Concepts in Calculus III

Beta Version

UNIVERSITY PRESS OF FLORIDA

Florida A&M University, Tallahassee
Florida Atlantic University, Boca Raton
Florida Gulf Coast University, Ft. Myers
Florida International University, Miami
Florida State University, Tallahassee
New College of Florida, Sarasota
University of Central Florida, Orlando
University of Florida, Gainesville
University of North Florida, Jacksonville
University of South Florida, Tampa
University of West Florida, Pensacola

Orange Grove Texts *Plus*

Concepts in Calculus III

Multivariable Calculus, Beta Version

Sergei Shabanov

University of Florida Department of Mathematics

University Press of Florida
Gainesville • Tallahassee • Tampa • Boca Raton
Pensacola • Orlando • Miami • Jacksonville • Ft. Myers • Sarasota

Copyright 2012 by the University of Florida Board of Trustees on behalf of the University of Florida Department of Mathematics

This work is licensed under a modified Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/. You are free to electronically copy, distribute, and transmit this work if you attribute authorship. *However, all printing rights are reserved by the University Press of Florida (http://www.upf.com). Please contact UPF for information about how to obtain copies of the work for print distribution.* You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work). For any reuse or distribution, you must make clear to others the license terms of this work. Any of the above conditions can be waived if you get permission from the University Press of Florida. Nothing in this license impairs or restricts the author's moral rights.

ISBN 978-1-61610-157-2

Orange Grove Texts *Plus* is an imprint of the University Press of Florida, which is the scholarly publishing agency for the State University System of Florida, comprising Florida A&M University, Florida Atlantic University, Florida Gulf Coast University, Florida International University, Florida State University, New College of Florida, University of Central Florida, University of Florida, University of North Florida, University of South Florida, and University of West Florida.

University Press of Florida
15 Northwest 15th Street
Gainesville, FL 32611-2079
http://www.upf.com

Florida Distance
Learning Consortium

Contents

Chapter and section numbering continues from the previous volume in the series, *Concepts in Calculus II.*

CHAPTER 11

Vectors and the Space Geometry

Our space may be viewed as a collection of points. Every geometrical figure, such as a sphere, plane, or line, is a special subset of points in space. The main purpose of an algebraic description of various objects in space is to develop a systematic representation of these objects by numbers. Interestingly enough, our experience shows that so far real numbers and basic rules of their algebra appear to be sufficient to describe all fundamental laws of nature, model everyday phenomena, and even predict many of them. The evolution of the Universe, forces binding particles in atomic nuclei, and atomic nuclei and electrons forming atoms and molecules, star and planet formation, chemistry, DNA structures, and so on, all can be formulated as relations between quantities that are measured and expressed as real numbers. Perhaps, this is the most intriguing property of the Universe, which makes mathematics the main tool of our understanding of the Universe. The deeper our understanding of nature becomes, the more sophisticated are the mathematical concepts required to formulate the laws of nature. But they remain based on real numbers. In this course, basic mathematical concepts needed to describe various phenomena in a three-dimensional Euclidean space are studied. The very fact that the space in which we live is a three-dimensional Euclidean space should not be viewed as an absolute truth. All one can say is that this *mathematical model* of the physical space is sufficient to describe a rather large set of physical phenomena in everyday life. As a matter of fact, this model fails to describe phenomena on a large scale (e.g., our galaxy). It might also fail at tiny scales, but this has yet to be verified by experiments.

71. Rectangular Coordinates in Space

The elementary object in space is a point. So the discussion should begin with the question: How can one describe a point in space by real numbers? The following procedure can be adopted. Select a particular point in space called the *origin* and usually denoted O. Set up three mutually perpendicular lines through the origin. A real number is associated with every point on each line in the following way. The origin corresponds to 0. Distances to points on one side of the line from the origin are marked by positive real numbers, while distances to points on the other half of the line are marked by negative numbers (the absolute value of a negative number is the distance). The half-lines with the grid of positive numbers will be indicated by arrows pointing from the origin to distinguish the half-lines with the grid of negative numbers. The described system of lines with the grid of real numbers on them is called a *rectangular coordinate system* at the origin O. The lines with the constructed grid of real numbers are called *coordinate axes*.

