column of the fundamental solution matrix.) Hence, we can drop that term, and we are left with

$$\mathbf{x}_p = \begin{bmatrix} 0 \\ -1 \end{bmatrix},$$

which is the same particular solution we obtained previously by guessing.

If A is a constant $n \times n$ matrix, then $X(t) = e^{At}$ is always a fundamental solution matrix for $d\mathbf{x}/dt = A\mathbf{x}$. Also,

$$X(s)^{-1} = (e^{As})^{-1} = e^{-As}.$$

Hence, the general solution of the inhomogeneous equation is

$$\begin{aligned} \mathbf{x} &= e^{At} \mathbf{c} + e^{At} \int_{t_0}^t e^{-As} \mathbf{f}(s) ds \\ &= e^{At} \mathbf{c} + \int_{t_0}^t e^{At} e^{-As} \mathbf{f}(s) ds \\ &= e^{At} \mathbf{c} + \int_{t_0}^t e^{A(t-s)} \mathbf{f}(s) ds. \end{aligned}$$

Moreover, if $\mathbf{x}(t)$ assumes the initial value $\mathbf{x}(t_0)$ at $t = t_0$, we have

$$\mathbf{x}(t_0) = e^{At_0} \mathbf{c} + \int_{t_0}^{t_0} (\dots) ds = e^{At_0} \mathbf{c},$$

so $\mathbf{c} = e^{-At_0}\mathbf{x}(t_0)$. Thus,

$$\begin{aligned}\mathbf{x} &= e^{At}e^{-At_0}\mathbf{x}(t_0) + \int_{t_0}^t e^{A(t-s)}\mathbf{f}(s) ds \\ &= e^{A(t-t_0)}\mathbf{x}(t_0) + \int_{t_0}^t e^{A(t-s)}\mathbf{f}(s) ds\end{aligned}$$

is a solution of the inhomogeneous equation satisfying the desired initial condition at $t = t_0$. This formula sums everything up in a neat package, but it is not specially easy to use for a variety of reasons. First, e^{At} is not usually easy to calculate, and in addition, the integration may not be specially easy to do.

(See if you can simplify the calculations in the previous example by exploiting the fact that we were using e^{At} as our fundamental solution matrix.)

Exercises for 12.2.

1. Variation of parameters requires calculation of $X(s)^{-1}$ where $X(t)$ is a fundamental solution matrix. Suppose the coefficient matrix A is constant. Derive the formula

$$X(s)^{-1} = X(0)^{-1}X(-s)X(0)^{-1}.$$

Hint: Use $X(s) = e^{As}X(0)$.

2. (a) Find a particular solution of the system

$$\mathbf{x}' = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 2 \\ 5 \end{bmatrix}$$

by guessing.

(b) Solve the corresponding homogeneous system and find a fundamental solution matrix.

(c) Find a particular solution by using (202). (Take $t_0 = 0$.)

3. Find a general solution of the system

$$\mathbf{x}' = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 2 \end{bmatrix} \mathbf{x} + e^{3t} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

4. (a) Find a general solution of the differential equation $y''' - 2y'' - 5y' + 6y = 0$. Hint: solve the system you get by putting $x_1 = y$, $x_2 = y'$, and $x_3 = y''$.

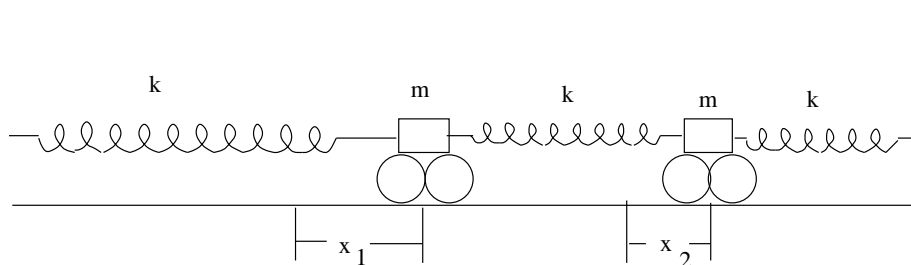
(b) Solve $y''' - 2y'' - 5y' + 6y = e^t$ given $y(0) = y'(0) = 0$, $y''(0) = 1$. Hint: use variation of parameters for the appropriate inhomogeneous system.

12.3 Normal Modes

Example 229 Recall the *second order system*

$$\begin{aligned} m \frac{d^2 x_1}{dt^2} &= -2kx_1 + kx_2 \\ m \frac{d^2 x_2}{dt^2} &= kx_1 - 2kx_2 \end{aligned}$$

which was discussed in Chapter X, Section 1, Example 2. This system arose from the configuration of particles and springs indicated below, where m is the common mass of the two particles and k is the common spring constant of the three springs.

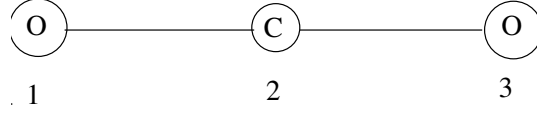


The system may also be rewritten in matrix form

$$m \frac{d^2 \mathbf{x}}{dt^2} = \begin{bmatrix} -2k & k \\ k & -2k \end{bmatrix} \mathbf{x}.$$

Systems of this kind abound in nature. For example, a molecule may be modeled as a system of particles connected by springs *provided one assumes all the displacements from equilibrium are small*. One is often very interested in determining the ways in which such a molecule may oscillate and in particular what the oscillatory frequencies are. These will tell us something about the spectral response of the molecule to infrared radiation. This *classical* model of a molecule is only an approximation, of course. and one must use quantum mechanics to get a more accurate picture of what happens. However, the classical model often illustrates important features of the problem, and it is usually more tractable mathematically.

Example 230 A CO_2 molecule may be represented as two Oxygen atoms connected by springs to a Carbon atom. In reality, the interatomic forces are quite complicated, but to a first approximation, they may be thought of as linear restoring forces produced by imaginary springs. Of course, the atoms in a real CO_2 molecule may be oriented relative to each other in space in quite complicated ways, but for the moment we consider only configurations in which all three atoms lie in a line.



If m is the mass of each Oxygen atom and m' is the mass of the Carbon atom then we have $m'/m \approx 12/16 = 3/4$. As in the diagram, let x_1 and x_3 denote the linear displacements of the two Oxygen atoms from some equilibrium position and let x_2 be the linear displacement of the Carbon atom. Then Newton's Second Law and an analysis of the forces yields the equations

$$\begin{aligned}
 m \frac{d^2 x_1}{dt^2} &= -k(x_1 - x_2) = -kx_1 + kx_2 \\
 m' \frac{d^2 x_2}{dt^2} &= -k(x_2 - x_1) - k(x_2 - x_3) = kx_1 - 2kx_2 + kx_3 \\
 m \frac{d^2 x_3}{dt^2} &= -k(x_3 - x_2) = kx_2 - kx_3
 \end{aligned}$$

which may be put in the following matrix form

$$\begin{bmatrix} m & 0 & 0 \\ 0 & m' & 0 \\ 0 & 0 & m \end{bmatrix} \frac{d^2}{dt^2} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -k & k & 0 \\ k & -2k & k \\ 0 & k & -k \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

This may be written more compactly as

$$M\mathbf{x}'' = K\mathbf{x} \quad (204)$$

where

$$M = \begin{bmatrix} m & 0 & 0 \\ 0 & m' & 0 \\ 0 & 0 & m \end{bmatrix}$$

is a *diagonal matrix* of masses, and

$$K = \begin{bmatrix} -k & k & 0 \\ k & -2k & k \\ 0 & k & -k \end{bmatrix}$$

is a matrix of spring constants. Note that K is a *symmetric matrix*, i.e. K is equal to its transpose K^t .

It will be the object of this and ensuing sections to solve second order systems of the form (204). Of course, we already have a method for doing that: convert to a first order system of twice the degree and solve that by the methods we developed in the previous chapter. It is more enlightening, however, to start from the beginning and apply the same principles directly to the second order system. As in the eigenvalue-eigenvector method for first order systems, we proceed by looking for complex vector valued solutions of the form

$$\mathbf{x} = e^{i\omega t} \mathbf{v} \quad (205)$$

where ω and $\mathbf{v} \neq 0$ are to be determined. (The rationale for replacing λ by $i\omega$ is that because of the nature of the physical problem it makes sense to look for oscillatory real solutions, and we know from our previous study of simple harmonic motion that the complex expression of such solutions will involve exponentials of the form $e^{i\omega t}$.) Then

$$\frac{d\mathbf{x}}{dt} = i\omega e^{i\omega t} \mathbf{v} \quad \text{and} \quad \frac{d^2\mathbf{x}}{dt^2} = (i\omega)^2 e^{i\omega t} \mathbf{v} = -\omega^2 e^{i\omega t} \mathbf{v}.$$

Hence, putting (205) in (204) yields

$$M(-\omega^2 e^{i\omega t} \mathbf{v}) = K e^{i\omega t} \mathbf{v}.$$

Factoring out the (non-zero) term $e^{i\omega t}$ yields in turn $-\omega^2 M\mathbf{v} = K\mathbf{v}$, which may be rewritten

$$K\mathbf{v} = \mu M\mathbf{v} \quad \text{where} \quad \mu = -\omega^2 \quad \text{and} \quad \mathbf{v} \neq \mathbf{0}. \quad (206)$$

This equation should look familiar. The quantity $\mu = -\omega^2$ looks like an eigenvalue for K , and the vector \mathbf{v} looks like an eigenvector, except of course for the presence of the diagonal matrix M . As previously, (206) may be rewritten as a system

$$(K - \mu M)\mathbf{v} = 0 \quad (207)$$

and since we need to find non-zero solutions, we need to require

$$\det(K - \mu M) = 0. \quad (208)$$

This equation is similar to the characteristic equation for K except that the identity matrix I has been replaced by the diagonal matrix M . It is called the *secular equation* of the system.

The strategy then is first to find the possible values of μ by solving (208), and for each such μ to find the possible $\mathbf{v} \neq 0$ by solving the system (207). The corresponding oscillatory (complex) solutions will be $e^{i\omega t} \mathbf{v}$ where $\omega = \sqrt{|\mu|}$.

Example 229, continued We have

$$M = mI = \begin{bmatrix} m & 0 \\ 0 & m \end{bmatrix} \quad K = k \begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix} = \begin{bmatrix} -2k & k \\ k & -2k \end{bmatrix}$$

so the secular equation is

$$\begin{aligned} \det \begin{bmatrix} -2k - m\mu & k \\ k & -2k - m\mu \end{bmatrix} &= (-2k - m\mu)^2 - k^2 \\ &= m^2\mu^2 + 2km\mu + 4k^2 - k^2 \\ &= m^2\mu^2 + 4km\mu + 3k^2 \\ &= (m\mu + k)(m\mu + 3k) = 0. \end{aligned}$$

Hence, the roots are $\mu = -k/m$ ($\omega = \sqrt{k/m}$) and $\mu = -3k/m$ ($\omega = \sqrt{3k/m}$).

For $\mu = -k/m$ ($\omega = \sqrt{k/m}$), we need to solve $(K + (k/m)M)\mathbf{v} = 0$. But

$$K + \frac{k}{m}M = \begin{bmatrix} -2k+k & k \\ k & -2k+k \end{bmatrix} = k \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}.$$

Hence, the solution is $v_1 = v_2$ with v_2 free. A basic solution vector for the subspace of solutions is

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

The corresponding complex solution is

$$e^{i\sqrt{k/m}t} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

If we take the real and imaginary parts, we get *two* real solutions.

$$\cos \sqrt{k/m}t \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \sin \sqrt{k/m}t \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

If we write the components out explicitly, we get

$$x_1 = \cos \sqrt{k/m}t, \quad x_2 = \cos \sqrt{k/m}t$$

for the first real solution, and

$$x_1 = \sin \sqrt{k/m}t, \quad x_2 = \sin \sqrt{k/m}t$$

for the second real solution. In either case, we have $x_1(t) = x_2(t)$ for all t , and the two particles move together in tandem with the same angular frequency $\sqrt{k/m}$. Note the behavior of the particles is a consequence of the fact that the components of the basic vector \mathbf{v}_1 are equal. Indeed, the same would be true for any linear combination

$$\begin{aligned} c_1 \cos \sqrt{k/m}t \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 \sin \sqrt{k/m}t \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ = (c_1 \cos \sqrt{k/m}t + c_2 \sin \sqrt{k/m}t) \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{aligned}$$

of the two real solutions obtained above. This two dimensional real subspace of solutions is called the *normal mode* of angular frequency $\sqrt{k/m}$.

Similarly, for $\mu = -3k/m$ ($\omega = \sqrt{3k/m}$), we have

$$K + 3k/mM = \begin{bmatrix} -2k+3k & k \\ k & -2k+3k \end{bmatrix} = k \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}.$$

The solution is $v_1 = -v_2$ with v_2 free, and a basic solution vector for the system is

$$\mathbf{v}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

The corresponding solution of the differential equation is

$$e^{i\sqrt{3k/m}t} \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

The corresponding normal mode is encompassed by the set of all real solutions of the form

$$(c_3 \cos \sqrt{3k/m}t + c_4 \sin \sqrt{3k/m}t) \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

which oscillate with angular frequency $\sqrt{3k/m}$.

The general real solution of the differential equation has the form

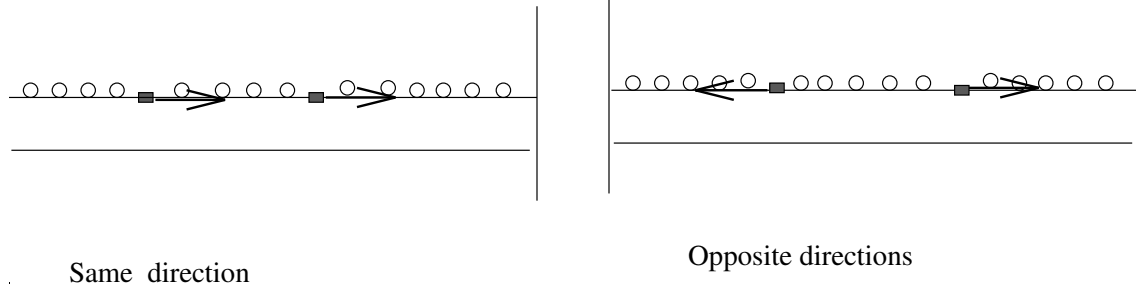
$$\begin{aligned} \mathbf{x} = & c_1 \cos \sqrt{k/m}t \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 \sin \sqrt{k/m}t \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ & + c_3 \cos \sqrt{3k/m}t \begin{bmatrix} -1 \\ 1 \end{bmatrix} + c_4 \sin \sqrt{3k/m}t \begin{bmatrix} -1 \\ 1 \end{bmatrix}. \end{aligned}$$

This example illustrates some features which need further discussion. First, we assumed implicitly in writing out the general solution that the 4 functions

$$\cos \sqrt{k/m}t \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \sin \sqrt{k/m}t \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \cos \sqrt{3k/m}t \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \sin \sqrt{3k/m}t \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

constitute a basis for the vector space of solutions. For this to be true we need to know first of all that the space of solutions is 4 dimensional, and secondly that the above functions form a linearly independent set. The first conclusion is clear if we recall that the second order system in 2 variables which we are considering is equivalent to a first order system of twice the size. Hence, the dimension of the solution space is 4. (In general, for a normal mode problem, the solution space should have dimension $2n$ where n is the number of variables.) It is not so obvious that the functions form a linearly independent set. We shall address this question in detail at the end of this section. Note, however, that the rule which worked for first order systems does not work here. If we evaluate the above vector functions at $t = 0$, we don't get a linearly independent set; in fact, two of the vectors so obtained are zero.

Another point is that we could have determined the two vectors \mathbf{v}_1 and \mathbf{v}_2 by inspection. The first corresponds to motion in which the particles move in tandem and the spring between them experiences no net change in length. The second corresponds to motion in which the particles move back and forth equal amounts in opposite directions but with the same frequency. In fact, it is often true that careful consideration of the physical arrangement of the particles, with particular attention to any symmetries that may be present, may suggest possible normal modes with little or no calculation.



Example 230, continued In this case, the secular equation is

$$\begin{aligned}
 \det(K - \mu M) &= \det \begin{bmatrix} -k - m\mu & k & 0 \\ k & -2k - \mu m' & k \\ 0 & k & -k - \mu m \end{bmatrix} \\
 &= (-m\mu - k)((-m'\mu - 2k)(-m\mu - k) - k^2) - k(k(-m\mu - k)) \\
 &= (-m\mu - k)((-m'\mu - 2k)(-m\mu - k) - 2k^2) \\
 &= (-m\mu - k)(mm'\mu^2 + k(2m + m')\mu + 2k^2 - 2k^2) \\
 &= -(m\mu + k)(mm'\mu + k(2m + m'))\mu = 0.
 \end{aligned}$$

This has 3 roots. (Don't worry about multiplicities at this point.) They are

$$\mu = -\frac{k}{m}, \quad \mu = -\frac{k(2m + m')}{mm'}, \quad \mu = 0$$

with corresponding angular frequencies

$$\omega = \sqrt{\frac{k}{m}}, \quad \omega = \sqrt{\frac{k(2m + m')}{mm'}}, \quad \omega = 0.$$

Let's find a pair of real solutions for each of these. That should provide an independent set of 6 basic real solutions, which is what we should expect since the solution space is 6 dimensional.

Start with $\mu = -k/m$ ($\omega = \sqrt{k/m}$). If we make the approximation $\frac{m'}{m} = \frac{3}{4}$, the coefficient matrix of the system $(K + k/mM)\mathbf{v} = 0$ becomes

$$\begin{aligned}
 \begin{bmatrix} -k + k & k & 0 \\ k & -2k + (k/m)m' & k \\ 0 & k & 0 \end{bmatrix} &= \begin{bmatrix} 0 & k & 0 \\ k & -2k + k(3/4) & k \\ 0 & k & 0 \end{bmatrix} \\
 &= \begin{bmatrix} 0 & k & 0 \\ k & -(5/4)k & k \\ 0 & k & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.
 \end{aligned}$$

The general solution is $v_1 = -v_3, v_2 = 0$ with v_3 free. A basic solution is

$$\mathbf{v}_1 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$

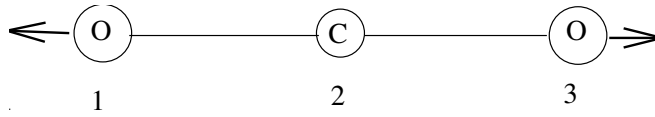
The corresponding complex solution is

$$e^{i\sqrt{k/m}t} \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$

The corresponding normal mode is describe in real terms by

$$(c_1 \cos \sqrt{k/m}t + c_2 \sin \sqrt{k/m}t) \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$

The physical interpretation is clear. The two Oxygen atoms move in equal and opposite directions while the Carbon atom stays fixed.



For $\mu = -k(2/m' + 1/m)$ ($\omega = \sqrt{k(2/m' + 1/m)}$), the coefficient matrix of the relevant system is

$$\begin{aligned} & \begin{bmatrix} -k + mk(2/m' + 1/m) & k & 0 \\ k & -2k + m'k(2/m' + 1/m) & k \\ 0 & k & -k + mk(2/m' + 1/m) \end{bmatrix} \\ &= \begin{bmatrix} -k + k(8/3 + 1) & k & 0 \\ k & -2k + k(2 + 3/4) & k \\ 0 & k & -k + k(8/3 + 1) \end{bmatrix} \\ &= k \begin{bmatrix} 8/3 & 1 & 0 \\ 1 & 3/4 & 1 \\ 0 & 1 & 8/3 \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 1 & 3/8 & 0 \\ 0 & 3/8 & 1 \\ 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 8/3 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

The corresponding solution is $v_1 = v_3, v_2 = -(8/3)v_3$ with v_3 free. The corresponding basic vector is

$$\mathbf{v}_2 = \begin{bmatrix} 1 \\ -8/3 \\ 1 \end{bmatrix},$$

and the corresponding normal mode is described in real terms by

$$(c_3 \cos \sqrt{k(2/m' + 1/m)} t + c_4 \sin \sqrt{k(2/m' + 1/m)} t) \begin{bmatrix} 1 \\ -8/3 \\ 1 \end{bmatrix}.$$

The physical interpretation is that the two Oxygen atoms move together in tandem while the Carbon atom moves in the opposite direction in such a way that the center of mass always stays fixed.



Finally, consider $\mu = 0$. This does not correspond to an oscillatory solution at all since in this case $\omega = 0$ and $e^{i\omega t} = 1$. Let's solve the system $(K - \mu M)\mathbf{v} = K\mathbf{v} = 0$ in any case, although it is not exactly clear what the physical interpretation should be.

$$\begin{bmatrix} -k & k & 0 \\ k & -2k & k \\ 0 & k & -k \end{bmatrix} \rightarrow \begin{bmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & -1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}.$$

The corresponding solution is $v_1 = v_3, v_2 = v_3$ with v_3 free. The corresponding basic vector is

$$\mathbf{v}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

but all we get this way for a real solution is

$$c_5 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

What does this mean and where is the second real solution in this case? Since all three displacements are equal, it appears that the particles are displaced an equal distance in the same direction and there is no oscillation. A little thought suggests that what this corresponds to is a uniform motion of the center of mass and no relative motion of the individual particles about the center of mass. This tells us that we should add the additional solution $t\mathbf{v}_3$, and the corresponding 'normal mode' is

$$(c_5 + c_6 t) \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

which is a mathematical description of such uniform motion. We now have a set of 6 independent real solutions.

This example also illustrates the principle that understanding the physical nature of the problem and the underlying symmetries can often lead to appropriate guesses for the vectors \mathbf{v} . Note also that once you have picked out such a vector, you can check if you are right and also determine the corresponding root $\mu = -\omega^2$ of the secular equation by using the relation

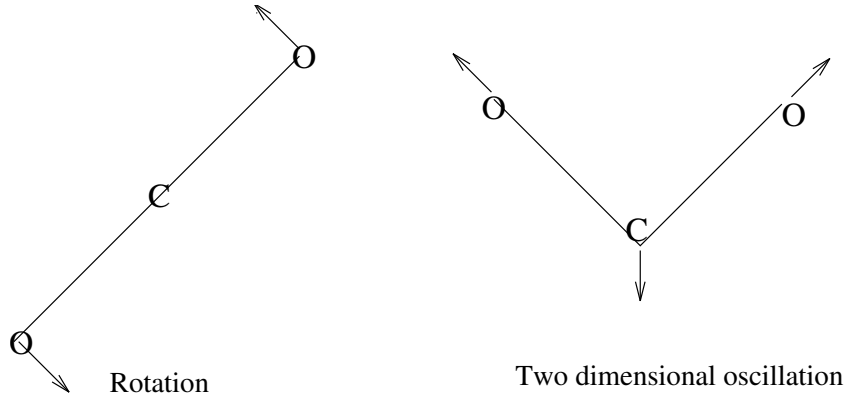
$$K\mathbf{v} = \mu M\mathbf{v}.$$

Some General Remarks The whole method depends on the fact that the roots μ of the secular equation

$$\det(K - \mu M) = 0$$

can be represented $\mu = -\omega^2$ where $\omega \geq 0$. This is the same as assuming all the roots μ are *negative* or at worst zero. However, if you pick a random symmetric $n \times n$ matrix and a random diagonal matrix M —even assume the diagonal entries of M are positive—there is no way to be sure that some of the roots μ may not be positive. Hence, for the problem to be a bona-fide normal mode problem, we must impose this as an additional assumption. This is not entirely arbitrary, however, because it can be shown from energy considerations that if some of the roots are positive, then the physical configuration will tend to fly apart instead of oscillating about a stable equilibrium.

As in Example 2, solutions associated with the root zero correspond to non-oscillatory uniform motions. However, if we allow more than one spatial dimension, the situation is much more complicated. Consider, for example, plane motions of the CO_2 molecule. For each of the three particles, there are two spatial coordinates, and so there are altogether 6 displacement variables. Thus the vector space of all solutions is 12 dimensional, and it has 6 possible basic ‘normal modes’. Some of these will involve two dimensional oscillations—see the diagram—but others will be uniform motions corresponding to a zero root of the secular equation. Some non-oscillatory solutions will consist of motion of the center of mass in some direction at a constant velocity with no relative motion of the particles about the center of mass. The problem is that other solutions will correspond to uniform *rotations* about the center of mass, but that won’t be apparent from the mathematical representation. The point is that in our analysis of the problem, we assumed that the components of the vector $\mathbf{x}(t)$ are small, since we were concentrating on oscillations. Consider for example the motion in which the two Oxygen atoms rotate at a constant rate about the Carbon atom which stays fixed at the center of mass of the system. *For small displacements*, this is not distinguishable from a solution in which each Oxygen atom starts moving perpendicular to the line between the two Oxygen atoms (passing through the Carbon atom) but in opposite directions. This is what the mathematical solution $\mathbf{x} = (a + bt)\mathbf{v}$ will appear to describe, but it is only valid ‘infinitesimally’.