71.1. Points in Space as Ordered Triples of Real Numbers. The position of any point in space can be *uniquely* specified as an *ordered triple of real numbers* relative to a given rectangular coordinate system. Consider a rectangle whose two opposite vertices (the endpoints of the largest diagonal) are the origin and a point P, while its sides that are adjacent at the origin lie on the axes of the coordinate system. For every point P there is only one such rectangle. The rectangle is uniquely determined by its three sides adjacent at the origin. Let the number x marks the position of one such side that lies on the first axis, the numbers y and z do so for the second and third sides, respectively. Note that, depending on the position of P, the numbers x, y, and z may be negative, positive, or even 0. In other words, any point in space is associated with a unique *ordered triple* of real numbers (x, y, z) determined relative to a rectangular coordinate system. This ordered triple of numbers is called *rectangular coordinates* of a point. To reflect the order in (x, y, z), the axes of the coordinate system will be marked as x, y, and z axes. Thus, to find a point in space with rectangular coordinates $(1, 2, -3)$, one has to construct a rectangle with a vertex at the origin such that its sides adjacent at the origin occupy the intervals $[0, 1]$, $[0, 2]$, and $[-3, 0]$ along the x, y, and z axes, respectively. The point in question is the vertex opposite to the origin.

71.2. A Point as an Intersection of Coordinate Planes. The plane containing the x and y axes is called the *xy plane*. For all points in this plane, the z coordinate is 0. The condition that a point lies in the xy plane can therefore be stated as $z = 0$. The xz and yz planes can be defined similarly. The condition that a point lies in the xz or yz plane reads $y = 0$ or $x = 0$, respectively. The origin $(0, 0, 0)$ can be viewed as the intersection of three coordinate planes $x = 0$, $y = 0$, and $z = 0$. Consider all points in space whose z coordinate is fixed to a particular value $z = z_0$ (e.g., $z = 1$). They form a plane parallel to the xy plane that lies $|z_0|$ units of length above it if $z_0 > 0$ or below it if $z_0 < 0$.

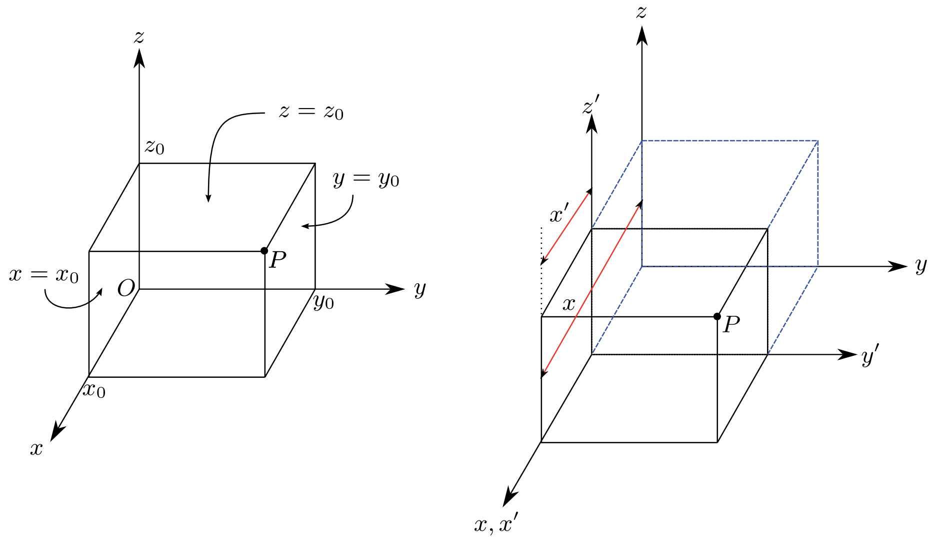

FIGURE 11.1. **Left**: Any point P in space can be viewed as the intersection of three coordinate planes $x = x_0$, $y = y_0$, and $z = z_0$; hence, P can be given an algebraic description as an ordered triple of numbers $P = (x_0, y_0, z_0)$. **Right**: Translation of the coordinate system. The origin is moved to a point (x_0, y_0, z_0) relative to the old coordinate system while the coordinate axes remain parallel to the axes of the old system. This is achieved by translating the origin first along the x axis by the distance x_0 (as shown in the figure), then along the y axis by the distance y_0, and finally along the z axis by the distance z_0. As a result, a point P that had coordinates (x, y, z) in the old system will have the coordinates $x' = x - x_0$, $y' = y - y_0$, and $z' = z - z_0$ in the new coordinate system.