Relation to Eigenvectors and Eigenvalues The normal mode problem may be restated as follows. Solve

$$\det(K - \mu M) = 0$$

and, for each root μ , find all solutions of the system

$$K\mathbf{v} = \mu M\mathbf{v}. \quad (209)$$

($\mu = -\omega^2$, but that doesn't matter here.) We noted that this looks very much like an eigenvalue-eigenvector problem, and by an appropriate change of variables, we can reduce it exactly to such a problem. Let

$$M = \begin{bmatrix} m_1 & 0 & \dots & 0 \\ 0 & m_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & m_n \end{bmatrix}.$$

Let

$$v_j = \frac{1}{\sqrt{m_j}} u_j \quad \text{for } j = 1, 2, \dots, n.$$

Thus, $u_j = \sqrt{m_j} v_j$ is weighted by the mass of the corresponding particle. This may be written in compact matrix form as

$$\mathbf{v} = (\sqrt{M})^{-1} \mathbf{u}$$

where \sqrt{M} is the matrix with diagonal entries $\sqrt{m_j}$. Putting this in (209) yields

$$K(\sqrt{M})^{-1} \mathbf{u} = \mu M(\sqrt{M})^{-1} \mathbf{u} = \sqrt{M} \mu \mathbf{u}$$

or

$$(\sqrt{M})^{-1} K (\sqrt{M})^{-1} \mathbf{u} = \mu \mathbf{u}.$$

This says that \mathbf{u} is an eigenvector for the matrix $A = (\sqrt{M})^{-1} K (\sqrt{M})^{-1}$ with μ as the eigenvalue. It is not hard to see that A is also a real symmetric matrix, so we

see that the normal mode problem is really a special case of the problem of finding eigenvalues and eigenvectors of real symmetric matrices.

Example 231 Consider a normal mode problem similar to that in Example 1 except that the second particle has mass $4m$ rather than m . Then the matrix K is the same, but

$$M = \begin{bmatrix} m & 0 \\ 0 & 4m \end{bmatrix}.$$

Hence,

$$\sqrt{M} = \begin{bmatrix} \sqrt{m} & 0 \\ 0 & 2\sqrt{m} \end{bmatrix}$$

and

$$\begin{aligned} A &= (\sqrt{M})^{-1} K (\sqrt{M})^{-1} \\ &= \begin{bmatrix} 1/\sqrt{m} & 0 \\ 0 & 1/(2\sqrt{m}) \end{bmatrix} \begin{bmatrix} -2k & k \\ k & -2k \end{bmatrix} \begin{bmatrix} 1/\sqrt{m} & 0 \\ 0 & 1/(2\sqrt{m}) \end{bmatrix} \\ &= \begin{bmatrix} -2k/m & k/(2m) \\ k/(2m) & -k/(2m) \end{bmatrix} \end{aligned}$$

Linear Independence of the Solutions Suppose that in solving the $n \times n$ normal mode problem $M\mathbf{x}'' = K\mathbf{x}$ we obtained a basis $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ for \mathbf{R}^n such that

$$K\mathbf{v}_j = -\omega_j^2 \mathbf{v}_j \quad j = 1, 2, \dots, n,$$

where each angular frequency ω_j is a non-negative root of the secular equation. We don't assume that the ω_j are distinct, and, as in Example 231, some of them may be zero. We want to show that the set of $2n$ real solutions of the normal mode problem

$$\begin{aligned} &\cos \omega_j t \mathbf{v}_j, \quad \sin \omega_j t \mathbf{v}_j \quad \text{if } \omega_j \neq 0 \\ &\text{or} \quad \mathbf{v}_j, \quad t\mathbf{v}_j \quad \text{if } \omega_j = 0 \\ &\text{for } j \text{ between 1 and } n \end{aligned}$$

is linearly independent.

Suppose not. Then there is a dependence relation of some sort. By transposing, we may assume this has the form

$$\sum_{j=1}^n (a_j \cos \omega_j t + b_j \sin \omega_j t) \mathbf{v}_j = 0, \quad (210)$$

and at least one of the coefficients $a_1, b_1, a_2, b_2, \dots, a_n, b_n$ is 1. (If $\omega_j = 0$, the appropriate term in the sum should be $(a_j + b_j t)\mathbf{v}_j$, but, as you will see, that will not affect the nature of the argument.)

Put $t = 0$ in (210). We obtain

$$\sum_{j=1}^n a_j \mathbf{v}_j = 0.$$

However, the set $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is linearly independent so the only such relation is the trivial one, i.e., we must have $a_j = 0$ for $j = 1, 2, \dots, n$. Rewrite (210)

$$\sum_{j=1}^n b_j \sin \omega_j t \mathbf{v}_j = 0.$$

(Again, if $\omega_j = 0$, the corresponding term will be $b_j t \mathbf{v}_j$ instead.) Differentiate (211) and set $t = 0$. We obtain

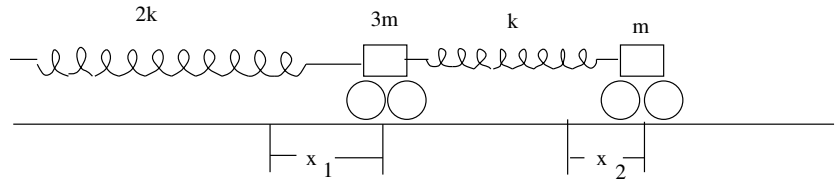
$$\sum_{j=1}^n b_j \omega_j \cos \omega_j t \mathbf{v}_j = 0.$$

(Note that if $\omega_j = 0$, the differentiated term would be $b_j \mathbf{v}_j$ without the factor of ω_j .) Again by linear independence, all the coefficients are zero. It follows that $b_j = 0$ for $j = 1, 2, \dots, n$. (Either $\omega_j \neq 0$ or the factor ω_j is not there.) This contradicts the assumption that we had a dependence relation in the first place.

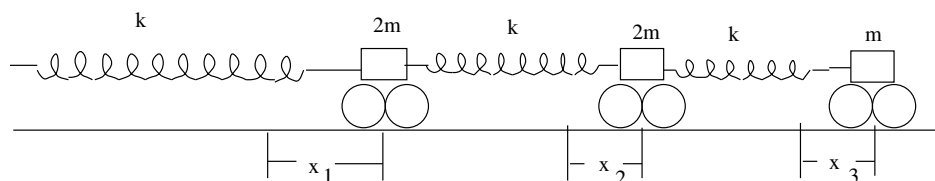
Exercises for 12.3.

1. Set up and solve the normal mode problem for each of the systems depicted below. Choose as variables the horizontal displacements from equilibrium of each of the particles. Assume these are measured so the positive direction is to the right. In each case identify the normal modes and their frequencies.

(a)



(b)



2. Consider the normal mode problem for the *plane* oscillations of a molecule consisting of three atoms of equal mass m connected by ‘springs’ with equal spring constants k . Without deriving the equations, see if you can figure out what some of the normal modes should look like. Which ‘normal modes’ will correspond to $\omega = 0$? Can you guess what the dimension of the space of all real solutions corresponding to $\omega = 0$ will be?
3. Read the proof of linear independence at the end of the section and write it out explicitly for the case of the CO_2 molecule as described in Example 2.
4. In our discussion of normal modes, we suggested a method for finding n basic complex solutions $e^{i\omega t}\mathbf{v}$ and then constructed an independent set of $2n$ basic real solutions by taking real and imaginary parts. However, there should be n additional basic complex solutions since the dimension of the vector space of all complex solutions should also be $2n$. What are those additional basic complex solutions?
5. (a) Calculate $A = (\sqrt{M})^{-1}K(\sqrt{M})^{-1}$ for K and M as in Example 2 (the CO_2 molecule).
 (b) Use $\frac{m'}{m} = \frac{3}{4}$ to simplify A , and find its eigenvalues. Check that you get the same roots as we did when we worked directly with the secular equation.
 (c) Are the eigenvectors of A the same as the solutions of $(K - \mu M)\mathbf{v} = 0$?
6. Let K be a symmetric $n \times n$ matrix and let Q be a diagonal $n \times n$ matrix. Show that $A = QKQ$ is symmetric. Hint: Calculate $(QKQ)^t$.

12.4 Real Symmetric and Complex Hermitian Matrices

We saw in the previous section that finding the normal modes of a system of particles is mathematically a special case of finding the eigenvalues and eigenvectors of a real

symmetric matrix. Many other physical problems reduce mathematically to the same problem. In this section we investigate that problem in greater detail.

Let A be a real symmetric matrix. The first thing we want to show is that *the roots of its characteristic equation*

$$\det(A - \lambda I) = 0$$

are all real. This is important in modern physics because we have to boil the predictions of a theory down to some numbers which can be checked against experiment, and it is easier if these are real. Many physical theories generate such numbers as the eigenvalues of a matrix.

It is not true that the characteristic equation of a real matrix must have real roots. (Try $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, for example, as in Chapter XI, Section 6.) Hence, the fact that the matrix is symmetric must play an important role. In order to see how this comes into play, we need a short digression.

The *dot product* in \mathbf{R}^n was defined by the formula

$$(\mathbf{u}, \mathbf{v}) = \mathbf{u}^t \mathbf{v} = \begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \sum_{j=1}^n u_j v_j. \quad (212)$$

(Note that we introduced a new notation (\mathbf{u}, \mathbf{v}) for the dot product $\mathbf{u} \cdot \mathbf{v}$. Another notation you will often see is $\langle \mathbf{u}, \mathbf{v} \rangle$.) In order to discuss complex eigenvalues, even if only to show there aren't any, we have to allow the possibility of complex eigenvectors, i.e., we need to work in \mathbf{C}^n , so we want to generalize the notion of dot product to that domain. One's first thought is to just use the same formula (212), but there is a problem with that. In \mathbf{R}^n , the length of a vector is given by

$$|\mathbf{v}|^2 = (\mathbf{v}, \mathbf{v}) = \sum_{j=1}^m v_j^2$$

and this has all the properties you expect from a good 'length function'. Unfortunately, in the complex case, the quantity $\sum_{i=1}^n v_i^2$ can vanish *without* \mathbf{v} being zero. For example, with $n = 2$, for

$$\mathbf{v} = \begin{bmatrix} i \\ 1 \end{bmatrix}, \quad \text{we have } v_1^2 + v_2^2 = i^2 + 1 = 0.$$

Clearly, it would be much better to use the formula

$$|\mathbf{v}|^2 = \sum_{j=1}^n |v_j|^2 \quad (213)$$

where $|v_j|^2 = \bar{v}_j v_j$ is the square of the absolute value of the complex number v_j . Unfortunately, this is not consistent with the definition (212) of the dot product,

but it is easy to remedy that. Define

$$(\mathbf{u}, \mathbf{v}) = \overline{\mathbf{u}}^t \mathbf{v} = \sum_{j=1}^n \overline{u}_j v_j \quad (214)$$

for two vectors \mathbf{u} and \mathbf{v} in \mathbf{C}^n . If the vectors are real this gives the same dot product as before, but it also gives

$$|\mathbf{v}|^2 = (\mathbf{v}, \mathbf{v})$$

if the left hand side is defined by (213).

We may now use this extended dot product to derive the promised result.

Theorem 12.24 Let A be a real symmetric matrix. The roots of $\det(A - \lambda I) = 0$ are all real.

Proof. Let λ be a possibly complex eigenvalue for A , i.e., assume there is a non-zero \mathbf{v} in \mathbf{C}^n such that $A\mathbf{v} = \lambda\mathbf{v}$. Consider the expression

$$(A\mathbf{v}, \mathbf{v}) = \overline{A\mathbf{v}}^t \mathbf{v}.$$

We have

$$\overline{(A\mathbf{v})}^t = (\overline{A\mathbf{v}})^t = \overline{\mathbf{v}}^t \overline{A}^t, \quad (215)$$

but since A is real and symmetric, we have

$$\overline{A}^t = A^t = A. \quad (216)$$

Hence,

$$\overline{(A\mathbf{v})}^t \mathbf{v} = \overline{\mathbf{v}}^t A\mathbf{v}.$$

Now put $A\mathbf{v} = \lambda\mathbf{v}$ in the last equation to get

$$\begin{aligned} \overline{(\lambda\mathbf{v})}^t \mathbf{v} &= \overline{\mathbf{v}}^t \lambda\mathbf{v} \\ \overline{\lambda} \overline{\mathbf{v}}^t \mathbf{v} &= \lambda \overline{\mathbf{v}}^t \mathbf{v} \end{aligned}$$

However, $\overline{\mathbf{v}}^t \mathbf{v} = (\mathbf{v}, \mathbf{v}) = |\mathbf{v}|^2 \neq 0$ since $\mathbf{v} \neq 0$. Hence,

$$\overline{\lambda} = \lambda.$$

That tells us λ is real.

Note that paradoxically we have to consider the *possibility* that λ is complex in order to show it is real. It is possible to prove this result without mentioning \mathbf{C}^n , but the argument is much more difficult. We will mention it again when we discuss the subject of Lagrange multipliers. \square \square

One crucial step in the above argument was in (216) where we concluded that

$$\overline{A}^t = A$$

from the fact that A is real and symmetric. However, the proof would work just as well if A were complex and satisfied $\overline{A}^t = A$. Such matrices are called *Hermitian* (after the 19th century French mathematician Hermite). Thus, we may extend the previous result to

Theorem 12.25 Let A be a complex Hermitian $n \times n$ matrix. Then the eigenvalues of A are real.

Example 232 The matrix

$$A = \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix}$$

which is used in quantum mechanics is Hermitian since

$$\overline{A} = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad \overline{A}^t = \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix} = A.$$

Its eigenvalues are the roots of

$$\det \begin{bmatrix} -\lambda & i \\ -i & -\lambda \end{bmatrix} = \lambda^2 - 1 = 0,$$

which are $\lambda = \pm 1$ so they are certainly real.

Note that a real $n \times n$ matrix A is Hermitian matrix if and only if it is symmetric. (The conjugation has no effect so the condition becomes $\overline{A}^t = A^t = A$.)

More about the dot product in \mathbf{C}^n The new dot product $(\mathbf{u}, \mathbf{v}) = \overline{\mathbf{u}}^t \mathbf{v}$ has all the usual properties we expect of a dot product except that because of the complex conjugation of the first factor, it obeys the rule

$$(c\mathbf{u}, \mathbf{v}) = \overline{c}(\mathbf{u}, \mathbf{v}). \quad (217)$$

However, for the second factor it obeys the usual rule $(\mathbf{u}, c\mathbf{v}) = c(\mathbf{u}, \mathbf{v})$. It is also not commutative, but obeys the following rule when the factors are switched.

$$(\mathbf{u}, \mathbf{v}) = \overline{(\mathbf{v}, \mathbf{u})}.$$

(These formulas follow easily from the definition $(\mathbf{u}, \mathbf{v}) = \overline{\mathbf{u}}^t \mathbf{v}$. See the Exercises.)

In \mathbf{R}^3 we chose the basis vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ (formerly called $\mathbf{i}, \mathbf{j}, \mathbf{k}$) by picking unit vectors along the coordinate axes. It is usual in \mathbf{R}^3 to pick mutually perpendicular coordinate axes, so the basis vectors are *mutually perpendicular unit vectors*. It makes sense to do the same thing in \mathbf{R}^n (or in the complex case \mathbf{C}^n). That is, we should attach special significance to bases $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ satisfying

$$\begin{aligned} (\mathbf{u}_i, \mathbf{u}_j) &= 0 && \text{for } i \neq j \\ |\mathbf{u}_i|^2 &= (\mathbf{u}_i, \mathbf{u}_i) = 1 && \text{otherwise.} \end{aligned}$$

Such a basis is called an *orthonormal basis*. The ‘ortho’ part of ‘orthonormal’ refers to the fact that the vectors are mutually *orthogonal*, i.e., perpendicular, and the ‘normal’ part refers to the fact that they are unit vectors.

Orthogonality plays an important role for eigenvectors of Hermitian matrices. **Theorem 12.26** Let A be a real symmetric $n \times n$ matrix or, in the complex case, a Hermitian matrix. Then eigenvectors of A associated with different eigenvalues are perpendicular.

Proof. Assume

$$A\mathbf{u} = \lambda\mathbf{u} \quad \text{and} \quad A\mathbf{v} = \mu\mathbf{v} \quad \text{where } \lambda \neq \mu.$$

We have

$$(\overline{A\mathbf{u}})^t \mathbf{v} = \overline{\mathbf{u}}^t \overline{A}^t \mathbf{v} = \overline{\mathbf{u}}^t (A\mathbf{v})$$

since $\overline{A}^t = A$. Hence,

$$\begin{aligned} \overline{\lambda\mathbf{u}}^t \mathbf{v} &= \overline{\mathbf{u}}^t (\mu\mathbf{v}) \\ \overline{\lambda}\overline{\mathbf{u}}^t \mathbf{v} &= \mu\overline{\mathbf{u}}^t \mathbf{v}. \end{aligned}$$

But the eigenvalues of A are real, so $\overline{\lambda} = \lambda$, and

$$\begin{aligned} \lambda\overline{\mathbf{u}}^t \mathbf{v} &= \mu\overline{\mathbf{u}}^t \mathbf{v} \\ (\lambda - \mu)\overline{\mathbf{u}}^t \mathbf{v} &= 0 \\ (\lambda - \mu)(\mathbf{u}, \mathbf{v}) &= 0. \end{aligned}$$

Since, $\lambda \neq \mu$, this implies that $(\mathbf{u}, \mathbf{v}) = 0$ as required. \square \square

Because of the above theorems, orthogonality plays an important role in finding the eigenvectors of a real symmetric (or a complex Hermitian) $n \times n$ matrix A . Suppose first of all that A has n distinct eigenvalues. Then we know in any case that we can choose a basis for \mathbf{R}^n (or \mathbf{C}^n in the complex case) consisting of eigenvectors for A . However, by Theorem 12.26, we know in addition in the symmetric (Hermitian) case that these basis vectors are mutually perpendicular. To get an orthonormal basis, it suffices to *normalize* each basic eigenvector by *dividing it by its length*.

Example 232, continued We saw that the eigenvalues of

$$A = \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix}$$

are $\lambda = \pm 1$.

For $\lambda = 1$, we find the eigenvectors by reducing

$$A - I = \begin{bmatrix} -1 & i \\ -i & -1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -i \\ 0 & 0 \end{bmatrix}.$$

The general solution is $v_1 = iv_2$ with v_2 free, and a basic eigenvector is

$$\mathbf{v}_1 = \begin{bmatrix} i \\ 1 \end{bmatrix}.$$

To get a unit vector, divide this by its length $|\mathbf{v}_1| = \sqrt{|i|^2 + 1^2} = \sqrt{2}$; this gives

$$\mathbf{u}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} i \\ 1 \end{bmatrix}.$$

Similarly, for $\lambda = -1$, reduce

$$A + I = \begin{bmatrix} 1 & i \\ -i & - \end{bmatrix} \rightarrow \begin{bmatrix} 1 & i \\ 0 & 0 \end{bmatrix}$$

which yields as above the unit basic eigenvector

$$\mathbf{u}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} -i \\ 1 \end{bmatrix}.$$

Note that

$$(\mathbf{u}_1, \mathbf{u}_2) = \frac{1}{2}((-i)(-i) + (1)(1)) = 0$$

as expected.

Exercises for 12.4.

1. (a) Find a basis for the subspace of \mathbf{R}^4 of all vectors perpendicular to $\mathbf{v} = \langle 1, 2, -1, 4 \rangle$. It need not be an orthonormal basis.
 (b) Find a basis for the subspace of \mathbf{C}^3 of all vectors perpendicular to $\mathbf{v} = \langle i, -i, 1 \rangle$. It need not be an orthonormal basis.
2. (Optional) Derive the following properties of the dot product in \mathbf{C}^n . Use the rules of matrix algebra derived earlier and the additional rules $\overline{B + C} = \overline{B} + \overline{C}$, $\overline{BC} = \overline{B} \overline{C}$, and $(B + C)^t = B^t + C^t$. Note also that $(\mathbf{u}^t \mathbf{v})^t = \mathbf{u}^t \mathbf{v}$ since either is a scalar.
 (a) $(\mathbf{u} + \mathbf{v}, \mathbf{w}) = (\mathbf{u}, \mathbf{v}) + (\mathbf{u}, \mathbf{w})$, $(\mathbf{u}, \mathbf{v} + \mathbf{w}) = (\mathbf{u}, \mathbf{v}) + (\mathbf{u}, \mathbf{w})$.
 (b) $(c\mathbf{u}, \mathbf{v}) = \overline{c}(\mathbf{u}, \mathbf{v})$, $(\mathbf{u}, c\mathbf{v}) = c(\mathbf{u}, \mathbf{v})$.
 (c) $(\mathbf{u}, \mathbf{v}) = \overline{(\mathbf{v}, \mathbf{u})}$.
3. Let A be Hermitian. Prove the *self-adjoint property*

$$(A\mathbf{u}, \mathbf{v}) = (\mathbf{u}, A\mathbf{v}).$$

Note. This formula was used implicitly at several points in the text. See if you can find where.

4. Let $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k\}$ be a set of mutually perpendicular non-zero vectors. Show that the set is linearly independent. Hint. Assume there is a dependence relation which after renumbering takes the form

$$\mathbf{u}_1 = c_2 \mathbf{u}_2 + \dots + c_k \mathbf{u}_k$$

and take the dot product of both sides with \mathbf{u}_1 .

5. (a) Show that the matrix $A = \begin{bmatrix} 0 & 3i & 0 \\ -3i & 0 & 4i \\ 0 & -4i & 0 \end{bmatrix}$ is Hermitian.

(b) Find the eigenvalues and eigenvectors for A .

(c) Find an orthonormal basis of eigenvectors for A .

6. (a) Find a basis of eigenvectors for $A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$.

(b) Find an orthonormal basis of eigenvectors for A .

12.5 The Principal Axis Theorem

One of the most important results in linear algebra asserts that *if A is a real symmetric (a complex Hermitian) $n \times n$ matrix then there is an orthonormal basis for \mathbf{R}^n (\mathbf{C}^n) consisting of eigenvectors for A* . This is usually called the *Principal Axis Theorem*. The reason for the name is that the case $n = 2, 3$ may be used to find the orientation or ‘principal axes’ of an arbitrary conic in the plane or quadric surface in space. There is an important generalization of the result to infinite dimensional spaces which is called the *Spectral Theorem*, so the Principal Axis Theorem is also called the ‘finite dimensional Spectral Theorem’.

In this section we shall work an example and also explore some concepts related to the use of the theorem. The proof will be deferred for the moment.

Let A be a complex Hermitian $n \times n$ matrix. (If it is real it will automatically be symmetric, so we don’t need to discuss the real case separately.) As we saw in the previous section, if the eigenvalues of A are distinct, then we already know that there is a basis consisting of eigenvectors and they are automatically perpendicular to one another. Hence, the Principal Axis Theorem really only tells us something new in case of repeated eigenvalues.

Example 233 Consider

$$A = \begin{bmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix}.$$

This example is real, so we shall work in \mathbf{R}^3 .

The characteristic equation is

$$\begin{aligned} \det \begin{bmatrix} -1-\lambda & 1 & 1 \\ 1 & -1-\lambda & 1 \\ 1 & 1 & -1-\lambda \end{bmatrix} \\ = -(1+\lambda)((1+\lambda)^2 - 1) - 1(-1-\lambda-1) + 1(1+1+\lambda) \\ = -(1+\lambda)(\lambda^2 + 2\lambda) + 2(\lambda+2) \\ = -(\lambda^3 + 3\lambda^2 - 4) = 0. \end{aligned}$$

Using the method suggested at the end of Chapter XI, Section 5, we may find the roots of this equation by trying the factors of the constant term. The roots are $\lambda = 1$, which has multiplicity 1, and $\lambda = -2$, which has multiplicity 2.

For $\lambda = 1$, we need to reduce

$$A - I = \begin{bmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & -2 \\ 0 & -3 & 3 \\ 0 & 3 & -3 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}.$$

The general solution is $v_1 = v_3, v_2 = v_3$ with v_3 free. A basic eigenvector is

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

but we should normalize this by dividing it by $|\mathbf{v}_1| = \sqrt{3}$. This gives

$$\mathbf{u}_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

For $\lambda = -2$, the situation is more complicated. Reduce

$$A + 2I = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

which yields the general solution $v_1 = -v_2 - v_3$ with v_2, v_3 free. This gives basic eigenvectors

$$\mathbf{v}_2 = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$

Unfortunately, \mathbf{v}_2 and \mathbf{v}_3 are *not perpendicular*, but this is easy to remedy. All we have to do is pick *another basis* for the subspace spanned by $\{\mathbf{v}_2, \mathbf{v}_3\}$. The

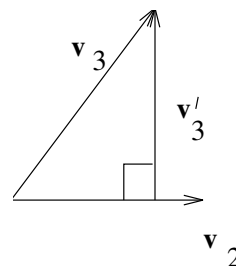
eigenvectors with eigenvalue -2 are exactly the non-zero vectors in this subspace, so any basis will do as well.

It is easy to construct the new basis. Indeed we need only replace one of the two vectors. Keep \mathbf{v}_2 , and let $\mathbf{v}'_3 = \mathbf{v}_3 - c\mathbf{v}_2$ where c is chosen so that

$$(\mathbf{v}_2, \mathbf{v}'_3) = (\mathbf{v}_2, \mathbf{v}_3) - c(\mathbf{v}_2, \mathbf{v}_2) = 0,$$

i.e., take $c = \frac{(\mathbf{v}_2, \mathbf{v}_3)}{(\mathbf{v}_2, \mathbf{v}_2)}$. (See the diagram to get some idea of the geometry behind this calculation.) We have

$$\mathbf{v}'_3 = \mathbf{v}_3 - \frac{1}{2}\mathbf{v}_2 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ 1 \end{bmatrix}.$$



We should also normalize this basis by choosing

$$\mathbf{u}_2 = \frac{1}{|\mathbf{v}_2|}\mathbf{v}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{u}_3 = \frac{1}{|\mathbf{v}'_3|}\mathbf{v}'_3 = \sqrt{\frac{2}{3}} \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ 1 \end{bmatrix}.$$

Putting this all together, we see that

$$\mathbf{u}_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{u}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{u}_3 = \sqrt{\frac{2}{3}} \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ 1 \end{bmatrix}$$

form an orthonormal basis for \mathbf{R}^3 consisting of eigenvectors for A . Notice that \mathbf{u}_1 is automatically perpendicular to \mathbf{u}_2 and \mathbf{u}_3 as the theory predicts.

The Gram-Schmidt Process In Example 233, we used a special case of a more general algorithm in order to construct an orthonormal basis of eigenvectors. The algorithm, called the *Gram-Schmidt Process* works as follows. Suppose

$$\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$$

is a *linearly independent set* spanning a certain subspace W . We construct an

orthonormal basis for W as follows. Let

$$\begin{aligned}\mathbf{v}'_1 &= \mathbf{v}_1 \\ \mathbf{v}'_2 &= \mathbf{v}_2 - \frac{(\mathbf{v}'_1, \mathbf{v}_2)}{(\mathbf{v}'_1, \mathbf{v}'_1)} \mathbf{v}'_1 \\ \mathbf{v}'_3 &= \mathbf{v}_3 - \frac{(\mathbf{v}'_1, \mathbf{v}_3)}{(\mathbf{v}'_1, \mathbf{v}'_1)} \mathbf{v}'_1 - \frac{(\mathbf{v}'_2, \mathbf{v}_3)}{(\mathbf{v}'_2, \mathbf{v}'_2)} \mathbf{v}'_2 \\ &\vdots \\ \mathbf{v}'_k &= \mathbf{v}_k - \sum_{j=1}^{k-1} \frac{(\mathbf{v}'_j, \mathbf{v}_k)}{(\mathbf{v}'_j, \mathbf{v}'_j)} \mathbf{v}'_j.\end{aligned}$$

It is not hard to see that each new \mathbf{v}'_j is perpendicular to those constructed before it. For example,

$$(\mathbf{v}'_1, \mathbf{v}'_3) = (\mathbf{v}'_1, \mathbf{v}_3) - \frac{(\mathbf{v}'_1, \mathbf{v}_3)}{(\mathbf{v}'_1, \mathbf{v}'_1)} (\mathbf{v}'_1, \mathbf{v}'_1) - \frac{(\mathbf{v}'_2, \mathbf{v}_3)}{(\mathbf{v}'_2, \mathbf{v}'_2)} (\mathbf{v}'_1, \mathbf{v}'_2).$$

However, we may suppose that we already know that $(\mathbf{v}'_1, \mathbf{v}'_2) = 0$ (from the previous stage of the construction), so the above becomes

$$(\mathbf{v}'_1, \mathbf{v}'_3) = (\mathbf{v}'_1, \mathbf{v}_3) - (\mathbf{v}'_1, \mathbf{v}_3) = 0.$$

The same argument works at each stage.

It is also not hard to see that at each stage, replacing \mathbf{v}_j by \mathbf{v}'_j in

$$\{\mathbf{v}'_1, \mathbf{v}'_2, \dots, \mathbf{v}'_{j-1}, \mathbf{v}_j\}$$

does not change the subspace spanned by the set. Hence, for $j = k$, we conclude that $\{\mathbf{v}'_1, \mathbf{v}'_2, \dots, \mathbf{v}'_k\}$ is a basis for W consisting of mutually perpendicular vectors. Finally, to complete the process simply divide each \mathbf{v}'_j by its length

$$\mathbf{u}_j = \frac{1}{|\mathbf{v}'_j|} \mathbf{v}'_j.$$

Then $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ is an orthonormal basis for W .

Example 234 Consider the subspace of \mathbf{R}^4 spanned by

$$\mathbf{v}_1 = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} -1 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Then

$$\begin{aligned}\mathbf{v}'_1 &= \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \\ \mathbf{v}'_2 &= \begin{bmatrix} -1 \\ 1 \\ 1 \\ 0 \end{bmatrix} - \frac{2}{3} \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{3} \\ \frac{1}{3} \\ 1 \\ -\frac{2}{3} \end{bmatrix} \\ \mathbf{v}'_3 &= \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} - \frac{0}{3} \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix} - \frac{-1}{\frac{15}{9}} \begin{bmatrix} -\frac{1}{3} \\ \frac{1}{3} \\ 1 \\ -\frac{2}{3} \end{bmatrix} = \begin{bmatrix} \frac{4}{3} \\ \frac{1}{3} \\ \frac{13}{3} \\ \frac{5}{3} \end{bmatrix}.\end{aligned}$$

Normalizing, we get

$$\begin{aligned}\mathbf{u}_1 &= \frac{1}{\sqrt{3}} \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \\ \mathbf{u}_2 &= \frac{3}{\sqrt{15}} \begin{bmatrix} -\frac{1}{3} \\ \frac{1}{3} \\ 1 \\ -\frac{2}{3} \end{bmatrix} = \frac{1}{\sqrt{15}} \begin{bmatrix} -1 \\ 1 \\ 3 \\ -2 \end{bmatrix} \\ \mathbf{u}_3 &= \frac{5}{\sqrt{35}} \begin{bmatrix} \frac{4}{3} \\ \frac{1}{3} \\ \frac{13}{3} \\ \frac{5}{3} \end{bmatrix} = \frac{1}{\sqrt{35}} \begin{bmatrix} 4 \\ 1 \\ 13 \\ 5 \end{bmatrix}.\end{aligned}$$

Exercises for 12.5.

1. Apply the Gram–Schmidt Process to each of the following sets of vectors.

(a) $\left\{ \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \right\}$

(b) $\left\{ \begin{bmatrix} 1 \\ 0 \\ 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 2 \\ 1 \\ -1 \end{bmatrix} \right\}.$

2. Let $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ be a linearly independent set. Suppose $\{\mathbf{v}'_1, \mathbf{v}'_2, \mathbf{v}'_3\}$ is the set obtained (before normalizing) by the Gram–Schmidt Process. Show that none of the \mathbf{v}'_j is zero.

The generalization of this to an arbitrary linearly independent set is one reason the Gram-Schmidt Process works. The vectors produced by that process are mutually perpendicular *provided they are non-zero*, and so they form a linearly independent set. Since they are in the subspace W spanned by the original set of vectors and there are just enough of them, they must form a basis for W .

3. Find an orthonormal basis of eigenvectors for $A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$.

4. Find an orthonormal basis of eigenvectors for

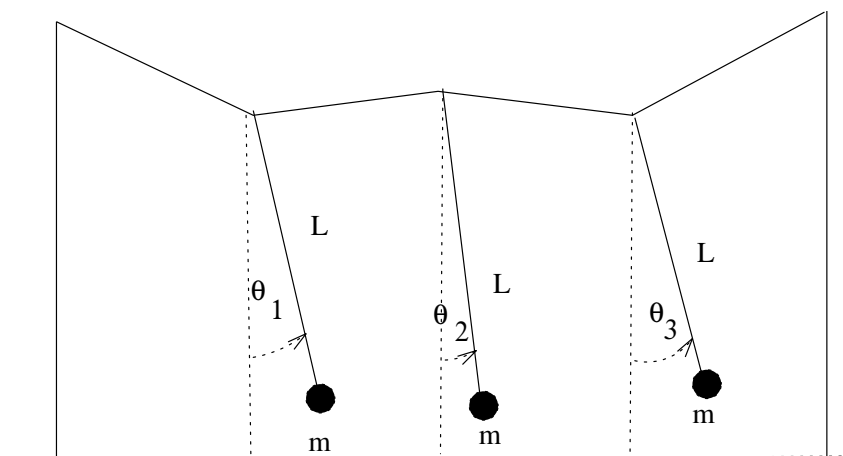
$$A = \begin{bmatrix} -1 & k & k \\ k & -1 & k \\ k & k & -1 \end{bmatrix}.$$

Hint: $2k - 1$ is an eigenvalue.

Use the results to solve the differential equation $\frac{d^2 \mathbf{x}}{dt^2} = A\mathbf{x}$ where

$$\mathbf{x} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}.$$

This system describes small oscillations of the triple pendulum system pictured below if the units are chosen so that $m = 1$ and $L = g$. (What is k ?) Find the normal modes.



Note that there are at two obvious normal modes, and if you choose the appropriate eigenvectors for those modes, you can determine the corresponding eigenvalues, one of which happens to be $2k - 1$.

12.6 Change of Coordinates and the Principal Axis Theorem

One way to understand the Principal Axis Theorem and other such theorems about a special choice of basis is to think of how a given problem would be expressed relative to that basis. For example, if we look at the linear operator L defined by $L(\mathbf{x}) = A\mathbf{x}$, then if $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is a basis of eigenvectors for A , we have by definition $L(\mathbf{v}_i) = \lambda_i \mathbf{v}_i$. Thus, for any vector \mathbf{x} , its coordinates x'_1, x'_2, \dots, x'_n with respect to this basis are the coefficients in

$$\mathbf{x} = \mathbf{v}_1 x'_1 + \mathbf{v}_2 x'_2 + \cdots + \mathbf{v}_n x'_n = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{bmatrix} \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix},$$

so we have

$$\begin{aligned} L(\mathbf{x}) &= L(\mathbf{v}_1)x'_1 + L(\mathbf{v}_2)x'_2 + \cdots + L(\mathbf{v}_n)x'_n \\ &= \mathbf{v}_1 \lambda_1 x'_1 + \mathbf{v}_2 \lambda_2 x'_2 + \cdots + \mathbf{v}_n \lambda_n x'_n = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{bmatrix} \begin{bmatrix} \lambda_1 x'_1 \\ \lambda_2 x'_2 \\ \vdots \\ \lambda_n x'_n \end{bmatrix}. \end{aligned}$$

Thus, the effect of L on the *coordinates* of a vector *with respect to such a basis* is quite simple: each coordinate is just multiplied by the corresponding eigenvalue. (See Chapter X, Section 8

to review the concept of coordinates with respect to a basis.)

To study this in greater detail, we need to talk a bit more about changes of coordinates. Although the theory is quite general, we shall concentrate on \mathbf{R}^n and \mathbf{C}^n . In either of these vector spaces, we start implicitly with the standard basis $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$. The entries in a vector \mathbf{x} may be thought of as the coordinates x_1, x_2, \dots, x_n of the vector with respect to that basis. Suppose $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is another basis. As above, the coordinates of \mathbf{x} with respect to the new basis are obtained by solving

$$\mathbf{x} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{bmatrix} \mathbf{x}' \quad (218)$$

for

$$\mathbf{x}' = \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix}.$$

Let

$$P = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{bmatrix}.$$

Then the relation (218) becomes

$$\mathbf{x} = P\mathbf{x}' \quad (219)$$

which may be thought of as a rule relating the ‘old’ coordinates of a vector to its ‘new’ coordinates. P is called the *change of coordinates matrix*, and its j th column is \mathbf{v}_j which may also be thought of as *the set of ‘old’ coordinates of the j th ‘new’ basis vector*.

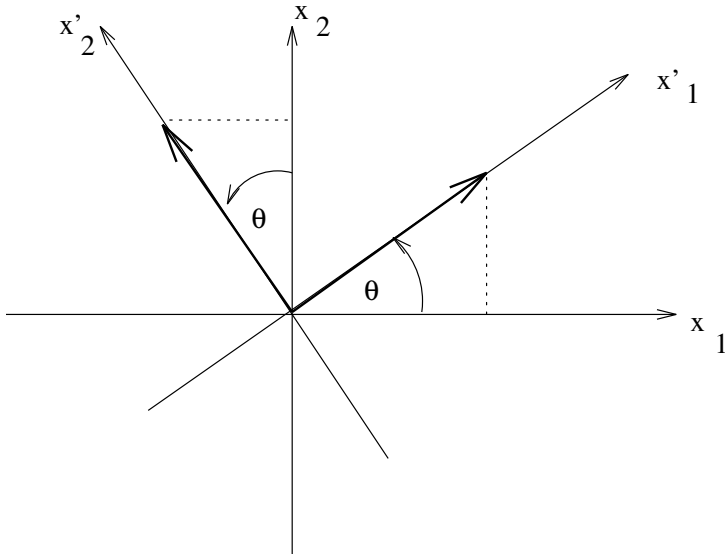
(219) is backwards in that the ‘old’ coordinates are expressed in terms of the ‘new’ coordinates. However, it is easy to turn this around. Since the columns of P are linearly independent, P is invertible and we may write instead

$$\mathbf{x}' = P^{-1}\mathbf{x}.$$

These rules have been stated for the case in which we change from the standard basis to some other basis, but they work quite generally for any change of basis. (They even work in cases where there is no obvious ‘standard basis’.) Just use the rule enunciated above: the j th column of P is the set of ‘old’ coordinates of the j th ‘new’ basis vector.

Example 235 Suppose in \mathbf{R}^2 we pick a new set of coordinate axes by rotating each of the old axes through angle θ in the counterclockwise direction. Call the old coordinates (x_1, x_2) and the new coordinates (x'_1, x'_2) . According to the above discussion, the columns of the change of basis matrix P come from the old coordinates of the new basis vectors, i.e., of unit vectors along the new axes. From the diagram, these are

$$\begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} \quad \begin{bmatrix} -\sin \theta \\ \cos \theta \end{bmatrix}.$$



Hence,

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix}.$$

The change of basis matrix is easy to invert in this case. (Use the special rule which applies to 2×2 matrices.)

$$\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}^{-1} = \frac{1}{\cos^2 \theta + \sin^2 \theta} \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}$$

(You could also have obtained this by using the matrix for rotation through angle $-\theta$.) Hence, we may express the ‘new’ coordinates in terms of the ‘old’ coordinates through the relation

$$\begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

The significance of the Principal Axis Theorem is clarified somewhat by thinking in terms of changes of coordinates. Suppose A is diagonalizable and $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is a basis of eigenvectors for A . Suppose $P = [\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n]$ is the corresponding change of basis matrix. We

showed in Chapter 11, Section 8 that

$$P^{-1}AP = D \tag{221}$$

where D is a diagonal matrix with the eigenvalues of A on the diagonal. To see how this might be used, consider a second order system of the form

$$\frac{d^2 \mathbf{x}}{dt^2} = A\mathbf{x}.$$

Assume we make the change of coordinates

$$\mathbf{x} = P\mathbf{x}'.$$

Then

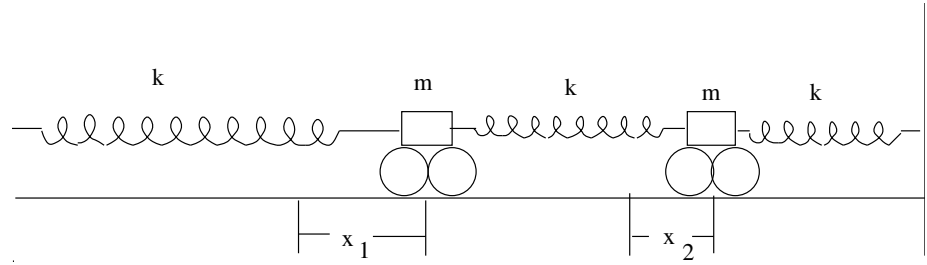
$$\begin{aligned} \frac{d^2 P\mathbf{x}'}{dt^2} &= AP\mathbf{x}' \\ P \frac{d^2 \mathbf{x}'}{dt^2} &= AP\mathbf{x}' \\ \frac{d^2 \mathbf{x}'}{dt^2} &= P^{-1}AP\mathbf{x}' = D\mathbf{x}'. \end{aligned}$$

However, since D is diagonal, this last equation may be written as n scalar equations

$$\frac{d^2 x'_j}{dt^2} = \lambda_j x'_j \quad j = 1, 2, \dots, n.$$

In the original coordinates, the motions of the particles are ‘coupled’ since the motion of each particle may affect the motion of the other particles. In the new coordinate system, these motions are ‘decoupled’. If we do this for a normal modes problem, the new coordinates are called *normal* coordinates. Each x'_j may be thought of as the displacement of one of n fictitious particles, each of which oscillates independently of the others in one of n mutually perpendicular directions. The physical significance in terms of the original particles of each normal coordinate is a bit murky, but they presumably represent underlying structure of some importance.

Example 236 Recall the normal modes problem in Section 3, Example 1.



Since the masses are equal, the problem can be reformulated as

$$\frac{d^2 \mathbf{x}}{dt^2} = \frac{k}{m} \begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix} \mathbf{x}.$$

This doesn't change anything in the solution process, and a basis of eigenvectors for the coefficient matrix is as before

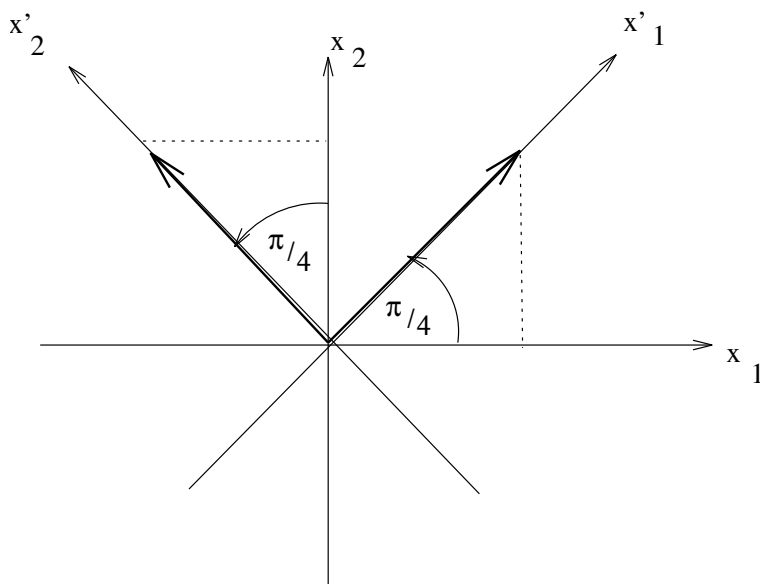
$$\left\{ \mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}.$$

If we divide the vectors by their lengths, we obtain the orthonormal basis

$$\left\{ \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}.$$

This in turn leads to the change of basis matrix

$$P = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$



If you look carefully, you will see this represents a rotation of the original x_1, x_2 -axes through an angle $\pi/4$. However, this has nothing to do with the original geometry of the problem. x_1 and x_2 stand for displacements of two different particles along the same one dimensional axis. The x_1, x_2 plane is a fictitious configuration space in which a single point represents the pair of particles. It is not absolutely clear what a rotation of axes means for this plane, but the new normal coordinates x'_1, x'_2 obtained thereby give us a formalism in which the normal modes appear as decoupled oscillations.

Orthogonal and Unitary Matrices You may have noticed that the matrix P obtained in Example 236 has the property $P^{-1} = P^t$. This is no accident. It is a consequence of the fact that its columns are mutually perpendicular unit vectors.

Theorem 12.27 Let P be an $n \times n$ real matrix. Then the columns of P form an orthonormal basis for \mathbf{R}^n if and only if $P^{-1} = P^t$. Similarly if P is an $n \times n$ complex matrix, its columns form an orthonormal basis for \mathbf{C}^n if and only if $P^{-1} = \overline{P}^t$.

A matrix with this property is called *orthogonal* in the real case and *unitary* in the complex case. The complex case subsumes the real case since a real matrix is unitary if and only if it is orthogonal.

Proof. We consider the real case. (The argument in the complex case is similar except that dot products need a complex conjugation on the first factor.) Let

$$P = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \dots \quad \mathbf{v}_n].$$

Then

$$P^t = \begin{bmatrix} \mathbf{v}_1^t \\ \mathbf{v}_2^t \\ \vdots \\ \mathbf{v}_n^t \end{bmatrix}.$$

Hence, the j, k -term of the product $P^t P$ is

$$\mathbf{v}_j^t \mathbf{v}_k = (\mathbf{v}_j, \mathbf{v}_k)$$

Thus, $P^t P = I$ if and only if

$$(\mathbf{v}_j, \mathbf{v}_k) = \delta_{jk}$$

where δ_{jk} , the ‘Kronecker δ ’, gives the entries of the identity matrix. However, this just says that the vectors are mutually perpendicular (for $j \neq k$) and have length 1 (for $j = k$). □

Since the Principal Axis Theorem asserts that there is an *orthonormal* basis consisting of eigenvectors for the Hermitian matrix A , that means that the corresponding change of basis matrix is always unitary (orthogonal in the real case). Putting this in (221), we get the following equivalent form of the Principal Axis Theorem.

Theorem 12.28 If A is a complex Hermitian $n \times n$ matrix, there is a unitary matrix P such that

$$\overline{P}^t A P = P^{-1} A P = D$$

is diagonal. If A is real symmetric, P may be chosen to be real orthogonal.

The Proof of the Principal Axis Theorem

Proof. We know that we can always find a basis for \mathbf{C}^n of generalized eigenvectors for a complex $n \times n$ matrix A . The point of the Principal Axis Theorem is that if A is Hermitian, ordinary eigenvectors suffice. The issue of orthogonality may be dealt with separately since, for a Hermitian matrix, eigenvectors for different eigenvalues are perpendicular and the Gram–Schmidt Process is available for repeated eigenvalues. Unfortunately there does not seem to be a simple direct way to eliminate the possibility of generalized eigenvectors which are not eigenvectors. The proof we shall give proceeds by induction, and it shares with many inductive proofs the feature that, while you can see that it is correct, you may not find it too enlightening. You might want to skip the proof the first time you study this material.

We give the proof in the real case. The only difference in the complex case is that you need to put complex conjugates over the appropriate terms in the formulas.

Let A be an $n \times n$ symmetric matrix. We shall show that there is a real orthogonal $n \times n$ matrix P such that

$$AP = PD \quad \text{or equivalently} \quad P^t AP = D$$

where D is a diagonal matrix with the eigenvalues of A (possibly repeated) on its diagonal.

If $n = 1$ there really isn't anything to prove. (Take $P = [1]$.) Suppose the theorem has been proved for $(n - 1) \times (n - 1)$ matrices. Let \mathbf{u}_1 be a unit eigenvector for A with eigenvalue λ_1 . Consider the subspace W consisting of all vectors perpendicular to \mathbf{u}_1 . It is not hard to see that W is an $n - 1$ dimensional subspace. Choose (by the Gram-Schmidt Process) an orthonormal basis $\{\mathbf{w}_2, \mathbf{w}_2, \dots, \mathbf{w}_n\}$ for W . Then $\{\mathbf{u}_1, \mathbf{w}_2, \dots, \mathbf{w}_n\}$ is an orthonormal basis for \mathbf{R}^n , and

$$A\mathbf{u}_1 = \mathbf{u}_1\lambda_1 = \underbrace{\begin{bmatrix} \mathbf{u}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_n \end{bmatrix}}_{P_1} \begin{bmatrix} \lambda_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

This gives the first column of AP_1 , and we want to say something about its remaining columns

$$A\mathbf{w}_2, \quad A\mathbf{w}_2, \dots, A\mathbf{w}_n.$$

To this end, note that if \mathbf{w} is any vector in W , then $A\mathbf{w}$ is also a vector in W . For, starting with the *self adjoint property* (Section 4, Problem A3),

we have

$$(\mathbf{u}_1, A\mathbf{w}) = (A\mathbf{u}_1, \mathbf{w}) = (\lambda_1\mathbf{u}_1, \mathbf{w}) = \lambda_1(\mathbf{u}_1, \mathbf{w}) = 0,$$

which is to say, $A\mathbf{w}$ is perpendicular to \mathbf{u}_1 if \mathbf{w} is perpendicular to \mathbf{u}_1 . It follows that each $A\mathbf{w}_j$ is a linear combination just of $\mathbf{w}_2, \mathbf{w}_3, \dots, \mathbf{w}_n$, i.e.,

$$A\mathbf{w}_j = \begin{bmatrix} \mathbf{u}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_n \end{bmatrix} \begin{bmatrix} 0 \\ * \\ \vdots \\ * \end{bmatrix}$$

where '*' denotes some unspecified entry. Putting this all together, we see that

$$AP_1 = \underbrace{P_1 \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & A' & \\ 0 & & & \end{bmatrix}}_{A_1}$$

where A' is an $(n - 1) \times (n - 1)$ matrix. P_1 is orthogonal (since its columns form an orthonormal basis) so

$$P_1^t AP_1 = A_1,$$

and it is not hard to derive from this the fact that A_1 is symmetric. Because of the structure of A_1 , this implies that A' is symmetric. Hence, by induction we may

assume there is an $(n-1) \times (n-1)$ orthogonal matrix P' such that $A'P' = P'D'$ with D' diagonal. It follows that

$$\begin{aligned}
 A_1 \underbrace{\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & P' & \\ 0 & & & \end{bmatrix}}_{P_2} &= \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & A' & \\ 0 & & & \end{bmatrix} \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & P' & \\ 0 & & & \end{bmatrix} \\
 &= \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & A'P' & \\ 0 & & & \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & P'D' & \\ 0 & & & \end{bmatrix} \\
 &= \underbrace{\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & P' & \\ 0 & & & \end{bmatrix}}_{P_2} \underbrace{\begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & D' & \\ 0 & & & \end{bmatrix}}_D = P_2 D.
 \end{aligned}$$

Note that P_2 is orthogonal and D is diagonal. Thus,

$$\begin{aligned}
 A \underbrace{P_1 P_2}_P &= P_1 A_1 P_2 = \underbrace{P_1 P_2}_P D \\
 \text{or } AP &= PD.
 \end{aligned}$$

However, a product of orthogonal matrices is orthogonal—see the Exercises—so P is orthogonal as required. \square \square

Exercises for 12.6.