A point P with coordinates (x_0, y_0, z_0) can therefore be viewed as an intersection of three *coordinate planes* $x = x_0$, $y = y_0$, and $z = z_0$ as shown in Figure 11.1. The faces of the rectangle introduced to specify the position of P relative to a rectangular coordinate system lie in the coordinate planes. The coordinate planes are perpendicular to the corresponding coordinate axes: the plane $x = x_0$ is perpendicular to the x axis, and so on.

71.3. Changing the Coordinate System. Since the origin and directions of the axes of a coordinate system can be chosen arbitrarily, the coordinates of a point depend on this choice. Suppose a point P has coordinates (x, y, z). Consider a new coordinate system whose axes are parallel to the corresponding axes of the old coordinate system, but whose origin is shifted to the point O' with coordinates $(x_0, 0, 0)$. It is straightforward to see that the point P would have the coordinates $(x - x_0, y, z)$ relative to the new coordinate system (Figure 11.1, right panel). Similarly, if the origin is shifted to a point O' with coordinates (x_0, y_0, z_0), while the axes remain parallel to the corresponding axes of the old coordinate system, then the coordinates of P are transformed as

$$(x, y, z) \longrightarrow (x - x_0, y - y_0, z - z_0)\,. \tag{11.1}$$

One can change the orientation of the coordinate axes by rotating them about the origin. The coordinates of the same point in space are different in the original and rotated rectangular coordinate systems. Algebraic relations between old and new coordinates, similar to (11.1), can be established. A simple case, when a coordinate system is rotated about one of its axes, is discussed at the end of this section.

It is important to realize that no physical or geometrical quantity should depend on the choice of a coordinate system. For example, the length of a straight line segment must be the same in any coordinate system, while the coordinates of its endpoints depend on the choice of the coordinate system. When studying a practical problem, a coordinate system can be chosen in any way convenient to describe objects in space. Algebraic rules for real numbers (coordinates) can then be used to compute physical and geometrical characteristics of the objects. The numerical values of these characteristics do not depend on the choice of the coordinate system.

71.4. Distance Between Two Points. Consider two points in space, P_1 and P_2. Let their coordinates relative to some rectangular coordinate system be (x_1, y_1, z_1) and (x_2, y_2, z_2), respectively. How can one calculate the distance between these points, or the length of a straight line segment with endpoints P_1 and P_2? The point P_1 is the intersection point of three coordinate planes $x = x_1$, $y = y_1$, and $z = z_1$. The point P_2 is the intersection point of three coordinate planes $x = x_2$, $y = y_2$, and $z = z_2$. These six planes contain faces of the rectangle whose largest diagonal is the straight line segment between the points P_1 and P_2. The question therefore is how to find the length of this diagonal.

Consider three sides of this rectangle that are adjacent, say, at the vertex P_1. The side parallel to the x axis lies between the coordinate planes $x = x_1$ and $x = x_2$ and is perpendicular to them. So the length of this side is $|x_2 - x_1|$. The absolute value is necessary as the difference $x_2 - x_1$ may be negative, depending on the values of x_1 and x_2, whereas the distance must be nonnegative. Similar arguments lead to the conclusion that the lengths of the other two adjacent sides are $|y_2 - y_1|$ and $|z_2 - z_1|$. If a rectangle has adjacent sides of length a, b, and c, then the length d of its largest diagonal satisfies the equation

$$d^2 = a^2 + b^2 + c^2 .$$

Its proof is based on the Pythagorean theorem (see Figure 11.2). Consider the rectangle face that contains the sides a and b. The length f of its diagonal is determined by the Pythagorean theorem $f^2 = a^2 + b^2$. Consider the cross section of the rectangle by the plane that contains the face diagonal f and the side c. This cross section is a rectangle with two adjacent sides c and f and the diagonal d. They are related as $d^2 = f^2 + c^2$ by the Pythagorean theorem, and the desired conclusion follows.