1. An inclined plane makes an angle of 30 degrees with the horizontal. Change to a coordinate system with x'_1 axis parallel to the inclined plane and x'_2 perpendicular to it. Use the change of variables formula derived in the section to find the components of the gravitational acceleration vector $-g\mathbf{j}$ in the new coordinate system. Compare this with what you would get by direct geometric reasoning.
2. Show that the product of two orthogonal matrices is orthogonal. Show that the product of two unitary matrices is unitary. How about the inverse of an orthogonal or unitary matrix?
3. Let $A = \begin{bmatrix} 1 & -i \\ i & 1 \end{bmatrix}$. Find an orthonormal basis for \mathbf{C}^2 consisting of eigenvectors for A . Use this to find a unitary matrix P such that $P^{-1}AP$ is diagonal. (The diagonal entries should be the eigenvalues.)

4. Let $A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$. Find a 2×2 orthogonal matrix P such that $P^t A P$ is diagonal. What are the diagonal entries?

5. Let $A = \begin{bmatrix} 1 & 4 & 3 \\ 4 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}$. Find a 3×3 orthogonal matrix P such that $P^t A P$ is diagonal. What are the diagonal entries?

12.7 Classification of Conics and Quadrics

As mentioned earlier, the Principal Axis Theorem derives its name from its relation to classifying conics, quadric surfaces, and their higher dimensional analogues.

A level curve in \mathbf{R}^2 defined by an equation of the form

$$f(\mathbf{x}) = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 = C$$

is called a *central conic*. (The reason for the 2 will be clear shortly.) As we shall see, a central conic is either an ellipse or a hyperbola (for which the principal axes need not be the coordinate axes) or a degenerate ‘conic’ consisting of a pair of lines.

The most general conic is the locus of an arbitrary quadratic equation which may have linear as well as quadratic terms. Such curves may be studied by applying the methods discussed in this section to the quadratic terms and then completing squares to eliminate linear terms. Parabolas are included in the theory in this way.

To study a central conic, it is convenient to express the function f as follows.

$$\begin{aligned} f(\mathbf{x}) &= (x_1a_{1,1} + x_2a_{21})x_1 + (x_1a_{12} + x_2a_{22})x_2 \\ &= x_1(a_{11}x_1 + a_{12}x_2) + x_2(a_{21}x_1 + a_{22}x_2), \end{aligned}$$

where we have introduced $a_{21} = a_{12}$. The above expression may also be written in matrix form

$$f(\mathbf{x}) = \sum_{j,k=1}^2 x_j a_{jk} x_k = \mathbf{x}^t A \mathbf{x}$$

where A is the symmetric matrix of coefficients.

This may be generalized to $n > 2$ in a rather obvious manner. Let A be a real symmetric $n \times n$ matrix, and define

$$f(\mathbf{x}) = \sum_{j,k=1}^n x_j a_{jk} x_k = \mathbf{x}^t A \mathbf{x}.$$

For $n = 3$ this may be written explicitly

$$\begin{aligned} f(\mathbf{x}) &= (x_1 a_{11} + x_2 a_{21} + x_3 a_{31})x_1 \\ &\quad + (x_1 a_{12} + x_2 a_{22} + x_3 a_{32})x_2 \\ &\quad + (x_1 a_{13} + x_2 a_{23} + x_3 a_{33})x_3 \\ &= a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2 \\ &\quad + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + 2a_{23}x_2x_3. \end{aligned}$$

The level set defined by

$$f(\mathbf{x}) = C$$

is called a *central hyperquadric*. It should be visualized as an $n - 1$ dimensional curved object in \mathbf{R}^n . For $n = 3$ it will be an ellipsoid or a hyperboloid (of one or two sheets) or perhaps a degenerate ‘quadric’ like a cone. (As in the case of conics, we must also allow linear terms to encompass paraboloids.)

If the above contentions are true, we expect the locus of the equation $f(\mathbf{x}) = C$ to have certain axes of symmetry which we shall call its *principal axes*. It turns out that these axes are determined by an *orthonormal basis of eigenvectors* for the coefficient matrix A . To see this, suppose $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ is such a basis and $P = [\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_n]$ is the corresponding orthogonal matrix. By the Principal Axis Theorem, $P^t A P = D$ is diagonal with the eigenvalues, $\lambda_1, \lambda_2, \dots, \lambda_n$, of A appearing on the diagonal. Make the change of coordinates $\mathbf{x} = P\mathbf{x}'$ where \mathbf{x} represents the ‘old’ coordinates and \mathbf{x}' represents the ‘new’ coordinates. Then

$$f(\mathbf{x}) = \mathbf{x}^t A \mathbf{x} = (P\mathbf{x}')^t A (P\mathbf{x}') = (\mathbf{x}')^t P^t A P \mathbf{x}' = (\mathbf{x}')^t D \mathbf{x}'.$$

Since D is diagonal, the quadratic expression on the right has no cross terms, i.e.

$$\begin{aligned} (\mathbf{x}')^t D \mathbf{x}' &= \begin{bmatrix} x'_1 & x'_2 & \cdots & x'_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix} \\ &= \lambda_1(x'_1)^2 + \lambda_2(x'_2)^2 + \cdots + \lambda_n(x'_n)^2. \end{aligned}$$

In the new coordinates, the equation takes the form

$$\lambda_1(x'_1)^2 + \lambda_2(x'_2)^2 + \cdots + \lambda_n(x'_n)^2 = C$$

and its graph is usually quite easy to describe.

Example 237 We shall determine the level curve $f(x, y) = x^2 + 4xy + y^2 = 1$. First rewrite the equation

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 1.$$

Next, find the eigenvalues of the coefficient matrix by solving

$$\det \begin{bmatrix} 1-\lambda & 2 \\ 2 & 1-\lambda \end{bmatrix} = (1-\lambda)^2 - 4 = \lambda^2 - 2\lambda - 3 = 0.$$

This equation is easy to factor, and the roots are $\lambda = 3, \lambda = -1$.

For $\lambda = 3$, to find the eigenvectors, we need to solve

$$\begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0.$$

Reduction of the coefficient matrix yields

$$\begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}$$

with the general solution $v_1 = v_2, v_2$ free. A basic *normalized* eigenvector is

$$\mathbf{u}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

For $\lambda = -1$, a similar calculation (which you should make) yields the basic normalized eigenvector

$$\mathbf{u}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

(Note that $\mathbf{u}_1 \perp \mathbf{u}_2$ as expected.)

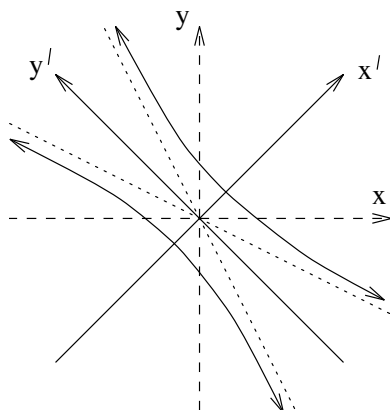
From this we can form the corresponding orthogonal matrix P and make the change of coordinates

$$\begin{bmatrix} x \\ y \end{bmatrix} = P \begin{bmatrix} x' \\ y' \end{bmatrix},$$

and, according to the above analysis, the equation of the level curve in the new coordinate system is

$$3(x')^2 - (y')^2 = 1.$$

It is clear that this is a hyperbola with principal axes pointing along the new axes.



Example 238 Consider the quadric surface defined by

$$x_1^2 + x_2^2 + x_3^2 - 2x_1x_3 = 1.$$

We take

$$f(\mathbf{x}) = x_1^2 + x_2^2 + x_3^2 - 2x_1x_3 = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

The characteristic equation of the coefficient matrix is

$$\det \begin{bmatrix} 1-\lambda & 0 & -1 \\ 0 & 1-\lambda & 0 \\ -1 & 0 & 1-\lambda \end{bmatrix} = (1-\lambda)^3 - (1-\lambda) = -(\lambda-2)(\lambda-1)\lambda = 0$$

Thus, the eigenvalues are $\lambda = 2, 1, 0$.

For $\lambda = 2$, reduce

$$\begin{bmatrix} -1 & 0 & -1 \\ 0 & -1 & 0 \\ -1 & 0 & -1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

to obtain $v_1 = -v_3, v_2 = 0$ with v_3 free. Thus,

$$\mathbf{v}_1 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

is a basic eigenvector for $\lambda = 2$, and

$$\mathbf{u}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

is a basic unit eigenvector.

Similarly, for $\lambda = 1$ reduce

$$\begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

which yields $v_1 = v_3 = 0$ with v_2 free. Thus a basic unit eigenvector for $\lambda = 1$ is

$$\mathbf{u}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

Finally, for $\lambda = 0$, reduce

$$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

This yields $v_1 = x_3, v_2 = 0$ with v_3 free. Thus, a basic unit eigenvector for $\lambda = 0$ is

$$\mathbf{u}_3 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

The corresponding orthogonal change of basis matrix is

$$P = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \mathbf{u}_3] = \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Moreover, putting $\mathbf{x} = P\mathbf{x}'$, we can express the equation of the quadric surface in the new coordinate system

$$2x_1'^2 + 1x_2'^2 + 0x_3'^2 = 2x_1'^2 + x_2'^2 = 1. \quad (222)$$

Thus it is easy to see what the level surface is: an elliptical cylinder perpendicular to the x_1', x_2' plane. The three ‘principal axes’ in this case are the two axes of the ellipse in the x_1', x_2' plane and the x_3' axis, which is the central axis of the cylinder.

Representing the graph in the new coordinates makes it easy to understand its geometry. Suppose, for example, that we want to find the points on the graph which are closest to the origin. These are the points at which the x_1' -axis intersects the surface. These are the points with new coordinates $x_1' = \pm \frac{1}{\sqrt{2}}, x_2' = x_3' = 0$. If you want the coordinates of these points in the original coordinate system, use the change of coordinates formula

$$\mathbf{x} = P\mathbf{x}'.$$

Thus, the old coordinates of the minimum point with new coordinates $(1/\sqrt{2}, 0, 0)$ are given by

$$\begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} \\ 0 \\ \frac{1}{2} \end{bmatrix}.$$

Exercises for 12.7.

1. Find the principal axes and classify the central conic $x^2 + xy + y^2 = 1$.
2. Identify the conic defined by $x^2 + 4xy + y^2 = 4$. Find its principal axes, and find the points closest and furthest (if any) from the origin.
3. Identify the conic defined by $2x^2 + 72xy + 23y^2 = 50$. Find its principal axes, and find the points closest and furthest (if any) from the origin.

4. Find the principal axes and classify the central quadric defined by

$$x^2 - y^2 + z^2 - 4xy - 4yz = 1.$$

5. (Optional) Classify the surface defined by

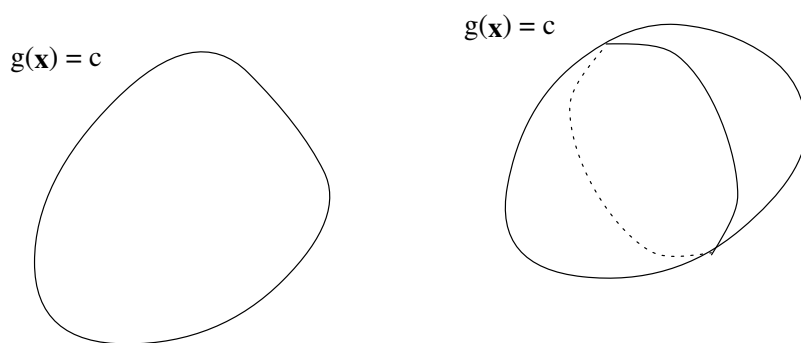
$$x^2 + 2y^2 + z^2 + 2xy + 2yz - z = 0.$$

Hint: This is not a central quadric. To classify it, first apply the methods of the section to the quadratic expression $x^2 + 2y^2 + z^2 + 2xy + 2yz$ to find a new coordinate system in which this expression has the form $\lambda_1 x'^2 + \lambda_2 y'^2 + \lambda_3 z'^2$. Use the change of coordinates formula to express z in terms of x' , y' , and z' and then complete squares to eliminate all linear terms. At this point, it should be clear what the surface is.

12.8 A Digression on Constrained Maxima and Minima

There is another approach to finding the principal axes of a conic, quadric, or hyperquadric. Consider for an example an ellipse in \mathbf{R}^2 centered at the origin. One of the principal axes intersects the conic in the two points at greatest distance from the origin, and the other intersects it in the two points at least distance from the origin. Similarly, two of the three principal axes of a central ellipsoid in \mathbf{R}^3 may be obtained in this way. Thus, if we didn't know about eigenvalues and eigenvectors, we might try to find the principal axes by maximizing (or minimizing) the function giving the distance to the origin *subject to* the quadratic equation defining the conic or quadric. In such a problem, we need to minimize a function assuming there are one or more relations or *constraints* among the variables. In this section we shall consider problems of this kind in general.

We start by considering the case of a single constraint. Suppose we want to maximize (minimize) the real valued function $f(\mathbf{x}) = f(x_1, x_2, \dots, x_n)$ subject to the constraint $g(\mathbf{x}) = g(x_1, x_2, \dots, x_n) = c$. For $n = 2$, this has a simple geometric interpretation. The locus of the equation $g(x_1, x_2) = c$ is a level curve of the function g , and we want to maximize (minimize) the function f *on that curve*. Similarly, for $n = 3$, the level set $g(x_1, x_2, x_3) = c$ is a surface in \mathbf{R}^3 , and we want to maximize (minimize) f *on that surface*. In \mathbf{R}^n , we call the level set defined by $g(x_1, x_2, \dots, x_n) = c$ a *hypersurface*, and the problem is to maximize (minimize) the function f *on that hypersurface*.



n = 2. Level curve in the plane.

n = 3. Level surface in space.

Examples Maximize $f(x, y) = x + 3y$ on the hyperbola $g(x, y) = x^2 - y^2 = 1$.

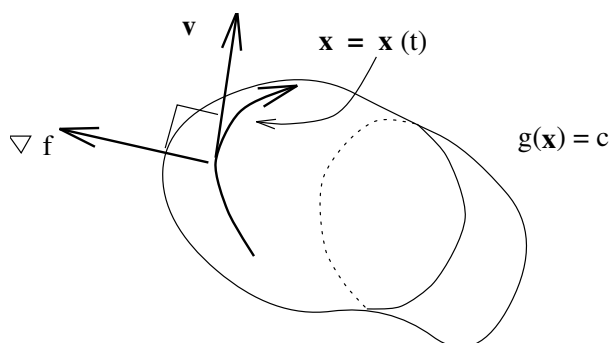
Maximize $f(x, y) = x^2 + y^2$ on the ellipse $g(x, y) = x^2 + 4y^2 = 3$. (This is easy if you draw the picture.)

Minimize $f(x, y, z) = 2x^2 + 3xy + y^2 + xz - 4z^2$ on the sphere $g(x, y, z) = x^2 + y^2 + z^2 = 1$.

Minimize $f(x, y, z, t) = x^2 + y^2 + z^2 - t^2$ on the hypersphere $g(x, y, z, t) = x^2 + y^2 + z^2 + t^2 = 1$.

We shall concentrate on the case of $n = 3$ variables, but the reasoning for any n is similar. We want to maximize (or minimize) $f(\mathbf{x})$ on a level set $g(\mathbf{x}) = c$ in \mathbf{R}^3 , where as usual we abbreviate $\mathbf{x} = (x_1, x_2, x_3)$. Assume that both f and g are smooth functions defined on open sets in \mathbf{R}^3 , and that the level set $g(\mathbf{x}) = c$ has a well defined tangent plane at a potential maximum point. The latter assumption means that the normal vector ∇g does not vanish at the point. It follows from this assumption that every vector \mathbf{v} perpendicular to ∇g at the point is a tangent vector for some curve in the level set passing through the point. (Refer back to the discussion of tangent planes and the implicit function theorem in Chapter III, Section 8.)

Suppose such a curve is given by the parametric representation $\mathbf{x} = \mathbf{x}(t)$.



By the chain rule we have

$$\frac{df}{dt} = \nabla f \cdot \frac{d\mathbf{x}}{dt} = \nabla f \cdot \mathbf{v}$$

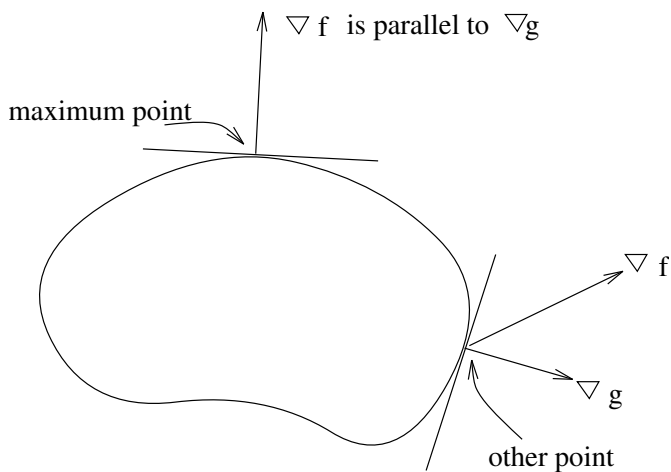
where $\mathbf{v} = d\mathbf{x}/dt$. If the function attains a maximum on the level set at the given point, it also attains a maximum along this curve, so we conclude that

$$\frac{df}{dt} = \nabla f \cdot \mathbf{v} = 0.$$

As above, we can arrange that the vector \mathbf{v} is any possible vector in the tangent plane at the point. Since there is a unique direction perpendicular to the tangent plane, that of ∇g , we conclude that ∇f is parallel to ∇g , i.e.,

$$\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x}) \quad (223)$$

for some scalar λ .



(223) is a *necessary condition* which must hold at any maximum point where f and g are smooth and $\nabla g \neq 0$. (It doesn't by itself guarantee that there is a maximum at the point. There could be a minimum or even no extreme value at all at the point.) Taking components, we obtain 3 scalar equations for the 4 variables x_1, x_2, x_3, λ . We would not expect, even in the best of circumstances to get a unique solution from this, but the defining equation for the level surface

$$g(\mathbf{x}) = c$$

provides a 4th equation. We still won't generally get a unique solution, but we will usually get at most a finite number of possible solutions. Each of these can be examined further to see if f attains a maximum (or minimum) at that point in the level set. Notice that the variable λ plays an auxiliary role since we really only want the coordinates of the point \mathbf{x} . (In some applications, λ has some significance beyond that.) This method is due to the 19th century French mathematician Lagrange and λ is called a *Lagrange multiplier*.

Example 239 Suppose we want to maximize the function $f(x, y, z) = x + y - z$ on the sphere $x^2 + y^2 + z^2 = 1$. We take $g(x, y, z) = x^2 + y^2 + z^2$. Then, $\nabla f = \langle 1, 1, -1 \rangle$ and $\nabla g = \langle 2x, 2y, 2z \rangle$, so the relation $\nabla f = \lambda \nabla g$ yields

$$\begin{aligned} 1 &= \lambda(2x) \\ 1 &= \lambda(2y) \\ -1 &= \lambda(2z) \end{aligned}$$

to which we add the equation

$$x^2 + y^2 + z^2 = 1.$$

From the first three equations, we obtain

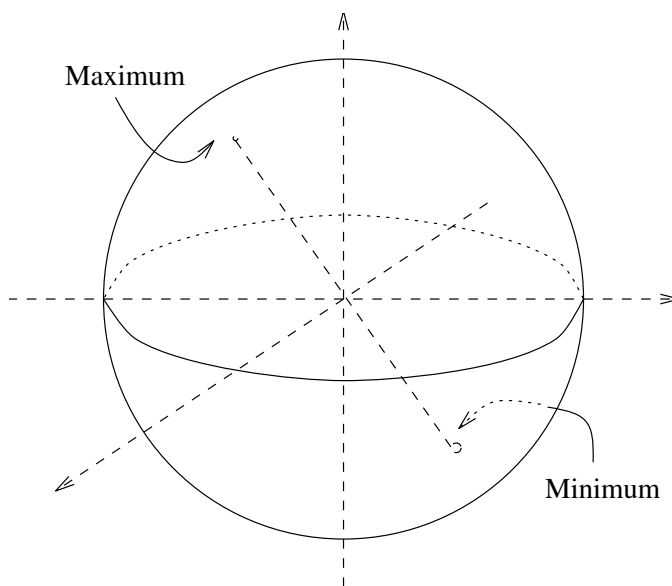
$$\begin{aligned} x &= \frac{1}{2\lambda} & y &= \frac{1}{2\lambda} & z &= -\frac{1}{2\lambda} \\ \frac{1}{4\lambda^2} + \frac{1}{4\lambda^2} + \frac{1}{4\lambda^2} &= 1 \\ \frac{3}{4} &= \lambda^2 \\ \lambda &= \pm \frac{\sqrt{3}}{2}. \end{aligned}$$

Thus we have two possible solutions. For $\lambda = \sqrt{3}/2$, we obtain the point

$$(1/\sqrt{3}, 1/\sqrt{3}, -1/\sqrt{3}) \quad \text{at which } f = x + y - z = \sqrt{3}.$$

For $\lambda = -\sqrt{3}/2$, we obtain the point

$$(-1/\sqrt{3}, -1/\sqrt{3}, 1/\sqrt{3}) \quad \text{at which } f = -\sqrt{3}.$$



Since the level set $x^2 + y^2 + z^2 = 1$ is a closed bounded set, and since the function f is continuous, both maximum and minimum values must be attained somewhere on the level set. The only two candidates we have come up with are the two points given above, so it is clear the first is a maximum point and the second is a minimum point.

The method of Lagrange multipliers often leads to a set of equations which is difficult to solve. Sometimes a great deal of ingenuity is required, so you should treat each problem as unique and expect to have to be creative about solving it.

Example 240 Suppose we want to minimize the function $f(x, y) = x^2 + 4xy + y^2$ on the circle $x^2 + y^2 = 1$. For this problem $n = 2$, and the level set is a curve. Take $g(x, y) = x^2 + y^2$. Then $\nabla f = \langle 2x + 4y, 4x + 2y \rangle$, $\nabla g = \langle 2x, 2y \rangle$, and $\nabla f = \lambda \nabla g$ yields the equations

$$2x + 4y = \lambda(2x)$$

$$4x + 2y = \lambda(2y)$$

to which we add

$$x^2 + y^2 = 1.$$

After canceling a common factor of 2, the first two equations may be written in matrix form

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix}$$

which says that

$$\begin{bmatrix} x \\ y \end{bmatrix}$$

is an *eigenvector* for the eigenvalue λ , and the equation $x^2 + y^2 = 1$ says it is a *unit eigenvector*. You should know how to solve such problems, and we leave it to you to make the required calculations. (See also Example 237 in the previous section where we made these calculations in another context.) The eigenvalues are $\lambda = 3$ and $\lambda = -1$. For $\lambda = 3$, a basic unit eigenvector is

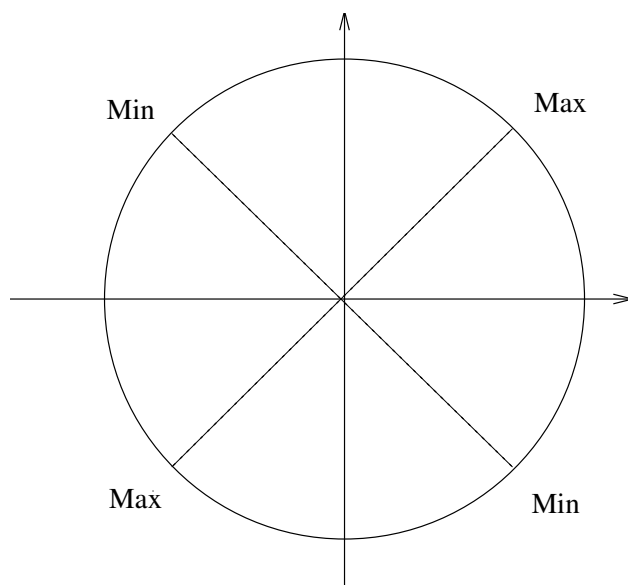
$$\mathbf{u}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

and every other eigenvector is of the form $c\mathbf{u}_1$. The latter will be a *unit* vector if and only $|c| = 1$, i.e., $c = \pm 1$. We conclude that $\lambda = 3$ yields two solutions of the Lagrange multiplier problem: $(1/\sqrt{2}, 1/\sqrt{2})$ and $(-1/\sqrt{2}, -1/\sqrt{2})$. At each of these points $f(x, y) = x^2 + 4xy + y^2 = 3$.

For $\lambda = -1$, we obtain the basic unit eigenvector

$$\mathbf{u}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix},$$

and a similar analysis (which you should do) yields the two points: $(1/\sqrt{2}, -1/\sqrt{2})$ and $(-1/\sqrt{2}, 1/\sqrt{2})$. At each of these points $f(x, y) = x^2 + 4xy + y^2 = -1$.



Hence, the function attains its maximum value at the first two points and its minimum value at the second two.

Example 241 Suppose we want to minimize the function $g(x, y) = x^2 + y^2$ (which is the square of the distance to the origin) on the conic $f(x, y) = x^2 + 4xy + y^2 = 1$. Note that this is basically the same as the previous example except that the roles of the two functions are reversed. The Lagrange multiplier condition $\nabla g = \lambda \nabla f$ is the same as the condition $\nabla f = (1/\lambda) \nabla g$ provided $\lambda \neq 0$. ($\lambda \neq 0$ in this case since otherwise $\nabla g = 0$, which yields $x = y = 0$. However, $(0, 0)$ is not a point on the conic.) We just solved that problem and found eigenvalues $1/\lambda = 3$ or $1/\lambda = -1$. In this case, we don't need unit eigenvectors, so to avoid square roots we choose basic eigenvectors

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

corresponding respectively to $\lambda = 3$ and $\lambda = -1$. The endpoint of \mathbf{v}_1 does not lie on the conic, but any other eigenvector for $\lambda = 3$ is of the form $c\mathbf{v}_1$, so all we need to do is adjust c so that the point satisfies the equation $f(x, y) = x^2 + 4xy + y^2 = 1$. Substituting $(x, y) = (c, c)$ yields $6c^2 = 1$ or $c = \pm 1/\sqrt{6}$. Thus, we obtain the two points $(1/\sqrt{6}, 1/\sqrt{6})$ and $(-1/\sqrt{6}, -1/\sqrt{6})$. For $\lambda = -1$, substituting $(x, y) = (-c, c)$ in the equation yields $-2c^2 = 1$ which has no solutions.

Thus, the only candidates for a minimum (or maximum) are the first pair of points: $(1/\sqrt{6}, 1/\sqrt{6})$ and $(-1/\sqrt{6}, -1/\sqrt{6})$. A simple calculation shows these are both $1/\sqrt{3}$ units from the origin, but without further analysis, we can't tell if this is the maximum, the minimum, or neither. However, it is not hard to classify this conic—see the previous section—and discover that it is a hyperbola. Hence, the two points are minimum points.

The Rayleigh-Ritz Method Example 240 above is typical of a certain class of Lagrange multiplier problems. Let A be a real symmetric $n \times n$ matrix, and consider the problem of maximizing (minimizing) quadratic function $f(\mathbf{x}) = \mathbf{x}^t A \mathbf{x}$ subject to the constraint $g(\mathbf{x}) = |\mathbf{x}|^2 = 1$. This is called the *Rayleigh-Ritz problem*. For $n = 2$ or $n = 3$, the level set $|\mathbf{x}|^2 = 1$ is a circle or sphere, and for $n > 3$, it is called a *hypersphere*.

Alternatively, we could reverse the roles of the functions f and g , i.e., we could try to maximize (minimize) the square of the distance to the origin $g(\mathbf{x}) = |\mathbf{x}|^2$ on the level set $f(\mathbf{x}) = 1$. Because the Lagrange multiplier condition in either case asserts that the two gradients ∇f and ∇g are parallel, these two problems are very closely related. The latter problem—finding the points on a conic, quadric, or hyperquadric furthest from (closest to) the origin—is easier to visualize, but the former problem—maximizing or minimizing the quadratic function f on the hypersphere $|\mathbf{x}| = 1$ —is easier to compute with.

Let's go about applying the Lagrange Multiplier method to the Rayleigh-Ritz problem. The components of ∇g are easy:

$$\frac{\partial g}{\partial x_i} = 2x_i, \quad i = 1, 2, \dots, n.$$

The calculation of ∇f is harder. First write

$$f(\mathbf{x}) = \sum_{j=1}^n x_j \left(\sum_{k=1}^n a_{jk} x_k \right)$$

and then carefully apply the product rule together with $a_{jk} = a_{kj}$. The result is

$$\frac{\partial f}{\partial x_i} = 2 \sum_{j=1}^n a_{ij} x_j \quad i = 1, 2, \dots, n.$$

(Work this out explicitly in the cases $n = 2$ and $n = 3$ if you don't believe it.) Thus, the Lagrange multiplier condition $\nabla f = \lambda \nabla g$ yields the equations

$$2 \sum_{j=1}^n a_{ij} x_j = \lambda (2x_i) \quad i = 1, 2, \dots, n$$

which may be rewritten in matrix form (after canceling the 2's)

$$A\mathbf{x} = \lambda\mathbf{x}. \quad (225)$$

To this we must add the equation of the level set

$$g(\mathbf{x}) = |\mathbf{x}|^2 = 1.$$

Thus, any potential solution \mathbf{x} is a *unit* eigenvector for the matrix A with eigenvalue λ . Note also that for such a unit eigenvector, we have

$$f(\mathbf{x}) = \mathbf{x}^t A \mathbf{x} = \mathbf{x}^t (\lambda \mathbf{x}) = \lambda \mathbf{x}^t \mathbf{x} = \lambda |\mathbf{x}|^2 = \lambda.$$

Thus the eigenvalue is the extreme value of the quadratic function at the point on the (hyper)sphere given by the unit eigenvector.

The upshot of this discussion is that for a real symmetric matrix A , the Rayleigh–Ritz problem is equivalent to the problem of finding an orthonormal basis of eigenvectors for A .

The Rayleigh–Ritz method may be used to show that a real symmetric matrix has real eigenvalues without invoking the use of complex vectors as we did previously in Section 4. (See Theorem 12.1.) Here is an outline of the argument. The hypersphere $g(\mathbf{x}) = |\mathbf{x}|^2 = 1$ is a closed bounded set in \mathbf{R}^n for any n . It follows from a basic theorem in analysis that any continuous function, in particular the quadratic function $f(\mathbf{x})$, must attain both maximum and minimum values on the hypersphere. Hence, the Lagrange multiplier problem always has solutions, which by the above algebra amounts to the assertion that the real symmetric matrix A must have at least one eigenvalue. This suggests a general procedure for showing that all the eigenvalues are real. First find the largest eigenvalue by maximizing the quadratic function $f(\mathbf{x})$ on the set $|\mathbf{x}|^2 = 1$. Let $\mathbf{x} = \mathbf{u}_1$ be the corresponding

eigenvector. Change coordinates by choosing an orthonormal basis starting with \mathbf{u}_1 . Then the additional basis elements will span the subspace perpendicular to \mathbf{u}_1 and we may obtain a lower dimensional quadratic function by restricting f to that subspace. We can now repeat the process to find the next smaller real eigenvalue. Continuing in this way, we will obtain an orthonormal basis of eigenvectors for A and each of the corresponding eigenvalues will be real.

The Rayleigh–Ritz Method generalizes nicely for complex Hermitian matrices and also for infinite dimensional analogues. In quantum mechanics, for example, one considers complex valued functions $\psi(x, y, z)$ defined on \mathbf{R}^3 satisfying the condition

$$\iiint_{\mathbf{R}^3} |\psi(x, y, z)|^2 dV < \infty.$$

Such functions are called *wave functions*, and the set of all such functions form a complex vector space. Certain operators A on this vector space represent observable quantities, and the eigenvalues of these operators represent the possible results of measurements of these observables. Since the vector space is infinite dimensional, one can't represent these operators by finite matrices, so the usual method of determining eigenvalues and eigenvectors breaks down. However, one can generalize many of the ideas we have developed here. For example, one may define the inner product of two wave functions by the formula

$$\langle \psi | \phi \rangle = \iiint_{\mathbf{R}^3} \overline{\psi(x, y, z)} \phi(x, y, z) dV.$$

Then, one may determine the eigenvalues of a Hermitian operator A by the studying the optimization problem for the quantity

$$\langle \psi | A\psi \rangle$$

subject to the condition $\langle \psi | \psi \rangle = 1$.

Lagrange Multipliers with More Than One Constraint In \mathbf{R}^n , suppose we want to maximize (minimize) a function $f(\mathbf{x})$ subject to m constraints

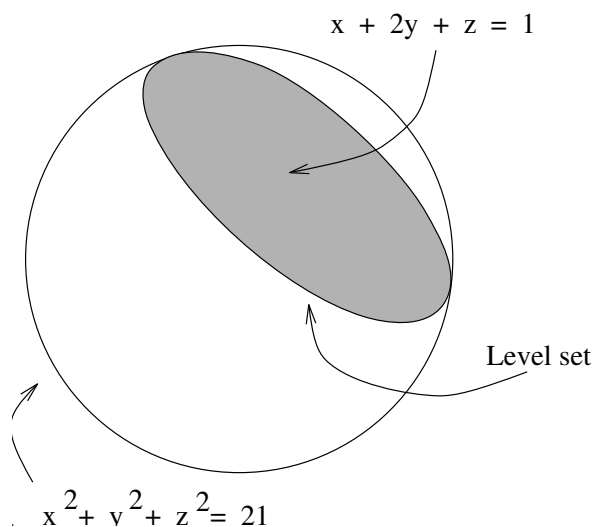
$$g_1(\mathbf{x}) = c_1, \quad g_2(\mathbf{x}) = c_2, \dots, g_r(\mathbf{x}) = c_m.$$

We can make the scalar functions $g_i(\mathbf{x})$ the components of a *vector* function $\mathbf{g} : \mathbf{R}^n \rightarrow \mathbf{R}^m$. Then the m constraining equations may be summarized by a single vector constraint

$$\mathbf{g}(\mathbf{x}) = \mathbf{c} = \langle c_1, c_2, \dots, c_m \rangle.$$

In this way, we may view the constraint as defining a level set (for \mathbf{g}) in \mathbf{R}^n , and the problem is to maximize f on this level set. The level set may also be viewed as the *intersection* of the m hypersurfaces in \mathbf{R}^n which are level sets of the component scalar functions $g_i(\mathbf{x}) = c_i$.

Example 242 Consider the problem of finding the highest point on the curve of intersection of the plane $x + 2y + z = 1$ with the sphere $x^2 + y^2 + z^2 = 21$.



Here we take $f(x, y, z) = z$ and

$$\mathbf{g}(x, y, z) = \begin{bmatrix} g_1(x, y, z) \\ g_2(x, y, z) \end{bmatrix} = \begin{bmatrix} x + 2y + z \\ x^2 + y^2 + z^2 \end{bmatrix}$$

$$\mathbf{c} = \begin{bmatrix} 1 \\ 21 \end{bmatrix}.$$

If we assume that f and \mathbf{g} are smooth, then just as before we obtain

$$\frac{df}{dt} = \nabla f \cdot \mathbf{v} = 0$$

for every vector \mathbf{v} tangent to a curve in the level set through the maximum point. Every such \mathbf{v} , since it is tangent to the level set, will be perpendicular to each of the normal vectors ∇g_i at the point, i.e.,

$$\nabla g_1 \cdot \mathbf{v} = 0$$

$$\nabla g_2 \cdot \mathbf{v} = 0$$

$$\vdots$$

$$\nabla g_m \cdot \mathbf{v} = 0.$$

If we make the gradient vectors into the rows of a matrix, this system may be rewritten

$$\begin{bmatrix} \nabla g_1 \\ \nabla g_2 \\ \vdots \\ \nabla g_m \end{bmatrix} \mathbf{v} = 0. \quad (226)$$

In the case of one constraint, we assumed that $\nabla g \neq 0$ so there would be a well defined tangent plane at the maximum point. Now, we need a more stringent condition: the gradients at the potential maximum point

$$\nabla g_1, \nabla g_2, \dots, \nabla g_m$$

should form a *linearly independent set*. This means that the $m \times n$ system (226) has rank m . Hence, the solution space of all vectors \mathbf{v} satisfying (226) is $n - m$ -dimensional. This solution space is called the *tangent space* to the level set at the point. In these circumstances, it is possible to show (from higher dimensional analogues of the implicit function theorem) that every vector \mathbf{v} in this tangent space is in fact tangent to a curve lying in the level set. Using this, we may conclude that, at a maximum point,

$$\nabla f \cdot \mathbf{v} = 0$$

for every vector \mathbf{v} in the tangent space. Consider then the $(m + 1) \times n$ system

$$\begin{bmatrix} \nabla g_1 \\ \nabla g_2 \\ \vdots \\ \nabla g_m \\ \nabla f \end{bmatrix} \mathbf{v} = 0.$$

This cannot have rank $m + 1$ since it has exactly the same solution space as the system (226). Hence, it has rank m , and the only way that could happen is if the last row is dependent on the m previous rows, i.e.,

$$\nabla f = \lambda_1 \nabla g_1 + \lambda_2 \nabla g_2 + \dots + \lambda_m \nabla g_m. \quad (227)$$

The scalars $\lambda_1, \lambda_2, \dots, \lambda_m$ are called Lagrange multipliers.

Example 242, continued We have $\nabla f = [0 \ 0 \ 1]$ and

$$\begin{bmatrix} \nabla g_1 \\ \nabla g_2 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 \\ 2x & 2y & 2z \end{bmatrix}.$$

Hence, the Lagrange multiplier condition $\nabla f = \lambda_1 \nabla g_1 + \lambda_2 \nabla g_2$ amounts to

$$[0 \ 0 \ 1] = \lambda_1 [1 \ 2 \ 1] + \lambda_2 [2x \ 2y \ 2z]$$

which yields

$$\begin{aligned} \lambda_1 + 2\lambda_2 x &= 0 \\ 2\lambda_1 + 2\lambda_2 y &= 0 \\ \lambda_1 + 2\lambda_2 z &= 1. \end{aligned}$$

To this we must add the constraints

$$\begin{aligned} x + 2y + z &= 1 \\ x^2 + y^2 + z^2 &= 21. \end{aligned} \quad (228)$$

In total, this gives 5 equations for the 5 unknowns $x, y, z, \lambda_1, \lambda_2$. We can solve these equations by being sufficiently ingenious, but there is a short cut. The multiplier condition just amounts to the assertion that $\{\nabla g_1, \nabla g_2, \nabla f\}$ is a dependent set. (But, it is assumed that $\{\nabla g_1, \nabla g_2\}$ is independent.) However, this set is dependent if and only if

$$\det \begin{bmatrix} \nabla g_1 \\ \nabla g_2 \\ \nabla f \end{bmatrix} = \det \begin{bmatrix} 0 & 0 & 1 \\ 1 & 2 & 1 \\ 2x & 2y & 2z \end{bmatrix} = 0$$

$$\text{i.e.} \quad 1(2(2x) - 2y) = 2(2x - y) = 0$$

$$\text{i.e.} \quad y = 2x.$$

Putting this in (228) yields

$$\begin{aligned} 5x + z &= 1 & 5x^2 + z^2 &= 21 \\ 5x^2 + (1 - 5x)^2 &= 30x^2 - 10x + 1 = 21 \\ 30x^2 - 10x - 20 &= 0 \\ x &= 1, -\frac{2}{3}. \end{aligned}$$

Using $z = 1 - 5x, y = 2x$ yields the following two points as possible maximum points:

$$(1, 2, -4) \quad \text{and} \quad (-2/3, -4/3, 13/3).$$

It is clear that the maximum value of $f(x, y, z) = z$ is attained at the second point.

There is one minor issue that was ignored in the above calculations. The reasoning is only valid at points at which the ‘tangent space’ to the level set is well defined as defined above. In this case the level set is a curve (in fact, it is a circle), and the tangent space is 1 dimensional, i.e., it is a line. The two gradients $\nabla g_1, \nabla g_2$ generally span the plane perpendicular to the tangent line, but it could happen at some point that one of the gradients is a multiple of the other. In that case the two level surfaces $g_1(\mathbf{x}) = c_1$ and $g_2(\mathbf{x}) = c_2$ are tangent to one another at the given point, so we would expect some problems. For example, consider the intersection of the hyperbolic paraboloid $z - x^2 + y^2 = 0$ with its tangent plane at the origin $z = 0$. This ‘curve’ consists two straight lines which intersect at the origin. At any point other than the origin, there is a well defined tangent line, i.e., whichever of the two lines is appropriate, but at the origin there is a problem.

In general, there is no way to know that a maximum or minimum does not occur at a point where the tangent space is not well defined. Hence, all such points must be considered possible candidates for maximum or minimum points. In Example 242, however, it is fairly clear geometrically that there are no such points. This can also be confirmed analytically by seeing that $\{\nabla g_1, \nabla g_2\}$ is independent at every point of the set. For, since $\nabla g_1 \neq 0$, the only way the pair could be dependent is by a

relation of the form $\nabla g_2 = c\nabla g_1$. This yields

$$\begin{aligned} \begin{bmatrix} 2x & 2y & 2z \end{bmatrix} &= c \begin{bmatrix} 1 & 2 & 1 \end{bmatrix} \\ 2x = c, \quad 2y = c2c, \quad 2z = c \\ x = z, y &= x/2 \end{aligned}$$

and it is easy to see these equations are not consistent with $x + 2y + z = 1, x^2 + y^2 + z^2 = 21$.

Exercises for 12.8.

1. Find the maximum value of $f(x, y) = 2x + y$ subject to the constraint $x^2 + y^2 = 4$.
2. Find the minimum value of $f(x, y, z) = x^2 + y^2 + z^2$ given the constraint $x + y + z = 10$.
3. Find the maximum and minimum values of the function $f(x, y) = x^2 + y^2$ given the constraint $x^2 + xy + y^2 = 1$.
4. Find the maximum and/or minimum value of $f(x, y, z) = x^2 - y^2 + z^2 - 4xy - 4yz$ subject to $x^2 + y^2 + z^2 = 1$.
5. The derivation of the Lagrange multiplier condition $\nabla f = \lambda \nabla g$ assumes that the $\nabla g \neq 0$, so there is a well defined tangent 'plane' at the potential maximum or minimum point. However, a maximum or minimum could occur at a point where $\nabla g = 0$, so all such points should also be checked. (Similarly, either f or g might fail to be smooth at a maximum or minimum point.) With these remarks in mind, find where $f(x, y, z) = x^2 + y^2 + z^2$ attains its minimum value subject to the constraint $g(x, y, z) = x^2 + y^2 - z^2 = 0$.
6. Consider as in Example 2 the problem of maximizing $f(x, y) = x^2 + 4xy + y^2$ given the constraint $x^2 + y^2 = 1$. This is equivalent to maximizing $F(x, y) = xy$ on the circle $x^2 + y^2 = 1$. (Why?) Draw a diagram showing the circle and selected level curves $F(x, y) = c$ of the function F . Can you see why $F(x, y)$ attains its maximum at $(1/\sqrt{2}, 1/\sqrt{2})$ and $(-1/\sqrt{2}, -1/\sqrt{2})$ without using any calculus? Hint: consider how the level curves of F intersect the circle and decide from that where F is increasing, and where it is decreasing on the circle.
7. Find the maximum and minimum values (if any) of the function $f(x, y, z) = x^2 + y^2 + z^2$ on the line of intersection of the two planes $x + y + 2z = 2$ and $2x - y + z = 4$.
8. Find the highest point, i.e., maximize $f(x, y, z) = z$, on the ellipse which is the intersection of the cylinder $x^2 + 2y^2 = 2$ with the plane $x + y - z = 10$.

9. For the problem of maximizing (minimizing) $f(x_1, \dots, x_n)$ subject to constraints $g_i(x_1, \dots, x_n) = c_i$, $i = 1, 2, \dots, m$, how many equations in how many unknowns does the method of Lagrange multipliers yield?

Chapter 13

Nonlinear Systems

13.1 Introduction

So far we have concentrated almost entirely on linear differential equations. This is appropriate for several reasons. First, most of classical physics is described by linear equations, so knowing how to solve them is fundamental in applying the laws of physics. Second, even in situations where nonlinear equations are necessary to describe phenomena, it is often possible to begin to understand the solutions by making linear approximations. Finally, linear equations are usually much easier to study than nonlinear equations. Be that as it may, nonlinear equations and nonlinear systems have become increasingly important in applications, and at the same time a lot of progress has been made in understanding their solutions. In this chapter we shall give a very brief introduction to the subject.

We start with a couple of typical examples.

Example 243 In your physics class you studied the behavior of an undamped pendulum. To make our analysis easier, we shall assume that the pendulum consists of a point mass m connected to a fixed pivot by a massless rigid rod of length L .

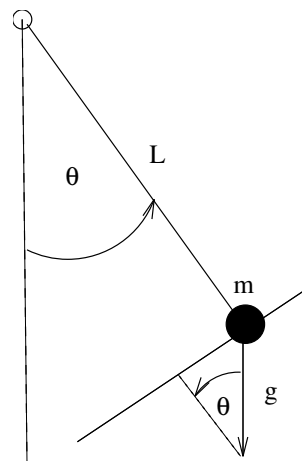
Then, using polar coordinates to describe the motion, we have for the tangential acceleration

$$a_\theta = r \frac{d^2\theta}{dt^2} + 2 \frac{dr}{dt} \frac{d\theta}{dt}.$$

(Refer to Chapter I, Section 2 for acceleration in polar coordinates.) Since $r = L$, $dr/dt = 0$, we obtain

$$a_\theta = L\theta''.$$

On the other hand, the component of acceleration in the tangential direction is



$-g \sin \theta$, so we obtain the second order differential equation

$$\theta'' = -\frac{g}{L} \sin \theta. \quad (229)$$

This is a second order nonlinear equation. It is usually solved as follows. *Assume θ is small.* Then

$$\sin \theta = \theta + O(\theta^3)$$

so (229) may be approximated by the second order linear equation

$$\theta'' = -\frac{g}{L} \theta.$$

It is easy to solve this equation:

$$\theta = A \cos \left(\sqrt{\frac{g}{L}} t + \delta \right).$$

However, this approximation is certainly not valid if θ is large. For example, if you give the mass a big enough shove, it will revolve about the pivot through a complete circuit and in the absence of friction will continue to do that indefinitely. We clearly need another approach if we want to understand what happens in general.