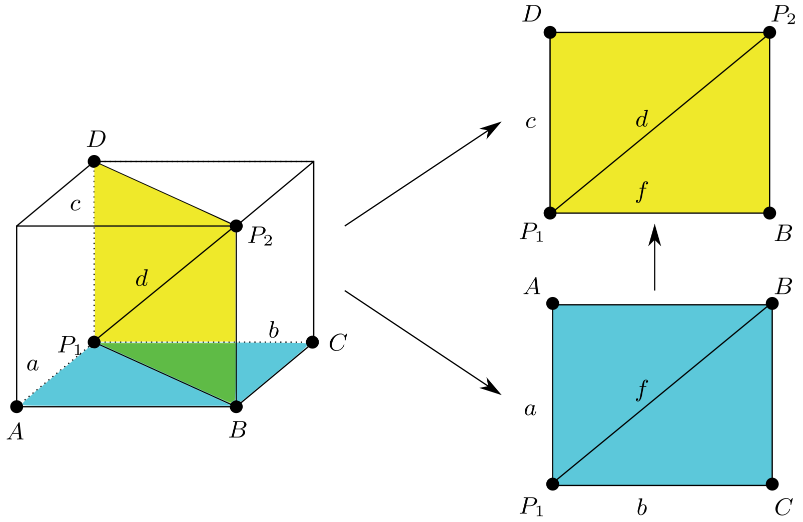

FIGURE 11.2. Distance between two points with coordinates $P_1 = (x_1, y_1, z_1)$ and $P_2 = (x_2, y_2, z_2)$. The line segment P_1P_2 is viewed as the largest diagonal of the rectangle whose faces are the coordinate planes corresponding to the coordinates of the points. Therefore, the distances between the opposite faces are $a = |x_1 - x_2|$, $b = |y_1 - y_2|$, and $c = |z_1 - z_2|$. The length of the diagonal d is obtained by the double use of the Pythagorean theorem in each of the indicated rectangles: $d^2 = c^2 + f^2$ (top right) and $f^2 = a^2 + b^2$ (bottom right).

Put $a = |x_2 - x_1|$, $b = |y_2 - y_1|$, and $c = |z_2 - z_1|$. Then $d = |P_1P_2|$ is the distance between P_1 and P_2. The distance formula is immediately found:

$$|P_1P_2| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}\,. \tag{11.2}$$

Note that the numbers (coordinates) (x_1, y_1, z_1) and (x_2, y_2, z_2) depend on the choice of the coordinate system, whereas the number $|P_1P_2|$ *remains the same* in any coordinate system! For example, if the origin of the coordinate system is translated to a point (x_0, y_0, z_0) while the orientation of the coordinate axes remains unchanged, then, according to rule (11.1), the coordinates of P_1 and P_2 relative to the new coordinate become $(x_1 - x_0, y_1 - y_0, z_1 - z_0)$ and $(x_2 - x_0, y_2 - y_0, z_2 - z_0)$, respectively. The numerical value of the distance does not change because the coordinate differences, $(x_2 - x_0) - (x_1 - x_0) = x_2 - x_1$ (similarly *for the y and z coordinates*), do not change.

Rotations in Space. The origin can always be translated to P_1 so that in the new coordinate system P_1 is $(0, 0, 0)$ and P_2 is $(x_2 - x_1, y_2 - y_1, z_2 - z_1)$. Since the distance should not depend on the orientation of the coordinate axes, any rotation can now be described algebraically as *a linear transformation of an ordered triple (x, y, z) under which the combination $x^2 + y^2 + z^2$ remains invariant.* A linear transformation means that the new coordinates are linear combinations of the old ones. It should be noted that reflections of the coordinate axes, $x \to -x$ (similarly for y and z), are linear and also preserve the distance. However, a coordinate system obtained by an odd number of reflections of the coordinate axes cannot be obtained by any rotation of the original coordinate system. So, in the above algebraic definition of a rotation, the reflections should be excluded.