Unfortunately equation (229) can't be solved explicitly in terms of known functions. However, it is possible to get a very good qualitative understanding of its solutions. To explore this, we consider the equivalent first order system. Let $x_1 = \theta$ and $x_2 = \theta'$. Then $x_2' = \theta'' = -(g/L) \sin \theta = -(g/L) \sin x_1$. Hence, the desired system is

$$\begin{aligned} x_1' &= x_2 \\ x_2' &= -\frac{g}{L} \sin x_1 \end{aligned}$$

or, in vector form

$$\mathbf{x}' = \mathbf{f}(\mathbf{x}) = \begin{bmatrix} x_2 \\ -\frac{g}{L} \sin x_1 \end{bmatrix}. \quad (230)$$

As noted earlier, we can't find $\mathbf{x} = \mathbf{x}(t)$ explicitly as a function of t , but it is possible to learn quite a lot about the solutions by looking at the geometry of the solution curves. In this case, we can describe the geometry by eliminating t from the differential equations. We have

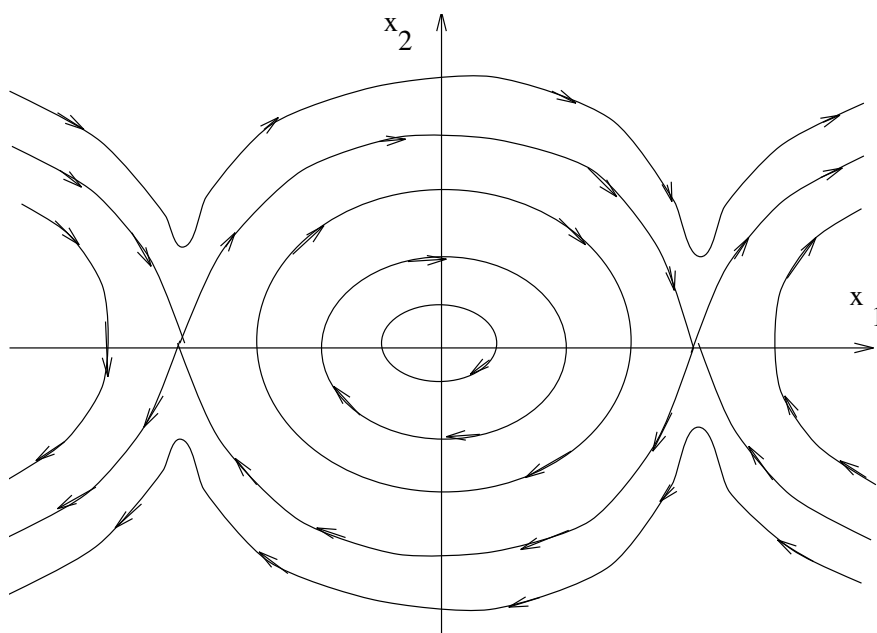
$$\frac{dx_2}{dx_1} = \frac{dx_2/dt}{dx_1/dt} = -\frac{g}{L} \frac{\sin x_1}{x_2}.$$

This equation may be solved by separation of variables. I leave the details to you, but the general solution is

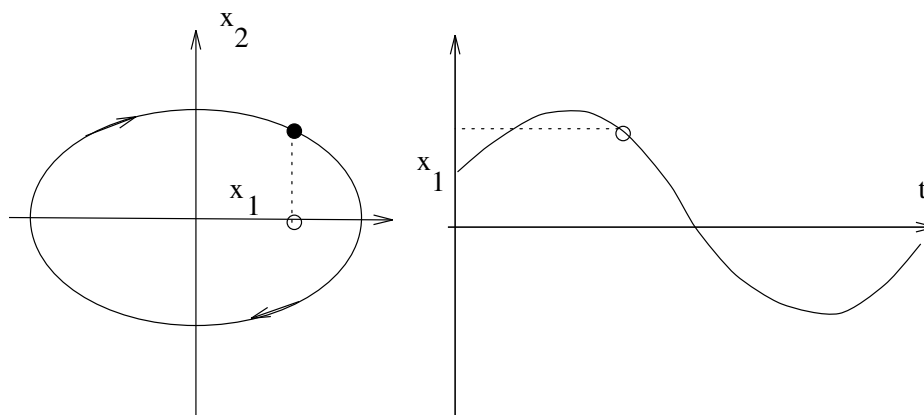
$$Lx_2^2 - 2g \cos x_1 = C. \quad (231)$$

This equation may also be derived quite easily from the law of *conservation of energy*. Namely, multiplying by $mL/2$ and expressing everything in terms of θ yields $\frac{1}{2}m(L\theta')^2 - mgL \cos \theta = C$. The first term on the left represents the kinetic energy and the second the potential energy.

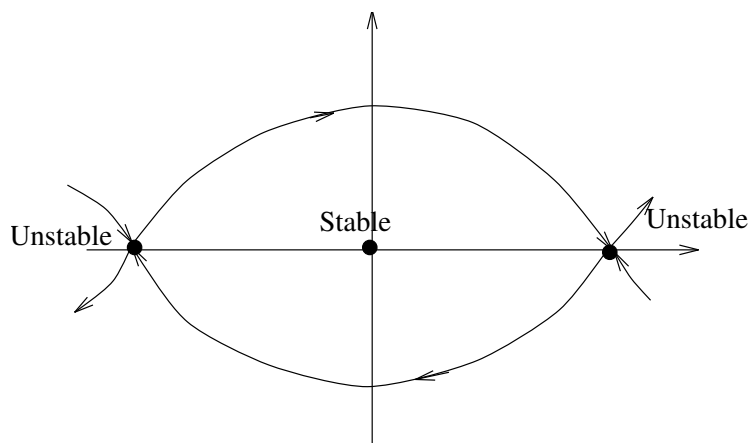
(231) gives a family of curves called the *orbits* of the system. (The term ‘orbit’ arises from celestial mechanics which has provided much of the motivation for the study of nonlinear systems.) You can sketch these orbits by hand, but a computer program will make it quite a bit easier. Such a diagram is called a *phase portrait* of the system. (The term ‘phase’ is used because the x_1, x_2 -plane is called the *phase plane* for historical reasons.) Note that (231) exhibits the orbits as the level curves of a function but it does not tell you the directions that solutions ‘flow’ along those orbits. It is easy to determine these directions by examining the vector $\frac{d\mathbf{x}}{dt}$ at typical points. For example, if $0 < x_1 < \pi$ and $x_2 > 0$, then from (230), we see that $\frac{dx_1}{dt} = x_2 > 0$ and $\frac{dx_2}{dt} = -\frac{g}{L} \sin x_1 < 0$. It follows that the solutions move along the orbits downward and to the right in that region.



Examination of the phase portrait exhibits some interesting phenomena. First look at the vaguely elliptical orbits which circle the origin. These represent periodic solutions in which the pendulum swings back and forth. (To see that, it suffices to follow what happens to $\theta = x_1$ as the solution moves around an orbit in the phase plane.)



Next look at the origin. This represents the constant solution $x_1 = 0, x_2 = 0$, in which the pendulum does not move at all ($\mathbf{x}'(t) = 0$ for all t). This may be obtained from (231) by taking $C = -2g$. There are two other constant solutions represented in the phase portrait: for $C = 2g$ we obtain $x_1 = \pi, x_2 = 0$ or $x_1 = -\pi, x_2 = 0$. These represent the same physical situation; the pendulum is precariously balanced on top of the rod ($\theta = \pi$ or $\theta = -\pi$). This physical situation is an example of an *unstable equilibrium*. Given a slight change in its position θ or velocity θ' , the pendulum will move off the balance point, and the corresponding solution will eventually move quite far from the equilibrium point. On the other hand, the constant solution $x_1 = x_2 = 0$ represents a *stable equilibrium*.



Consider the two orbits which appear to connect the two unstable equilibrium points. *Neither of the equilibrium points is actually part of either orbit.* For, once

the pendulum is in equilibrium it will stay there forever, and if it is not in equilibrium, it won't ever get there. You should think this out carefully for yourself and try to understand what actually happens to the pendulum for each of these solutions.

The remaining orbits represent motions in which the pendulum swings around the pivot in circuits which repeat indefinitely. (The orbits don't appear to repeat in the phase plane, but you should remember that values of $x_1 = \theta$ differing by 2π represent the same physical configuration of the pendulum.)

It should be noted in passing that different solutions of the system of differential equations can produce the same orbit. Indeed, we can start a solution off at $t = t_0$ at any point \mathbf{x}_0 on an orbit, and the resulting solution will trace out that orbit. Solutions obtained this way are the same except for a shift in the time scale.

Example 244 In the study of ecological systems, one is often interested in the dynamics of populations. Earlier in this course we considered the growth or decline of a single population. Consider now two populations x and y where the size of each depends on the other. For example, x might represent the number of caribou present in a given geographical region and y might represent the number of wolves which prey on the caribou. This is a so-called *prey-predator* problem. The mathematical model for such an interaction is often expressed as a system of differential equations of the form

$$\begin{aligned}\frac{dx}{dt} &= px - qxy \\ \frac{dy}{dt} &= -ry + sxy\end{aligned}\tag{232}$$

where p, q, r, s are *positive* constants. The justification for such a model is as follows. In absence of predators, the prey will follow a Malthusian law $dx/dt = px$ where p is the birthrate of the prey. However, in the presence of predators, there will be an additional term limiting the rate of growth of x which depends on the likelihood of an encounter between prey and predator. This likelihood is assumed to be proportional to the product xy of the two population sizes. Similarly, without prey, it is assumed that the population of predators will decline according to the Malthusian law $dy/dt = -ry$, but then the term sxy is added to account for the rate of population growth for the predators which can be supported from the existing prey. Note that this model is derived from rather simple minded considerations. Even so, the predictions of such a model may correspond quite well with observations. However, one should bear in mind that there may be other models which work just as well.

As in the previous example, the system (232) can't be solved explicitly at a function of t , but we can get a pretty good description of its phase portrait.

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{-ry + sxy}{px - qxy}$$

yields

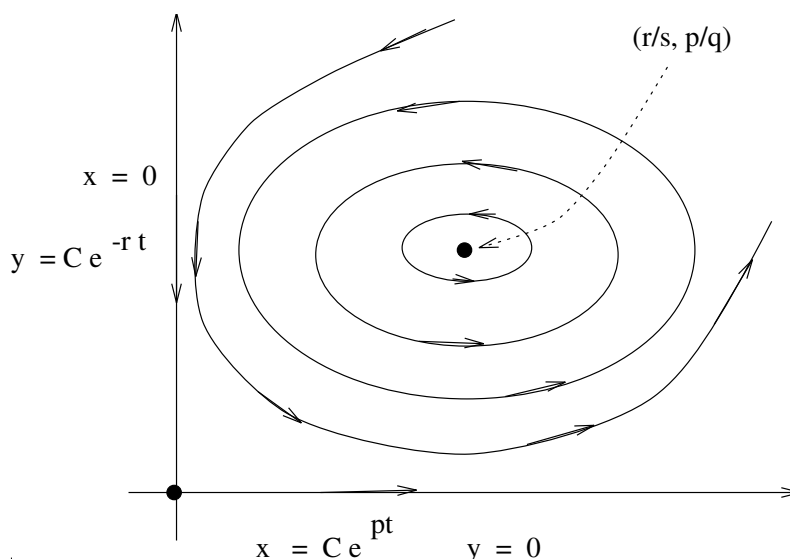
$$(ry - sxy)dx + (px - qxy)dy = 0$$

$$(r - sx)\frac{dx}{x} + (p - qy)\frac{dy}{y} = 0$$

$$r \ln x - sx + p \ln y - qy = C$$

$$\ln x^r y^p - (sx + qy) = C.$$

These curves may be graphed (after choosing plausible values of the constants p, q, r, s).



We only included orbits in the first quadrant since those are the only ones that have significance for population problems.

Note that there are two constant solutions. First, $x = y = 0$ is certainly an equilibrium point. The other constant solution may be determined from (232) by setting

$$\begin{aligned} \frac{dx}{dt} &= px - qxy = 0 \\ \frac{dy}{dt} &= -ry + sxy = 0. \end{aligned}$$

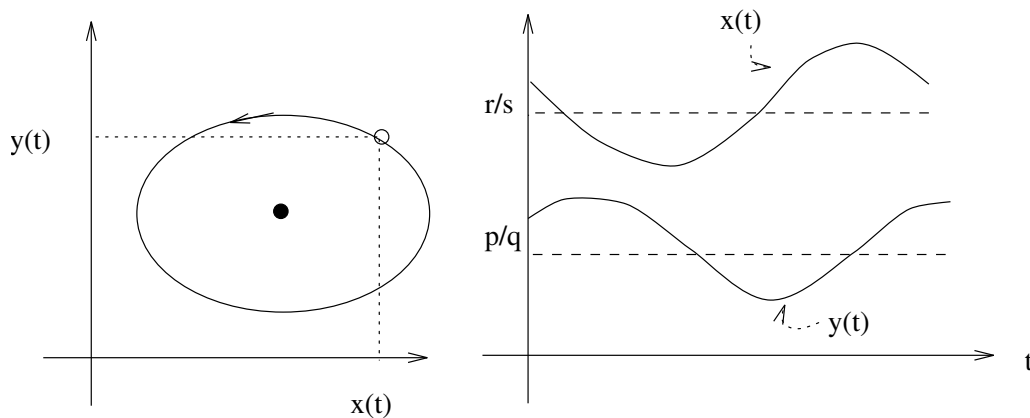
For, anything obtained this way will certainly be constant and also a solution of (232). In this case, we get

$$\begin{aligned} x(p - qy) &= 0 \\ y(r - sx) &= 0. \end{aligned}$$

From the first equation $x = 0$ or $y = p/q$. From the second $y = 0$ or $x = r/s$. However, $x = 0$ is not consistent with $x = r/s$ and similarly $y = 0$ is not consistent with $y = p/q$, so we obtain a second constant solution $x = r/s, y = p/q$. It represents an equilibrium in which both populations stay fixed at non-zero values.

The positive x axis is an orbit, and it corresponds to the situation where there are no predators ($x = Ce^{pt}, y = 0$). Note that this orbit does not contain the origin, but $\lim_{t \rightarrow -\infty} x(t) = 0$. Similarly, the positive y axis represents the situation with no prey ($x = 0, y = Ce^{-rt}$). This shows that the origin represents an unstable equilibrium. On the other hand, the point $(r/s, p/q)$ represents a stable equilibrium. (Can you see why?)

The remaining orbits correspond to solutions in which each population varies periodically. However, the exact relation between the times of maximum population for prey and predator may be quite subtle.



Orbit

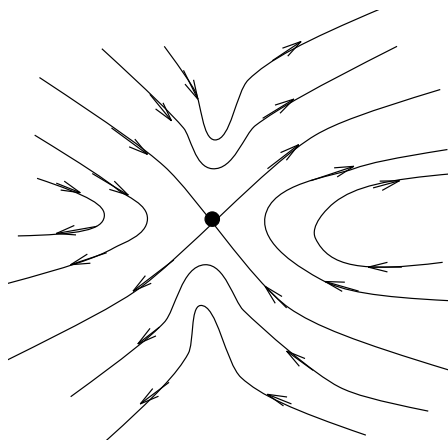
Two populations as functions of time

The general nonlinear first order system has the form

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, t)$$

where $\mathbf{f}(\mathbf{x}, t)$ is vector valued function taking values in \mathbf{R}^n and $\mathbf{x} = \mathbf{x}(t)$ is a (n -dimensional) vector valued solution. It is often the case, as in both examples, that \mathbf{f} doesn't depend explicitly on t . Such systems are called *time independent* or *autonomous*. The *critical points* of a system are those points \mathbf{a} satisfying $\mathbf{f}(\mathbf{a}, t) = 0$ for all t , or, in the autonomous case, just $\mathbf{f}(\mathbf{a}) = 0$. Each critical point corresponds to a constant solution, $\mathbf{x}(t) = \mathbf{a}$ for all t . The behavior of solutions

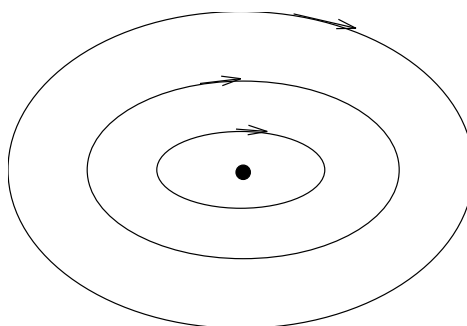
near a critical point in the phase portrait of the system tell us something about the stability of the equilibrium represented by the critical point. In the examples we noted at least two different types of behavior near a critical point.



Unstable equilibrium

Saddle point

In the above case, called a *saddle point*, The above situation occurred in the pendulum example at the unstable equilibria. Some solutions approach the critical point and then depart, some solutions approach the critical point asymptotically as $t \rightarrow \infty$, some solutions do the reverse for $t \rightarrow -\infty$. Such a critical point is called a *saddle point*,



Stable equilibrium

Center

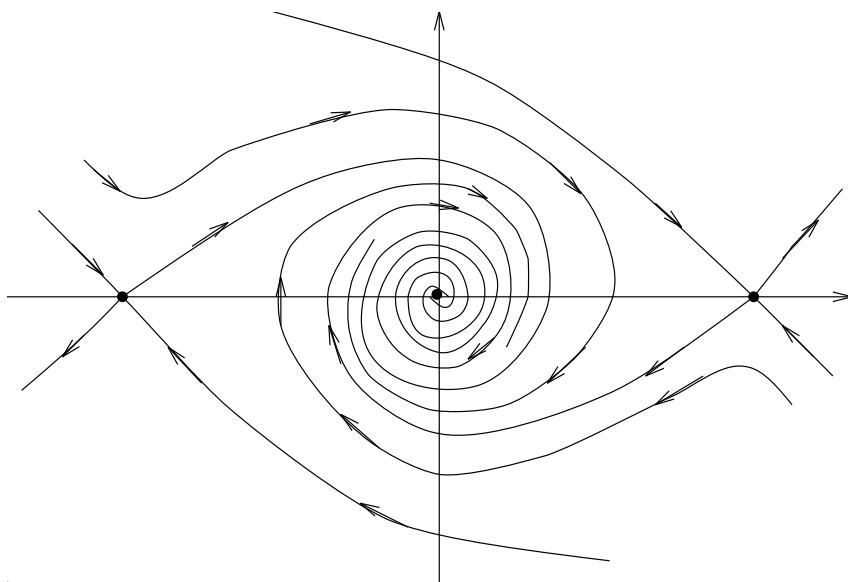
The above situation occurred in both examples. Nearby solutions repeat periodically. Such a critical point is called a *center*.

There are many other possibilities.

Example 245 Consider a damped rigid pendulum with damping dependent on velocity. Newton's second law results in a differential equation of the form

$$\theta'' = -\frac{g}{L} \sin \theta - a\theta'$$

where $a > 0$. It is fairly clear how the damping will affect the behavior of such a pendulum. There will still be unstable equilibria in the phase plane which correspond to the pendulum precariously balanced on end ($x_1 = \theta$ equal to an odd multiple of π and $x_2 = \theta' = 0$). There will be stable equilibria in the phase plane which correspond to the pendulum hanging downward at rest ($x_1 = \theta$ equal to an even multiple of π and $x_2 = \theta' = 0$). Near the stable equilibrium, the pendulum will oscillate with decreasing amplitude (and velocity), so the corresponding orbits in the phase plane will spiral inward toward the equilibrium point and approach it asymptotically. Here is what the phase portrait looks like in general.



The above physical reasoning is convincing, but it is also useful to have a more mathematical approach. To this end, we convert the second order equation for θ to the system

$$\begin{aligned} x_1' &= x_2 \\ x_2' &= -\frac{g}{L} \sin x_1 - ax_2. \end{aligned} \quad (233)$$

The critical points are $(n\pi, 0)$ where $n = 0, \pm 1, \pm 2, \dots$. The method we used for sketching the phase portrait of the undamped pendulum doesn't work in this case, but we can use what we learned previously to guide us to an understanding of the behavior near the critical point $(0, 0)$ as follows. Let

$$U = \frac{1}{2}x_2^2 - \frac{g}{L} \cos x_1.$$

This quantity was *constant* in the undamped case by the law of conservation of energy. Since some energy is lost because of the damping term, U is no longer

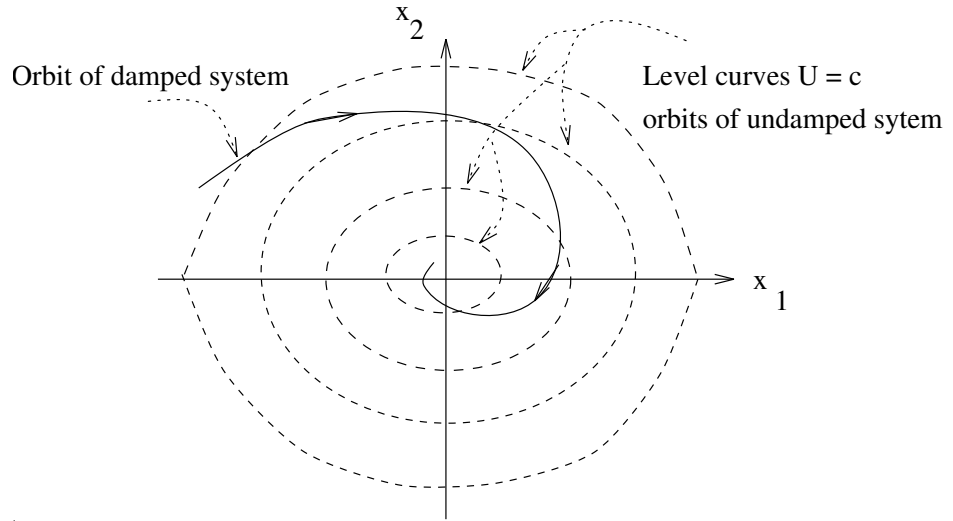
constant. Indeed, we have

$$\begin{aligned}\frac{dU}{dt} &= \frac{1}{2}2x_2x_2' + \frac{g}{L}\sin x_1 x_1' \\ &= x_2\left(-\frac{g}{L}\sin x_1 - ax_2\right) + \frac{g}{L}\sin x_1 x_2 \\ &= -ax_2^2.\end{aligned}$$

Thus, $dU/dt < 0$ except where the path crosses the x_1 -axis (i.e., $x_2 = 0$). However, at points on the x_1 -axis, the velocity $d\mathbf{x}/dt$ (given by (233)) is directed off the axis. Using this information, it is possible to see that U decreases steadily along any orbit of the damped system as the orbit descends inward crossing the level curves

$$Lx_2^2 - 2g\cos x_1 = 2LU = C$$

of the undamped system. In the limit, the orbit approaches the critical point at the origin. This confirms the physical interpretation described above near the critical point $x_1 = \theta = 0$, $x_2 = \theta' = 0$. Here is what the phase portrait looks like in general.



As noted above, the orbits around the critical point at the origin spiral in toward the critical point and approach it asymptotically. In general, such a critical point is called a *focus*. In this case, the critical point represents a stable equilibrium, but, in general, the reverse situation where the orbits spiral out from the origin is possible. Then the critical point represents an unstable equilibrium, but is still called a focus.

Generally, we may summarize the phenomena illustrated above by calling a critical point *stable* if any solution which comes sufficiently close to the critical point for one

time t stays close to it for all subsequent time. Otherwise, we shall call the critical point *unstable*. If *all nearby* solutions approach the critical point asymptotically as $t \rightarrow \infty$, it is called *asymptotically stable*.

As we shall see, the behavior of the phase portrait near a critical point can often be reduced to a problem in linear algebra.

Exercises for 13.1.

- In each case find the critical points of the indicated system.
 - $x'_1 = x_1 - x_1x_2$, $x'_2 = -3x_2 + 2x_1x_2$.
 - $x'_1 = -x_1 + 2x_1x_2$, $x'_2 = x_2 - x_2^2 + x_1x_2$.
 - $x'_1 = x_1 - x_1x_2$, $x'_2 = -x_2 + x_1x_2$, $x'_3 = x_3 + x_1x_2$.
- Consider the general 2×2 linear system $\mathbf{x}' = A\mathbf{x}$. Show that $x_1 = 0$, $x_2 = 0$ gives the only critical point if $\det A \neq 0$. What happens if $\det A = 0$? How does this generalize to an $n \times n$ linear system?
- For each of the following systems, first find the critical points. Then eliminate t , solve the ensuing first order differential equation in x_1 and x_2 and sketch some orbits. Use the system to determine the direction of 'flow' at typical points along the orbits.
 - $x'_1 = -x_2$, $x'_2 = 2x_1$.
 - $x'_1 = 2x_1 + x_2$, $x'_2 = x_1 - 2x_2$.
 - $x'_1 = x_1 - x_1x_2$, $x'_2 = -x_2 + x_1x_2$.

13.2 Linear Approximation

Suppose we want to study the behavior of a system

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x})$$

near a critical point $\mathbf{x} = \mathbf{a}$. (We assume the system is time independent for simplicity.) One way to do this is to approximate $\mathbf{f}(\mathbf{x})$ by a *linear* function in the vicinity of the point \mathbf{a} . In order to do this, we need a short digression about multidimensional calculus.

The Derivative of a Function $f : \mathbf{R}^n \rightarrow \mathbf{R}^m$ Let $\mathbf{f} : \mathbf{R}^n \rightarrow \mathbf{R}^m$ denote a smooth function. Since \mathbf{f} is an m -dimensional vector valued function, it may be specified

by m component functions

$$\mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix}$$

where each component is a scalar function $f_i(x_1, x_2, \dots, x_n)$ of n variables. Fix a point $\mathbf{a} = (a_1, a_2, \dots, a_n)$ in the domain of \mathbf{f} . For each component, we have the linear approximation

$$f_i(\mathbf{x}) = f_i(\mathbf{a}) + \nabla f_i(\mathbf{a}) \cdot (\mathbf{x} - \mathbf{a}) + o(|\mathbf{x} - \mathbf{a}|).$$

(See Chapter III, Section 4.) We may put these together in a single vector equation

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{a}) \\ f_2(\mathbf{a}) \\ \vdots \\ f_m(\mathbf{a}) \end{bmatrix} + \begin{bmatrix} \nabla f_1(\mathbf{a}) \\ \nabla f_2(\mathbf{a}) \\ \vdots \\ \nabla f_m(\mathbf{a}) \end{bmatrix} (\mathbf{x} - \mathbf{a}) + o(|\mathbf{x} - \mathbf{a}|).$$

Let

$$D\mathbf{f} = \begin{bmatrix} \nabla f_1 \\ \nabla f_2 \\ \vdots \\ \nabla f_m \end{bmatrix}$$

be the $m \times n$ matrix with rows the gradients of the component functions f_i . The i, j entry of $D\mathbf{f}$ is $\frac{\partial f_i}{\partial x_j}$. Then the above equation may be written more compactly

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{a}) + D\mathbf{f}(\mathbf{a})(\mathbf{x} - \mathbf{a}) + o(|\mathbf{x} - \mathbf{a}|). \quad (234)$$

The matrix $D\mathbf{f}$ is called the *derivative* of the function \mathbf{f} . It plays the same role for vector valued functions that the gradient plays for scalar valued functions.

Example 246 Let $m = n = 2$ and suppose

$$\mathbf{f}(x_1, x_2) = \begin{bmatrix} x_2 \\ -\sin x_1 \end{bmatrix}.$$

Let's consider the behavior of this function near $\mathbf{a} = (0, 0)$.

First calculate the derivative.

$$\begin{aligned} \nabla f_1 &= [0 \quad 1] \\ \nabla f_2 &= [-\cos x_1 \quad 0] \end{aligned}$$

so

$$D\mathbf{f}(0, 0) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Hence, formula (234) yields

$$\begin{aligned}\mathbf{f}(x_1, x_2) &= \mathbf{f}(0, 0) + D\mathbf{f}(0, 0)\mathbf{x} + o(|x|) \\ &\approx \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ -x_1 \end{bmatrix}.\end{aligned}$$

Example 247 Let $m = 2, n = 3$ and take

$$\mathbf{f}(x_1, x_2, x_3) = \begin{bmatrix} x_1 + 2x_2 + x_3 \\ x_1^2 + x_2^2 + x_3^2 \end{bmatrix}.$$

Then

$$D\mathbf{f}(x_1, x_2, x_3) = \begin{bmatrix} 1 & 2 & 1 \\ 2x_1 & 2x_2 & 2x_3 \end{bmatrix}.$$

Suppose we want to study the behavior of \mathbf{f} near the point $(1, 2, -4)$. We have

$$\mathbf{f}(1, 2, -4) = \begin{bmatrix} 1 \\ 21 \end{bmatrix}$$

and

$$D\mathbf{f}(1, 2, -4) = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & -8 \end{bmatrix}$$

so, according to (234),

$$\mathbf{f}(x_1, x_2, x_3) \approx \begin{bmatrix} 1 \\ 21 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & -8 \end{bmatrix} \begin{bmatrix} x_1 - 1 \\ x_2 - 2 \\ x_3 + 4 \end{bmatrix}.$$

The linear approximation is an invaluable tool for the study of functions $\mathbf{R}^n \rightarrow \mathbf{R}^m$. We have already seen its use in a variety of circumstances for scalar valued functions. It also arose implicitly when we were studying change of variables for multiple integrals. For example, a change of variables in \mathbf{R}^2 may be described by a function $\mathbf{g} : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ and the Jacobian determinant used in the correction factor for double integrals is just

$$\det D\mathbf{g} = \det \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{bmatrix}.$$

A similar remark applies in \mathbf{R}^3 , and indeed the change of variable rule may be generalized to integrals in \mathbf{R}^n by using the correction factor $|\det D\mathbf{g}|$ where $\mathbf{g} : \mathbf{R}^n \rightarrow \mathbf{R}^n$ is the function giving the change of variables.

Analysis of Non-Linear Systems near Critical Points We want to apply the above ideas to the analysis of an autonomous nonlinear system

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}) \tag{235}$$

near a critical point \mathbf{a} . Here, $m = n$ and $\mathbf{f} : \mathbf{R}^n \rightarrow \mathbf{R}^n$. By definition, $\mathbf{f}(\mathbf{a}) = 0$ at a critical point, so the linear approximation gives

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{a}) + D\mathbf{f}(\mathbf{a})(\mathbf{x} - \mathbf{a}) + o(|\mathbf{x} - \mathbf{a}|) = D\mathbf{f}(\mathbf{a})(\mathbf{x} - \mathbf{a}) + o(|\mathbf{x} - \mathbf{a}|).$$

If we put this in (235) and drop the error term, we obtain

$$\frac{d\mathbf{x}}{dt} = D\mathbf{f}(\mathbf{a})(\mathbf{x} - \mathbf{a}).$$

This may be simplified further by the change of variables $\mathbf{y} = \mathbf{x} - \mathbf{a}$ which in essence moves the critical point to the origin. Since $\frac{d\mathbf{y}}{dt} = \frac{d\mathbf{x}}{dt}$, this yields

$$\frac{d\mathbf{y}}{dt} = A\mathbf{y} \tag{236}$$

where $A = D\mathbf{f}(\mathbf{a})$ is the derivative matrix at the critical point. In this way, we have replaced the nonlinear system (235) by the *linear system* (236), at least near the critical point. If dropping the error term $o(|\mathbf{x} - \mathbf{a}|)$ doesn't affect things too severely, we expect the behavior of solutions of the linear system to give a pretty good idea of what happens to solutions of the nonlinear system near the critical point. At least in principle, we know how to solve linear systems.

Example 248 Consider the pendulum system

$$\begin{aligned} \frac{dx_1}{dt} &= x_2 \\ \frac{dx_2}{dt} &= -\sin x_1 \end{aligned}$$

where the units have been chosen so $g = L$. Consider the critical point $(\pi, 0)$ corresponding to the unstable equilibrium discussed earlier. We have

$$D\mathbf{f} = \begin{bmatrix} 0 & 1 \\ -\cos x_1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{at } (\pi, 0).$$

Hence, putting $y_1 = x_1 - \pi$, $y_2 = x_2$, near the critical point, the system is approximated by the linear system

$$\frac{d\mathbf{y}}{dt} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \mathbf{y}.$$

(In the ' y ' coordinates, the critical point is at the origin.) This system is quite easy to solve. I leave the details to you. The eigenvalues are $\lambda = 1, \lambda = -1$. A basis of eigenvectors is given by

$$\begin{aligned} \mathbf{v}_1 &= \begin{bmatrix} -1 \\ 1 \end{bmatrix} && \text{corresponding to } \lambda = 1 \\ \mathbf{v}_2 &= \begin{bmatrix} 1 \\ 1 \end{bmatrix} && \text{corresponding to } \lambda = -1. \end{aligned}$$

The general solution of the linear system is

$$\mathbf{y} = c_1 e^t \mathbf{v}_1 + c_2 e^{-t} \mathbf{v}_2 \quad (237)$$

or

$$\begin{aligned} y_1 &= -c_1 e^t + c_2 e^{-t} \\ y_2 &= c_1 e^t + c_2 e^{-t} \end{aligned}$$

The phase portrait for the solution near $(0, 0)$ can be worked out by trying various c_1 and c_2 and sketching the resulting orbits. However, it is easier to see what it looks like if we put $P = [\mathbf{v}_1 \ \mathbf{v}_2]$ and make the change of coordinates $\mathbf{y} = P\mathbf{z}$. As in Chapter XII, Section 6, this will have the effect of diagonalizing the coefficient matrix and yielding the system

$$\begin{aligned} \frac{dz_1}{dt} &= z_1 \\ \frac{dz_2}{dt} &= -z_2 \end{aligned}$$

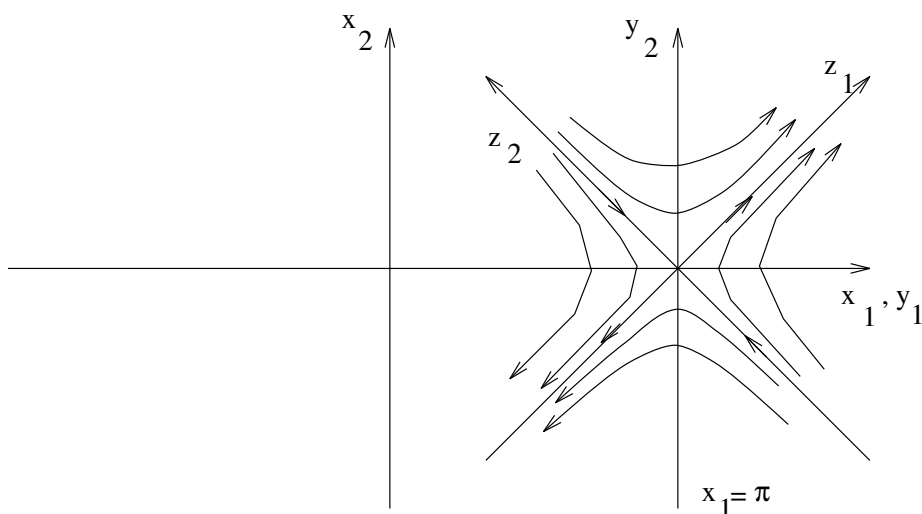
with solutions

$$z_1 = c_1 e^t \quad z_2 = c_2 e^{-t}. \quad (238)$$

From this, it is clear that the orbits are the family of hyperbolas given by

$$z_1 z_2 = c_1 c_2 = c.$$

This includes the case $c = 0$ which gives the degenerate ‘hyperbola’ consisting of the lines $z_1 = 0$ and $z_2 = 0$. (The degenerate case actually consists of five orbits: the critical point $(0, 0)$ and the positive and negative half lines on each axis.)



The z_1 and z_2 axes in this diagram are directed respectively along the vectors \mathbf{v}_1 and \mathbf{v}_2 . The direction in which each orbit is traversed may be determined from the explicit solutions (238).

You should compare the above picture with the phase portrait derived in Section 1 for the pendulum problem. In this example, the local analysis merely confirms what we already knew from the phase portrait that was derived in Section 1. However, in general, sketching the phase portrait of the nonlinear system may be very difficult or even impossible. Hence, deriving a picture near each critical point by linear analysis is a useful first step in understanding the whole picture.

The previous example illustrates generally what happens for a 2 dimensional system when the eigenvalues are *real* and of *opposite signs*. In this case the critical point is called a *saddle point*, and it is clearly unstable.

Example 249 Consider the prey-predator problem described by the system

$$\begin{aligned}\frac{dx_1}{dt} &= x_1 - x_1x_2 \\ \frac{dx_2}{dt} &= -x_2 + x_1x_2.\end{aligned}$$

This assumes that the populations are being measured in bizarre units, and with respect to these units the constants p, q, r, s are all 1. This is rather unrealistic, but it does simplify the algebra quite a lot. As before, there are two critical points in the first quadrant: $(0, 0)$ and $(r/s, p/q) = (1, 1)$. Let's study the linear approximation near $(1, 1)$. We have

$$D\mathbf{f} = \begin{bmatrix} 1 - x_2 & -x_1 \\ x_2 & -1 + x_1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \quad \text{at } (1, 1).$$

Hence, the approximating linear system is

$$\frac{d\mathbf{y}}{dt} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \mathbf{y}$$

where $y_1 = x_1 - 1, y_2 = x_2 - 1$. I leave it to you to work out the solutions of this system. The eigenvalues are $\lambda = \pm i$, so we need to use complex solutions. An eigenvector for $\lambda = i$ is

$$\mathbf{u} = \begin{bmatrix} i \\ 1 \end{bmatrix}$$

and a corresponding complex solution is

$$e^{it}\mathbf{u}.$$

As usual, we can find a linearly independent pair of real solutions by taking real and imaginary parts. To help you work similar problems in homework, we shall do this in slightly greater generality than necessary in the particular case.

$$\mathbf{u} = \mathbf{v} + i\mathbf{w} \quad \text{where } \mathbf{v} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{w} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Then

$$e^{it}\mathbf{u} = (\cos t + i \sin t)(\mathbf{v} + i\mathbf{w}) = \mathbf{v} \cos t - \mathbf{w} \sin t + i(\mathbf{v} \sin t + \mathbf{w} \cos t).$$

Taking real and imaginary parts yields the two solutions

$$\begin{aligned}\mathbf{v} \cos t - \mathbf{w} \sin t &= [\mathbf{v} \quad \mathbf{w}] \begin{bmatrix} \cos t \\ -\sin t \end{bmatrix} \\ \mathbf{v} \sin t + \mathbf{w} \cos t &= [\mathbf{v} \quad \mathbf{w}] \begin{bmatrix} \sin t \\ \cos t \end{bmatrix}.\end{aligned}$$

This suggests that we put $P = [\mathbf{v} \quad \mathbf{w}]$ (so that its columns are the real and imaginary parts of the eigenvector \mathbf{u}), and make the change of variables $\mathbf{y} = P\mathbf{z}$. Then, in the ‘ z ’ coordinate system, we obtain two linearly independent solutions

$$\begin{bmatrix} \cos t \\ -\sin t \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \sin t \\ \cos t \end{bmatrix}.$$

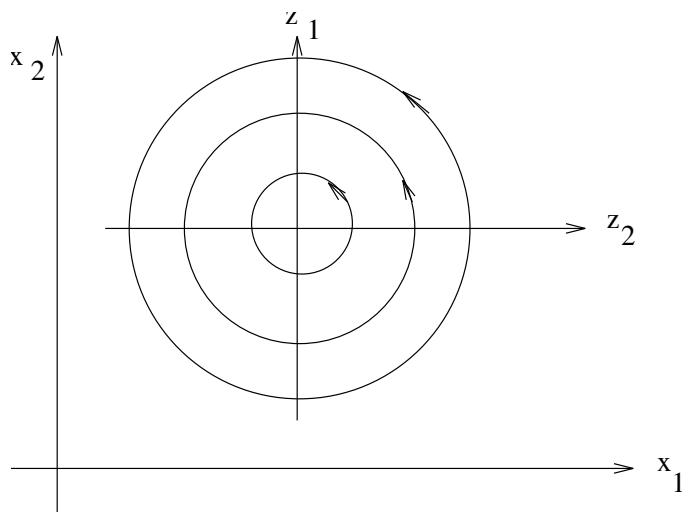
Any solution is then a linear combination

$$\mathbf{z} = c_1 \begin{bmatrix} \cos t \\ -\sin t \end{bmatrix} + c_2 \begin{bmatrix} \sin t \\ \cos t \end{bmatrix} = \begin{bmatrix} c_1 \cos t + c_2 \sin t \\ -c_1 \sin t + c_2 \cos t \end{bmatrix}.$$

Now put $c_1 = A \cos \delta$, $c_2 = A \sin \delta$. The above solution takes the form

$$\mathbf{z} = \begin{bmatrix} A \cos(t - \delta) \\ -A \sin(t - \delta) \end{bmatrix}.$$

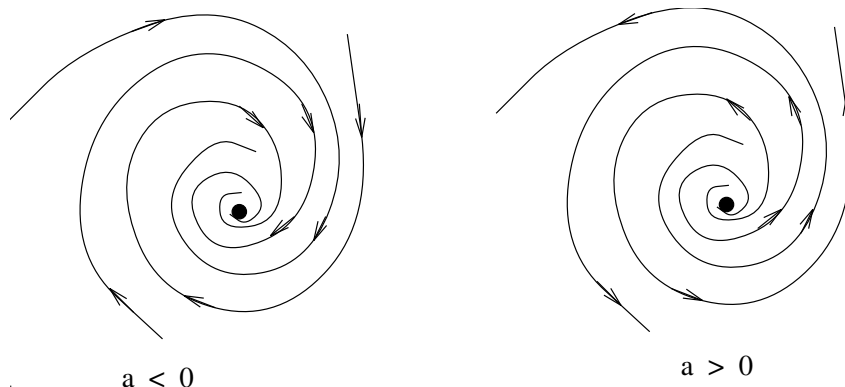
This gives a family of *circles* which are traversed clockwise with respect to the z_1 and z_2 axes. However, the z_1 and z_2 axes have orientation opposite to that of the original axes. (Look at $\mathbf{v} = \mathbf{e}_2$ and $\mathbf{w} = \mathbf{e}_1$ which are ‘unit’ vectors along the new axes.) Thus, with respect to the y_1, y_2 axes, the motion is counter-clockwise.



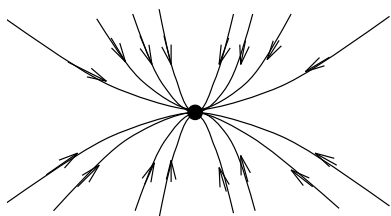
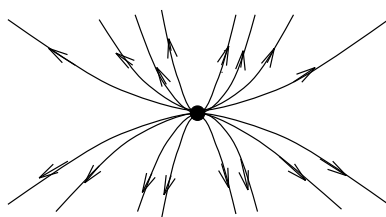
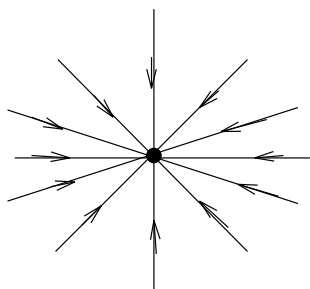
Note that the phase portrait we derived in Section 1 for the prey-predator system gives a similar picture near the critical point $(\pi, 0)$.

A similar analysis applies to any 2-dimensional real system if the eigenvalues are complex. If the eigenvalues are purely imaginary, i.e., of the form $\pm i\omega$, then the results are similar to what we got in Example 249. In the ' z ' coordinate system, the orbits look like circles and they are traversed clockwise. However, the change from the ' y ' coordinates to the ' z ' coordinates may introduce changes of scale, different for the two axes. Hence, the orbits are really ellipses when viewed in the original coordinate system. Also, as in Example 249, the change of coordinates may introduce a reversal of orientation. A critical point of this type is called a *center*.

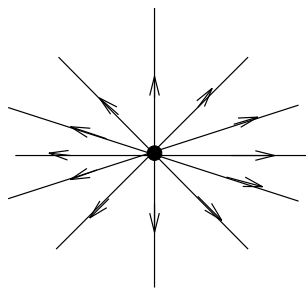
If the eigenvalues are not purely imaginary, then they are of the form $a \pm bi$ with $a \neq 0$, and all solutions have the additional factor e^{at} . This has the effect of turning the 'ellipses' into spirals. If $a < 0$, $e^{at} \rightarrow 0$ as $t \rightarrow \infty$, so all solutions spiral in towards the origin. If $a > 0$, they all spiral out. In either case, the critical point is called a *focus*. If $a < 0$ the critical point is stable, and if $a > 0$ it is not.



The above examples by no means exhaust all the possibilities even in the 2-dimensional case. In the remaining cases, the eigenvalues λ_1, λ_2 are real and of the same sign. The exact nature of the phase portrait will depend on how the linear algebra works out, but in all these cases the critical point is called a *node*. Here are some pictures of nodes:

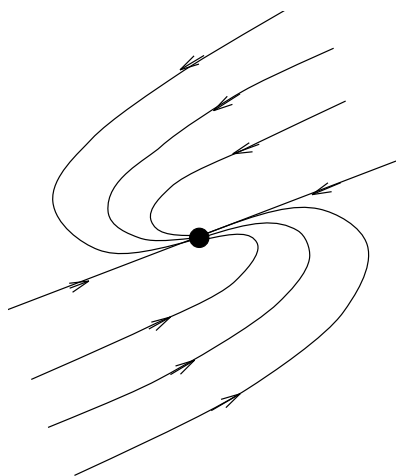
Stable node $\lambda_1 < \lambda_2 < 0$ Unstable node $\lambda_1 > \lambda_2 > 0$ 

Stable

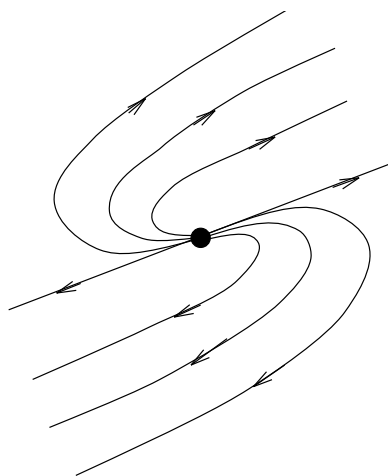


Unstable

Nodes with $\lambda_1 = \lambda_2$ and two linearly independent eigenvectors



Stable



Unstable

Nodes with $\lambda_1 = \lambda_2$, generalized eigenvectors needed

Note that in each case the critical point is stable if the eigenvalues are negative.

Example 250 Consider the linear system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} -3 & -1 \\ 1 & -1 \end{bmatrix} \mathbf{x}.$$

The only critical point is the origin, and of course the linear approximation there is just the original system.

It turns out that there is no basis of eigenvectors in this case, so we must use the method of generalized eigenvectors. In fact, this example was worked out in Chapter XI, Section 8, Example 2. We found there that

$$\begin{aligned} \mathbf{x}_1 &= e^{-2t}(\mathbf{e}_1 + t\mathbf{v}_2) = [\mathbf{e}_1 \quad \mathbf{v}_2] \begin{bmatrix} e^{-2t} \\ te^{-2t} \end{bmatrix} \\ \mathbf{x}_2 &= e^{-2t}\mathbf{v}_2 = [\mathbf{e}_1 \quad \mathbf{v}_2] \begin{bmatrix} 0 \\ e^{-2t} \end{bmatrix} \end{aligned}$$

form a basis for the solution space, where

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{v}_2 = (A + 2I)\mathbf{e}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

The general solution is

$$\mathbf{x} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 = [\mathbf{e}_1 \quad \mathbf{v}_2] \begin{bmatrix} c_1e^{-2t} \\ c_1te^{-2t} + c_2e^{-2t} \end{bmatrix}.$$

This suggests making the change of coordinates $\mathbf{x} = P\mathbf{z}$ where

$$P = [\mathbf{e}_1 \quad \mathbf{v}_2] = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}.$$

In the new coordinates, the general solution is given by

$$\mathbf{z} = \begin{bmatrix} c_1e^{-2t} \\ c_1te^{-2t} + c_2e^{-2t} \end{bmatrix}$$

which in components is

$$\begin{aligned} z_1 &= c_1e^{-2t} \\ z_2 &= c_1te^{-2t} + c_2e^{-2t}. \end{aligned}$$

As $t \rightarrow \infty$, both z_1 and z_2 approach zero because of the factor e^{-2t} . Also,

$$\frac{z_2}{z_1} = t + \frac{c_2}{c_1}$$

so for large t , both z_1 and z_2 have the same sign. On the other hand, for t sufficiently negative, z_1 and z_2 have opposite signs, i.e., \mathbf{z} starts off either in the fourth quadrant

or the second quadrant. (Note that this argument required $c_1 \neq 0$. What does the orbit look like if $c_1 = 0$?)

We leave it as an challenge for you to sketch the orbits of this system. You should get one of the diagrams sketched above. You should first do it in the z_1, z_2 coordinate system. However, to interpret this in the original x_1, x_2 coordinates, you should notice that the 'z' axes are not perpendicular. Namely, the z_1 -axis is the same as the x_1 axis, but the z_2 -axis points along the vector \mathbf{v}_2 , so it makes an angle of $3\pi/4$ with the positive x_1 axis.

If the *matrix* $D\mathbf{f}(\mathbf{a})$ is *singular*, the theory utilized above breaks down. The following example indicates how that might occur.

Example 251 Consider the system

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{x}.$$

The origin is a critical point, and since the system is linear, the linear approximation there is the same as the system itself.

The eigenvalues turn out to be $\lambda = -2$ and $\lambda = 0$. For $\lambda = -2$ a basic eigenvector is

$$\mathbf{v}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

For $\lambda = 0$, a basic eigenvector is

$$\mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Put these together to form a change of basis matrix

$$P = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Then, by our theory

$$P^{-1} \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} P = \begin{bmatrix} -2 & 0 \\ 0 & 0 \end{bmatrix}$$

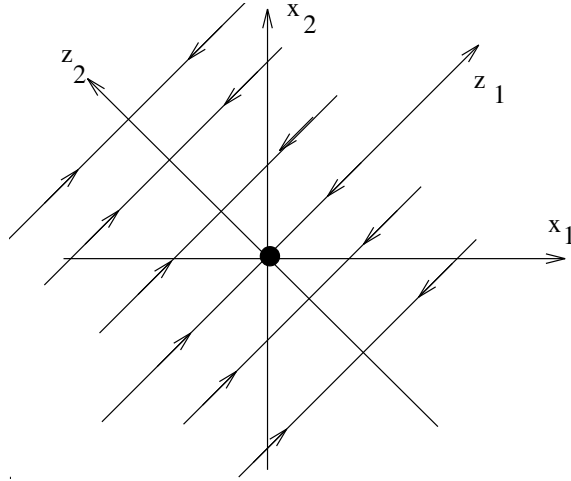
and the change of variables $\mathbf{x} = P\mathbf{z}$ yields the 'uncoupled' system

$$\begin{aligned} \frac{dz_1}{dt} &= -2z_1 \\ \frac{dz_2}{dt} &= 0. \end{aligned}$$

The solution is

$$z_1 = C_1 e^{-2t} \quad z_2 = C_2$$

This is a family of half lines perpendicular to the z_2 axis. Every point on the z_2 -axis is a critical point and is approached asymptotically from either side as $t \rightarrow \infty$.



In the above discussion, we have been assuming that the behavior of a non-linear system near a critical point may be determined from the behavior of the linear approximation. Unfortunately, that is not always the case. First of all, if the matrix $D\mathbf{f}(\mathbf{a})$ is singular, the behavior near the critical point depends strongly on the higher order terms which were ignored in forming the linear approximation. Even if $D\mathbf{f}(\mathbf{a})$ is non-singular, in some cases, the higher order terms can exert enough influence to change the nature of the critical point.