71.5. Spheres in Space. In this course, relations between two equivalent descriptions of objects in space—the geometrical and the algebraic—will always be emphasized. One of the course objectives is to learn how to interpret an algebraic equation by geometrical means and how to describe geometrical objects in space algebraically. The simplest example of this kind is a sphere.

Geometrical Description of a Sphere. A sphere is a set of points in space that are equidistant from a fixed point. The fixed point is called the *sphere center*. The distance from the sphere center to any point of the sphere is called the *sphere radius*.

Algebraic Description of a Sphere. An algebraic description of a sphere implies finding an algebraic condition on coordinates (x, y, z) of points in space that belong to the sphere. So let the center of the

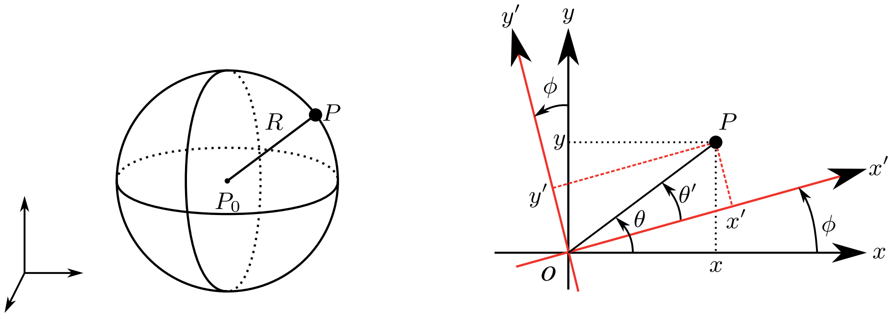

FIGURE 11.3. **Left**: A sphere is defined as a point set in space. Each point P of the set has a fixed distance R from a fixed point P_0. The point P_0 is the center of the sphere, and R is the radius of the sphere.
Right: Transformation of coordinates under a rotation of the coordinate system in a plane.

sphere be a point P_0 with coordinates (x_0, y_0, z_0) (defined relative to some rectangular coordinate system). If a point P with coordinates (x, y, z) belongs to the sphere, then the numbers (x, y, z) must be such that the distance $|PP_0|$ is the same for any such P and equal to the sphere radius, denoted R, that is, $|PP_0| = R$ or $|PP_0|^2 = R^2$ (see Figure 11.3, left panel). Using the distance formula, this condition can be written as

$$(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 = R^2 \,. \tag{11.3}$$

For example, the set of points with coordinates (x, y, z) that satisfy the condition $x^2 + y^2 + z^2 = 4$ is a sphere of radius $R = 2$ centered at the origin $x_0 = y_0 = z_0 = 0$.

71.6. Algebraic Description of Point Sets in Space. The idea of an algebraic description of a sphere can be extended to other sets in space. It is convenient to introduce some brief notation for an algebraic description of sets. For example, for a set $\mathcal{S}$ of points in space with coordinates (x, y, z) such that they satisfy the algebraic condition (11.3), one writes

$$\mathcal{S} = \left\{ (x, y, z) \;\middle|\; (x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 = R^2 \right\}.$$

This relation means that the set $\mathcal{S}$ is a collection of all points (x, y, z) such that (the vertical bar) their rectangular coordinates satisfy (11.3).

Similarly, the xy plane can be viewed as a set of points whose z coordinates vanish:

$$\mathcal{P} = \Big\{ (x, y, z) \;\Big|\; z = 0 \Big\}\,.$$

The solid region in space that consists of points whose coordinates are non negative is called the *first octant*:

$$\mathcal{O}_1 = \Big\{ (x, y, z) \;\Big|\; x \geq 0,\ y \geq 0,\ z \geq 0 \Big\}\,.$$

The spatial region

$$\mathcal{B} = \Big\{ (x, y, z) \;\Big|\; x > 0,\ \ y > 0,\ \ z > 0,\ \ x^2 + y^2 + z^2 < 4 \Big\}$$

is the collection of all points in the portion of a ball of radius 2 that lies in the first octant. The strict inequalities imply that the boundary of this portion of the ball does not belong to the set $\mathcal{B}$.

71.7. Study Problems.

Problem 11.1. *Show that the coordinates of the midpoint of a straight line segment are*

$$\left(\frac{x_1 + x_2}{2},\ \frac{y_1 + y_2}{2},\ \frac{z_1 + z_2}{2} \right)$$

if the coordinates of its endpoints are (x_1, y_1, z_1) and (x_2, y_2, z_2).