Example 252 Consider the nonlinear system

$$\begin{aligned}\frac{dx_1}{dt} &= x_2 - x_1(x_1^2 + x_2^2) \\ \frac{dx_2}{dt} &= -x_1 - x_2(x_1^2 + x_2^2).\end{aligned}\tag{239}$$

$(0, 0)$ is clearly a critical point. (Are there any more?) Also,

$$D\mathbf{f} = \begin{bmatrix} -3x_1^2 - x_2^2 & 1 - 2x_1x_2 \\ -1 - 2x_1x_2 & -x_1^2 - 3x_2^2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad \text{at } (0, 0).$$

Hence, the approximating linear system is

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \mathbf{x}.$$

The eigenvalues are $\lambda = \pm i$, so the critical point is a center. The phase portrait of the linear system consists of a family of closed loops centered at the origin.

On the other hand, we can solve the nonlinear system exactly in this case if we switch to polar coordinates in the x_1, x_2 -plane. Put $x_1 = r \cos \theta$, $x_2 = r \sin \theta$ in

(239). We get

$$\begin{aligned}\frac{dr}{dt} \cos \theta - r \sin \theta \frac{d\theta}{dt} &= r \sin \theta - r^3 \cos \theta \\ \frac{dr}{dt} \sin \theta + r \cos \theta \frac{d\theta}{dt} &= -r \cos \theta - r^3 \sin \theta.\end{aligned}$$

Multiply the first equation by $\cos \theta$, the second by $\sin \theta$ and add to obtain

$$\begin{aligned}\frac{dr}{dt} &= -r^3 \\ -\frac{dr}{r^3} &= dt \\ \frac{1}{2r^2} &= t + c \\ r &= \frac{1}{\sqrt{2t + c_1}}\end{aligned}$$

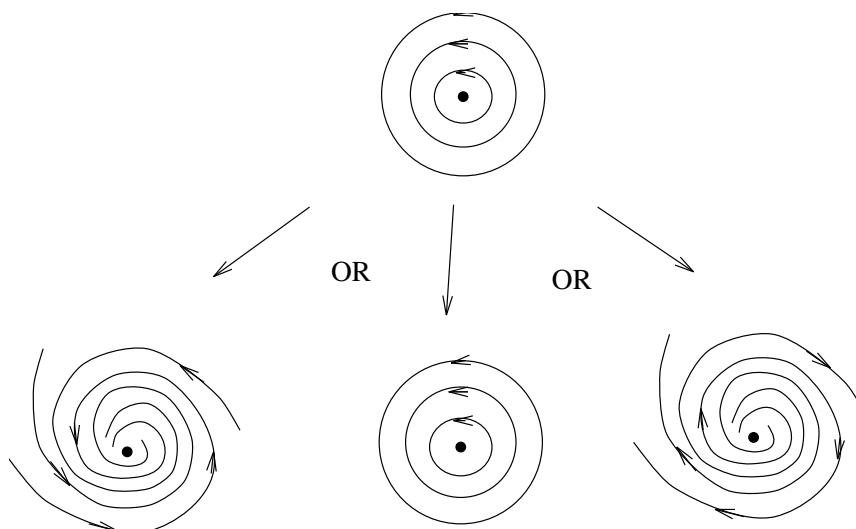
Similarly, multiplying the first equation by $\sin \theta$, the second by $\cos \theta$ and subtracting the first from the second yields

$$r \frac{d\theta}{dt} = -r.$$

However, $r = 0$ is the critical point which we already know yields a constant solution, so we may assume $r \neq 0$. Hence, we get

$$\begin{aligned}\frac{d\theta}{dt} &= -1 \\ \theta &= -t + c_2.\end{aligned}$$

This clearly represents a family of solutions which spiral in towards the origin, approaching it asymptotically at $t \rightarrow \infty$. Thus, *the additional nonlinear terms turned a center into a focus*. In this case, it is a stable focus with the orbits spiraling in toward the critical point. In other cases, the non-linear terms might perturb things in the other direction so that the orbits would spiral out from the origin



In general, for a two dimensional system, a center for the linear approximation can stay a center or can become a focus in the non-linear system. Fortunately, for a two dimensional system, if f is sufficiently smooth (i.e., C^2), and Df is non-singular at the critical point, *this is the only case* in which the nonlinear terms can significantly affect the structure of the phase portrait.

In every other non-singular case, i.e., a node, a saddle point, or a focus, a theorem of Poincaré assures us that the linear approximation gives a true picture of the phase portrait near the critical point.

Note that in the previous examples, the signs of the eigenvalues of the linear systems (or the signs of the real parts of complex eigenvalues) played an important role. It is clear why that would be the case. A basic solution will have a factor of the form e^{at} and if $a < 0$, the basic solution will necessarily converge to zero at $t \rightarrow \infty$. Hence, if these signs are negative for both basic solutions, every solution will converge to zero as $t \rightarrow \infty$, so the corresponding critical point will be asymptotically stable. On the other hand, if one eigenvalue is negative and one is positive (a saddle point), the situation is more complicated. (What if both signs are positive?)

Needless to say the situation is even more complicated for higher dimensional systems. It is still true that if all the eigenvalues of the linear approximation are negative, or, if complex, have negative real parts, then all solutions near the critical point converge to it as $t \rightarrow \infty$, i.e., the corresponding equilibrium is asymptotically stable. However, the basic linear algebra is considerably more complicated, so it is not so easy to classify exactly what happens even in the linear case. **Stable orbits**

and Attractors The behavior of a nonlinear system near its critical points helps us to understand the system, but it is certainly not the only thing of interest. For example, the solution of the prey-predator problem yields many periodic solutions.

The orbits traced out by such solutions are stable in the following sense: any solution which at some time t comes close to the orbit remains close to it for all subsequent t . That means that if we stray a little from such an orbit, we won't ever get very far from it. From this perspective, a critical point is just a stable orbit which happens to be a point. We should be interested in finding all stable orbits, but of course it is much harder to find the non-constant ones.

Much of the theory of nonlinear systems was motivated by questions in celestial mechanics. The general mathematical problem in that subject is the so-called n -body problem where we attempt to describe the motion of an arbitrary number of point masses subject to the gravitational forces between them. A complete solution of that problem still eludes us despite intense study over several centuries. Even fairly simple questions remain unanswered. For example, one would assume that the Solar System as a whole will continue to behave more or less as it does now as long as it is not disturbed by a significant perturbation such as a star passing nearby. However, no one has been able to prove, for example, that the entire system is 'stable' in the sense that it remains bounded for all time. Thus, it is conceivable that at some point in time a planet might cease to follow its normal orbit and leave the solar system altogether. (At least, it is conceivable to mathematicians, who generally believe that something may happen until they prove that it can't happen!)

Modern developments in the study of nonlinear systems are often concerned with 'stability' questions such as those mentioned above. One such result is a famous theorem of Poincaré and Bendixson. It asserts the following: if an orbit of a 2 dimensional nonlinear system enters a bounded region in the phase plane and remains there forever, and if that bounded region contains no critical points, then either the orbit is periodic itself, or it approaches a periodic orbit asymptotically. (If critical points were not excluded from the bounded region, the constant solutions represented by those points would violate the conclusion of the theorem. Also, the presence of critical points could 'disrupt' the behavior of other paths in rather subtle ways.) The following example illustrates this phenomenon.

Example 253 Consider the system

$$\begin{aligned}\frac{dx}{dt} &= -y + x(1 - x^2 - y^2) \\ \frac{dy}{dt} &= x + y(1 - x^2 - y^2).\end{aligned}$$

This system can be solved explicitly if we switch to polar coordinates. Putting $x = r \cos \theta$, $y = r \sin \theta$ in the above system yields

$$\begin{aligned}\frac{dr}{dt} \cos \theta - r \sin \theta \frac{d\theta}{dt} &= -r \sin \theta + r \cos \theta (1 - r^2) \\ \frac{dr}{dt} \sin \theta + r \cos \theta \frac{d\theta}{dt} &= r \cos \theta + r \sin \theta (1 - r^2).\end{aligned}$$

Multiply the first equation by $\sin \theta$ and subtract it from $\cos \theta$ times the second

equation to obtain

$$r \frac{d\theta}{dt} = r.$$

Thus, either $r = 0$ or

$$\begin{aligned} \frac{d\theta}{dt} &= 1 \\ \theta &= t + D, \end{aligned}$$

where D is an arbitrary constant. Similarly, multiplying the first equation by $\cos \theta$ and adding it to $\sin \theta$ times the second equation yields

$$\frac{dr}{dt} = r(1 - r^2).$$

We see from this that $r = 0$ and $r = 1$ are solutions in which $dr/dt = 0$. ($r = -1$ is not a solution because by definition $r \geq 0$.) $r = 0$ yields a critical point at the origin. $r = 1$ (together with $\theta = t + D$) yields a periodic solution for which the orbit is a circle of radius 1 centered at the origin. If we exclude these solutions, we may separate variables to obtain

$$\int \frac{dr}{r(1 - r^2)} = t + c_1.$$

The left hand side may be computed by the method of partial fractions which yields

$$\ln r - \frac{1}{2} \ln |1 - r| - \frac{1}{2} \ln(1 + r) = t + c_1.$$

I did it instead using Mathematica which neglected to include the absolute values in the second term, but fortunately I remembered them. (The absolute values are not necessary for the other terms since $r, r + 1 > 0$.) This may be further simplified as follows.

$$\begin{aligned} \ln \frac{r}{\sqrt{|1 - r^2|}} &= t + c_1 \\ \frac{r}{\sqrt{|1 - r^2|}} &= c_2 e^t \\ \frac{r^2}{|1 - r^2|} &= c_3 e^{2t}. \end{aligned}$$

Note that the constant necessarily satisfies $c_3 > 0$. We now consider two cases. If $0 < r < 1$, we have

$$\begin{aligned} \frac{r^2}{1 - r^2} &= c_3 e^{2t} \\ r^2 &= c_3 e^{2t} (1 - r^2) \\ r^2 (1 + c_3 e^{2t}) &= c_3 e^{2t} \\ r^2 &= \frac{c_3 e^{2t}}{1 + c_3 e^{2t}} \end{aligned}$$

Divide both numerator and denominator by $c_3 e^{2t}$ and take the square root to obtain

$$r = \sqrt{\frac{1}{C e^{-2t} + 1}}. \quad (240)$$

If you follow what happened to the constant at each stage, you will see that the constant we end up with satisfies $C > 0$.

If $r > 1$, we may continue instead as follows.

$$\begin{aligned} \frac{r^2}{r^2 - 1} &= c_3 e^{2t} \\ r^2 &= c_3 e^{2t} (r^2 - 1) \\ r^2 (c_3 e^{2t} - 1) &= c_3 e^{2t} \\ r^2 &= \frac{c_3 e^{2t}}{c_3 e^{2t} - 1}. \end{aligned}$$

Divide numerator and denominator by $c_3 e^{2t}$ to obtain

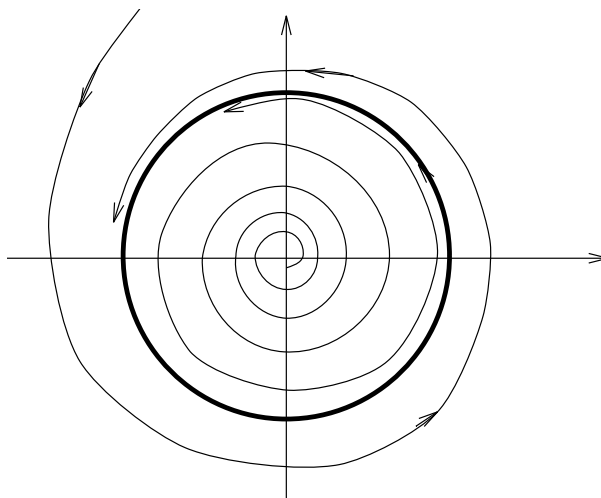
$$r = \sqrt{\frac{1}{1 - C e^{-2t}}}. \quad (241)$$

As above $C > 0$.

We may summarize all the above cases except the critical point $r = 0$ by writing

$$\begin{aligned} r &= \frac{1}{\sqrt{1 + C e^{-2t}}} \\ \theta &= t + D \end{aligned} \quad (242)$$

where $C > 0$ for $r < 1$, $C = 0$ for $r = 1$, and $C < 0$ for $r > 1$. Note that for $C > 0$, the solution spirals outward from the origin and approaches the periodic orbit $r = 1$ asymptotically as $t \rightarrow \infty$. Similarly, for $C < 0$, the solution spirals inward and approaches the periodic orbit $r = 1$ asymptotically as $t \rightarrow \infty$. All these paths behave as the Poincaré–Bendixson Theorem predicts.



You can choose for the bounded region any annular (ring shaped) region containing the circle $r = 1$. You can't allow the critical point at the origin in the bounded region because then the constant solution $\mathbf{x}(t) = 0$ would violate the conclusion of the theorem.

A periodic orbit that is asymptotically stable (i.e., all solutions which get sufficiently near it approach it asymptotically) is called an *attractor*. In the above example the orbit $r = 1$ is an attractor. On the other hand, the critical point $r = 0$ exhibits the opposite kind of behavior, so it might aptly be called a *repeller*.

Exercises for 13.2.

- Find $D\mathbf{f}$ for the functions $\mathbf{f} : \mathbf{R}^n \rightarrow \mathbf{R}^m$ given below. The function is given in a different form in each case, but you should assume it has been rewritten in standard form

$$\mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix}.$$

(a) $f_1(x, y) = x^2 + y^2, f_2(x, y) = 2xy.$

(b) $\mathbf{f}(x_1, x_2) = \begin{bmatrix} x_1 x_2 + x_2^2 \\ x_2^3 \\ x_1 + 3x_2 \end{bmatrix}$

(c) $\mathbf{f}(r, \theta) = \langle r \cos \theta, r \sin \theta \rangle.$

(d) $\mathbf{f}(x, y, z) = \frac{1}{\rho^2} \mathbf{u}_\rho.$

2. Let $\mathbf{F} : \mathbf{R}^3 \rightarrow \mathbf{R}^3$ denote a vector field on \mathbf{R}^3 . Show that the divergence of \mathbf{F} is the sum of the diagonal entries of the 3×3 derivative matrix $D\mathbf{F}$. Relate the curl of \mathbf{F} to the entries of the matrix $D\mathbf{F} - (D\mathbf{F})^t$.

Note that the sum of the diagonal entries of a square matrix A is usually called the *trace* of A .

3. For each of the following 2×2 linear systems, first solve the system, and then sketch the phase portrait in the vicinity of the origin (which is the critical point). You may have to change coordinates to get a good picture.

(a) $x'_1 = -x_2, x'_2 = 6x_1 - 5x_2$.

(b) $x'_1 = 3x_1 - 2x_2, x'_2 = 5x_1 - 3x_2$.

(c) $x'_1 = 2x_1 + x_2, x'_2 = -7x_1 - 3x_2$.

4. Sketch the phase portrait of the system in Example 252.

5. Consider the system $x'_1 = 2x_1 - x_1^2 - x_1x_2, x'_2 = 3x_2 - x_2^2 - 2x_1x_2$.

(a) Find the critical points.

(b) Find $D\mathbf{f}$ at each critical point.

(c) Solve the linear system $\frac{d\mathbf{x}}{dt} = D\mathbf{f}(\mathbf{a})\mathbf{x}$ at each critical point.

(d) Sketch the phase portraits resulting from part (c), and try to piece these together into a coherent phase portrait for the original non-linear system.

6. Repeat the above steps for the system $x'_1 = x_1^2 + x_2^2 - 2, x'_2 = x_1x_2 - 1$.

7. Consider the system $dx_1/dt = x_2, dx_2/dt = -(g/L)\sin x_1 - ax_2$ for the damped pendulum described in Section 1. Find the linear approximation at $(0, 0)$. Show that if a is a sufficiently small positive quantity, then $(0, 0)$ is a stable focus. What conclusion can you draw about the phase portrait of the nonlinear system near $(0, 0)$? Compare with the analysis in Section 1.

8. Two populations compete for the same resources and the competition is destructive to both. The following system provides a model governing their interaction.

$$\begin{aligned}\frac{dx}{dt} &= 10x - x^2 - 6xy \\ \frac{dy}{dt} &= 5y - y^2 - xy.\end{aligned}$$

Note that if $y = 0$, then x obeys a logistic law and similarly for y if $x = 0$.

(a) Find the critical points.

(b) Determine the linear system approximating the nonlinear system near each critical point.

(c) Solve each linear system at least as far as determining the eigenvalues of the coefficient matrix.

- (d) Sketch the phase portrait of each of these linear systems. You can look in the text to determine the general form of the phase portrait from the signs of the eigenvalues, but you may still have to figure out other details. You could do this by finding the general solution of each linear system in an appropriate new coordinate system, but other methods may suffice.
- (e) Check that each critical point is one of the types for which the phase portrait of the linear system is a good approximation of the phase portrait of the nonlinear system. Try to sketch the phase portrait of the nonlinear system in the first quadrant. Does your diagram tell you anything about the growth or decline of the two populations?
9. With reference to Example 7, find those values of t for which the solution $r = 1/\sqrt{1 - Ce^{-2t}}$ is defined. Assume $C > 0$. What significance does this have?

Appendices

Appendix A

Creative Commons Legal Text

Creative Commons Legal Code

Attribution-ShareAlike 3.0 Unported

CREATIVE COMMONS CORPORATION IS NOT A LAW FIRM AND DOES NOT PROVIDE LEGAL SERVICES. DISTRIBUTION OF THIS LICENSE DOES NOT CREATE AN ATTORNEY-CLIENT RELATIONSHIP. CREATIVE COMMONS PROVIDES THIS INFORMATION ON AN "AS-IS" BASIS. CREATIVE COMMONS MAKES NO WARRANTIES REGARDING THE INFORMATION PROVIDED, AND DISCLAIMS LIABILITY FOR DAMAGES RESULTING FROM ITS USE.

License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE ("CCPL" OR "LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

- a. "Adaptation" means a work based upon the Work, or upon the Work and

other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this License. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered an Adaptation for the purpose of this License.

- b. "Collection" means a collection of literary or artistic works, such as encyclopedias and anthologies, or performances, phonograms or broadcasts, or other works or subject matter other than works listed in Section 1(f) below, which, by reason of the selection and arrangement of their contents, constitute intellectual creations, in which the Work is included in its entirety in unmodified form along with one or more other contributions, each constituting separate and independent works in themselves, which together are assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined below) for the purposes of this License.
- c. "Creative Commons Compatible License" means a license that is listed at <https://creativecommons.org/compatiblelicenses> that has been approved by Creative Commons as being essentially equivalent to this License, including, at a minimum, because that license: (i) contains terms that have the same purpose, meaning and effect as the License Elements of this License; and, (ii) explicitly permits the relicensing of adaptations of works made available under that license under this License or a Creative Commons jurisdiction license with the same License Elements as this License.
- d. "Distribute" means to make available to the public the original and copies of the Work or Adaptation, as appropriate, through sale or other transfer of ownership.
- e. "License Elements" means the following high-level license attributes as selected by Licensor and indicated in the title of this License: Attribution, ShareAlike.
- f. "Licensor" means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License.
- g. "Original Author" means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i) in the case of a performance the actors, singers,

musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast.

- h. "Work" means the literary and/or artistic work offered under the terms of this License including without limitation any production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression including digital form, such as a book, pamphlet and other writing; a lecture, address, sermon or other work of the same nature; a dramatic or dramatico-musical work; a choreographic work or entertainment in dumb show; a musical composition with or without words; a cinematographic work to which are assimilated works expressed by a process analogous to cinematography; a work of drawing, painting, architecture, sculpture, engraving or lithography; a photographic work to which are assimilated works expressed by a process analogous to photography; a work of applied art; an illustration, map, plan, sketch or three-dimensional work relative to geography, topography, architecture or science; a performance; a broadcast; a phonogram; a compilation of data to the extent it is protected as a copyrightable work; or a work performed by a variety or circus performer to the extent it is not otherwise considered a literary or artistic work.
- i. "You" means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.
- j. "Publicly Perform" means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images.
- k. "Reproduce" means to make copies of the Work by any means including without limitation by sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage of a protected performance or phonogram in digital form or other electronic

medium.

2. Fair Dealing Rights. Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.

3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

- a. to Reproduce the Work, to incorporate the Work into one or more Collections, and to Reproduce the Work as incorporated in the Collections;
- b. to create and Reproduce Adaptations provided that any such Adaptation, including any translation in any medium, takes reasonable steps to clearly label, demarcate or otherwise identify that changes were made to the original Work. For example, a translation could be marked "The original work was translated from English to Spanish," or a modification could indicate "The original work has been modified.";
- c. to Distribute and Publicly Perform the Work including as incorporated in Collections; and,
- d. to Distribute and Publicly Perform Adaptations.
- e. For the avoidance of doubt:
 - i. Non-waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;
 - ii. Waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor waives the exclusive right to collect such royalties for any exercise by You of the rights granted under this License; and,
 - iii. Voluntary License Schemes. The Licensor waives the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License.

The above rights may be exercised in all media and formats whether now

known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. Subject to Section 8(f), all rights not expressly granted by Licensor are hereby reserved.

4. Restrictions. The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

- a. You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(c), as requested. If You create an Adaptation, upon notice from any Licensor You must, to the extent practicable, remove from the Adaptation any credit as required by Section 4(c), as requested.
- b. You may Distribute or Publicly Perform an Adaptation only under the terms of: (i) this License; (ii) a later version of this License with the same License Elements as this License; (iii) a Creative Commons jurisdiction license (either this or a later license version) that contains the same License Elements as this License (e.g., Attribution-ShareAlike 3.0 US)); (iv) a Creative Commons Compatible License. If you license the Adaptation under one of the licenses mentioned in (iv), you must comply with the terms of that license. If you license the Adaptation under the terms of any of the licenses mentioned in (i), (ii) or (iii) (the "Applicable License"), you must comply with the terms of the Applicable License generally and the following provisions: (I) You must include a copy of, or the URI for, the Applicable License with every copy of each Adaptation You

Distribute or Publicly Perform; (II) You may not offer or impose any terms on the Adaptation that restrict the terms of the Applicable License or the ability of the recipient of the Adaptation to exercise the rights granted to that recipient under the terms of the Applicable License; (III) You must keep intact all notices that refer to the Applicable License and to the disclaimer of warranties with every copy of the Work as included in the Adaptation You Distribute or Publicly Perform; (IV) when You Distribute or Publicly Perform the Adaptation, You may not impose any effective technological measures on the Adaptation that restrict the ability of a recipient of the Adaptation from You to exercise the rights granted to that recipient under the terms of the Applicable License. This Section 4(b) applies to the Adaptation as incorporated in a Collection, but this does not require the Collection apart from the Adaptation itself to be made subject to the terms of the Applicable License.

- c. If You Distribute, or Publicly Perform the Work or any Adaptations or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution ("Attribution Parties") in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if supplied; (iii) to the extent reasonably practicable, the URI, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and (iv) , consistent with Ssection 3(b), in the case of an Adaptation, a credit identifying the use of the Work in the Adaptation (e.g., "French translation of the Work by Original Author," or "Screenplay based on original Work by Original Author"). The credit required by this Section 4(c) may be implemented in any reasonable manner; provided, however, that in the case of a Adaptation or Collection, at a minimum such credit will appear, if a credit for all contributing authors of the Adaptation or Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate,

of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.

- d. Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Adaptations or Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author's honor or reputation. Licensor agrees that in those jurisdictions (e.g. Japan), in which any exercise of the right granted in Section 3(b) of this License (the right to make Adaptations) would be deemed to be a distortion, mutilation, modification or other derogatory action prejudicial to the Original Author's honor and reputation, the Licensor will waive or not assert, as appropriate, this Section, to the fullest extent permitted by the applicable national law, to enable You to reasonably exercise Your right under Section 3(b) of this License (right to make Adaptations) but not otherwise.

5. Representations, Warranties and Disclaimer

UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTIBILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

- a. This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Adaptations or Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full

compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

- b. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

- a. Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.
- b. Each time You Distribute or Publicly Perform an Adaptation, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.
- c. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.
- d. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.
- e. This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.
- f. The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced

according to the corresponding provisions of the implementation of those treaty provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights not granted under this License, such additional rights are deemed to be included in the License; this License is not intended to restrict the license of any rights under applicable law.

Creative Commons Notice

Creative Commons is not a party to this License, and makes no warranty whatsoever in connection with the Work. Creative Commons will not be liable to You or any party on any legal theory for any damages whatsoever, including without limitation any general, special, incidental or consequential damages arising in connection to this license. Notwithstanding the foregoing two (2) sentences, if Creative Commons has expressly identified itself as the Licensor hereunder, it shall have all rights and obligations of Licensor.

Except for the limited purpose of indicating to the public that the Work is licensed under the CCPL, Creative Commons does not authorize the use by either party of the trademark "Creative Commons" or any related trademark or logo of Creative Commons without the prior written consent of Creative Commons. Any permitted use will be in compliance with Creative Commons' then-current trademark usage guidelines, as may be published on its website or otherwise made available upon request from time to time. For the avoidance of doubt, this trademark restriction does not form part of the License.

Creative Commons may be contacted at <https://creativecommons.org/>.