Solution: Let P_1 and P_2 be the endpoints and let M be the midpoint. One has to verify the condition $|P_2M| = |MP_1|$ or $|P_2M|^2 = |MP_1|^2$ by means of the distance formula. The x-coordinate differences for the segments P_2M and MP_1 read $x_2 - (x_1 + x_2)/2 = (x_2 - x_1)/2$ and $(x_1 + x_2)/2 - x_1 = (x_2 - x_1)/2$, respectively; that is, they coincide. Similarly, the differences of the corresponding y and z coordinates are the same. By the distance formula, it is then concluded that $|P_2M|^2 = |MP_1|^2$. □

Problem 11.2. *Let (x, y, z) be coordinates of a point P. Consider a new coordinate system that is obtained by rotating the x and y axes about the z axis counterclockwise as viewed from the top of the z axis through an angle ϕ. Let (x', y', z') be coordinates of P in the new coordinate system. Find the relations between the old and new coordinates.*

Solution: The height of P relative to the xy plane does not change upon rotation. So $z' = z$. It is therefore sufficient to consider rotations in the xy plane, that is, for points P with coordinates $(x, y, 0)$. Let $r = |OP|$ (the distance between the origin and P) and let θ be the angle counted from the positive x axis toward the ray OP counterclockwise (see Figure 11.3, right panel). Then $x = r\cos\theta$ and $y = r\sin\theta$ (the polar coordinates of P). In the new coordinate system, the angle between the positive x' axis and the ray OP becomes $\theta' = \theta - \phi$. Therefore,

$$\begin{aligned} x' = r\cos\theta' = \cos(\theta - \phi) &= r\cos\theta\cos\phi + r\sin\theta\sin\phi \\ &= x\cos\phi + y\sin\phi\,, \\ y' = r\sin\theta' = r\sin(\theta - \phi) &= r\sin\theta\cos\phi - r\cos\theta\sin\phi \\ &= y\cos\phi - x\sin\phi\,. \end{aligned}$$

□

Problem 11.3. *Give a geometrical description of the set*

$$\mathcal{S} = \left\{(x, y, z) \;\middle|\; x^2 + y^2 + z^2 - 4z = 0\right\}\,.$$

Solution: The condition on the coordinates of points that belong to the set contains the sum of squares of the coordinates just like the equation of a sphere. The difference is that (11.3) contains the sum of perfect squares. So the squares must be completed in the above equation and the resulting expression compared with (11.3). One has $z^2-4z = (z-2)^2-4$ so that the condition becomes $x^2+y^2+(z-2)^2 = 4$. It describes a sphere of radius $R = 2$ that is centered at the point $(x_0, y_0, z_0) = (0, 0, 2)$; that is, the center of the sphere is on the z axis at a distance of 2 units above the xy plane. □

Problem 11.4. *Give a geometrical description of the set*

$$\mathcal{C} = \left\{(x, y, z) \;\middle|\; x^2 + y^2 - 2x - 4y = 4\right\}\,.$$

Solution: As in the previous problem, the condition can be written as the sum of perfect squares $(x - 1)^2 + (y - 2)^2 = 9$ by means the of relations $x^2 - 2x = (x - 1)^2 - 1$ and $y^2 - 4y = (y - 2)^2 - 4$. In the xy plane, this is nothing but the equation of a circle of radius 3 whose center is the point $(1, 2, 0)$. In any plane $z = z_0$ parallel to the xy plane, the x and y coordinates satisfy the same equation, and hence the corresponding points also form a circle of radius 3 with the center $(1, 2, z_0)$. Thus, the set is a cylinder of radius 3 whose axis is parallel to the z axis and passes through the point $(1, 2, 0)$. □

Problem 11.5. *Give a geometrical description of the set*

$$\mathcal{P} = \left\{(x, y, z) \;\middle|\; z(y - x) = 0\right\}\,.$$

SOLUTION: The condition is satisfied if either $z = 0$ or $y = x$. The former equation describes the xy plane, while the latter represents a line in the xy plane. Since it does not impose any restriction on the z coordinate, each point of the line can be moved up and down parallel to the z axis. The resulting set is a plane that contains the line $y = x$ in the xy plane and the z axis. Thus, the set $\mathcal{P}$ is the union of this plane and the xy plane. □

71.8. Exercises. **(1)** Find the distance between the following specified points:

(i) $A(1, 2, 3)$ and $B(-1, 0, 2)$
(ii) $A(-1, 3, -2)$ and $B(-1, 2, -1)$

(2) Let the set $\mathcal{S}$ consist of points $(t, 2t, 3t)$ where $-\infty < t < \infty$. Find the point of $\mathcal{S}$ that is the closest to the point $A(3, 2, 1)$. Describe the set $\mathcal{S}$ geometrically.

(3) Give a geometrical description of the following sets defined algebraically and sketch them:

(i) $x^2 + y^2 + z^2 - 2x + 4y - 6z = 0$
(ii) $x^2 + y^2 + z^2 \geq 4$
(iii) $x^2 + y^2 + z^2 \leq 4$, $z > 0$
(iv) $x^2 + y^2 - 4y < 0$, $z > 0$
(v) $4 \leq x^2 + y^2 + z^2 \leq 9$
(vi) $x^2 + y^2 \geq 1$, $x^2 + y^2 + z^2 \leq 4$
(vii) $x^2 + y^2 + z^2 - 2z < 0$, $z > 1$
(viii) $x^2 + y^2 + z^2 - 2z = 0$, $z = 1$
(ix) $(x - a)(y - b)(z - c) = 0$

(4) Sketch each of the following sets and give their algebraic description:

(i) A sphere whose diameter is the straight line segment AB, where $A = (1, 2, 3)$ and $B = (3, 2, 1)$.
(ii) A sphere centered at $(1, 2, 3)$ that lies in the first octant and touches one of the coordinate planes.
(iii) The largest solid cube that is contained in a ball of radius R centered at the origin. Solve the same problem if the ball is not centered at the origin.
(iv) The solid region that is a ball of radius R that has a cylindrical hole of radius $R/2$ whose axis is at a distance of $R/2$ from the center of the ball. Choose a convenient coordinate system.

(v) The portion of a ball of radius R that lies between two parallel planes each of which is at s distance of $a < R$ from the center of the ball. Choose a convenient coordinate system.

72. Vectors in Space

72.1. Oriented Segments and Vectors. Suppose there is a point like object moving in space with a constant rate, say, 5 m/s. If the object was initially at a point P_1, and in 1 second it arrives at a point P_2, then the distance traveled is $|P_1P_2| = 5$ m. The rate (or speed) 5 m/s does not provide a complete description of the motion of the object in space because it only answers the question "How fast does the object move?" but it does not say anything about "Where to does the object move?" Since the initial and final positions of the object are known, both questions can be answered, if one associates an *oriented segment* $\vec{P_1P_2}$ with the moving object. The arrow specifies the direction, "from P_1 to P_2," and the length $|P_1P_2|$ defines the rate (speed) at which the object moves. So, for every moving object, one can assign an oriented segment whose length equals its speed and whose direction coincides with the direction of motion. This oriented segment is called a *velocity*. The concept of velocity as an oriented segment still has a drawback. Indeed, consider two objects moving parallel with the same speed. The oriented segments corresponding to the velocities of the objects have the same length and the same direction, but they are still different because their initial points do not coincide. On the other hand, the velocity should describe a particular physical property of the motion itself ("how fast and where to"), and therefore the spatial position where the motion occurs should not matter. This observation leads to the concept of a *vector* as an abstract mathematical object that *represents all oriented segments that are parallel and have the same length.* If the velocity is a vector, then two objects have the same velocity if they move parallel with the same rate. The concept of velocity as a vector no longer has the aforementioned drawback.

Vectors will be denoted by boldface letters. *Two oriented segments $\vec{AB}$ and $\vec{CD}$ represent the same vector* $\mathbf{a}$ *if they are parallel and $|AB| = |CD|$; that is, they can be obtained from one another by transporting them parallel to themselves in space.* A representation of an abstract vector by a particular oriented segment is denoted by the equality $\mathbf{a} = \vec{AB}$ or $\mathbf{a} = \vec{CD}$. The fact that the oriented segments $\vec{AB}$ and $\vec{CD}$ represent the same vector is denoted by the equality $\vec{AB} = \vec{CD}$.

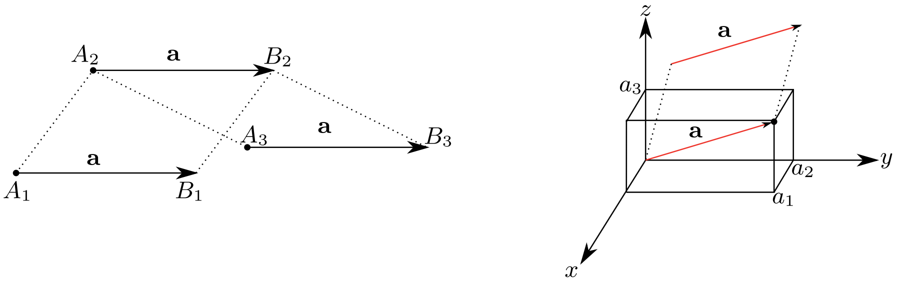

FIGURE 11.4. **Left**: Oriented segments obtained from one another by parallel transport. They all represent the same vector.
Right: A vector as an ordered triple of numbers. An oriented segment is transported parallel so that its initial point coincides with the origin of a rectangular coordinate system. The coordinates of the terminal point of the transported segment, (a_1, a_2, a_3), are components of the corresponding vector. So a vector can always be written as an ordered triple of numbers: $\mathbf{a} = \langle a_1, a_2, a_3\rangle$. By construction, the components of a vector depend on the choice of the coordinate system (the orientation of the coordinate axes in space).

72.2. Vector as an Ordered Triple of Numbers. Here an algebraic representation of vectors in space will be introduced. Consider an oriented segment $\vec{AB}$ that represents a vector $\mathbf{a}$ (i.e., $\mathbf{a} = \vec{AB}$). An oriented segment $\vec{A'B'}$ represents the same vector if it is obtained by transporting $\vec{AB}$ parallel to itself. In particular, let us take $A' = O$, where O is the origin of some rectangular coordinate system. Then $\mathbf{a} = \vec{AB} = \vec{OB'}$. The direction and length of the oriented segment $\vec{OB'}$ is uniquely determined by the coordinates of the point B'. Thus, we have the following algebraic definition of a vector.

DEFINITION 11.1. (Vectors).
A vector in space is an ordered triple of real numbers:

$$\mathbf{a} = \langle a_1\, , a_2\, , a_3\rangle\, .$$

The numbers a_1, a_2, and a_3 are called components *of the vector* $\mathbf{a}$.

Note that the numerical values of the components depend on the choice of coordinate system. From a geometrical point of view, the ordered triple (a_1, a_2, a_3) is the coordinates of the point B', that is, the endpoint of the oriented segment that represents $\mathbf{a}$ if the initial point coincides with the origin.

DEFINITION 11.2. (Equality of Two Vectors).
Two vectors $\mathbf{a}$ *and* $\mathbf{b}$ *are equal or coincide if their corresponding components are equal:*

$$\mathbf{a} = \mathbf{b} \quad \Longleftrightarrow \quad a_1 = b_1,\ a_2 = b_2,\ a_3 = b_3\,.$$

This definition agrees with the geometrical definition of a vector as a class of all oriented segments that are parallel and have the same length. Indeed, if two oriented segments represent the same vector, then, after parallel transport such that their initial points coincide with the origin, their final points coincide too and hence have the same coordinates.

EXAMPLE 11.1. *Find the components of a vector* $\vec{P_1P_2}$ *if the coordinates of* P_1 *and* P_2 *are* (x_1, y_1, z_1) *and* (x_2, y_2, z_2)*, respectively.*

SOLUTION: Consider a rectangle whose largest diagonal coincides with the segment P_1P_2 and whose sides are parallel to the coordinate axes. After parallel transport of the segment so that P_1 moves to the origin, the coordinates of the other endpoint are the components of $\vec{P_1P_2}$. Alternatively, the origin of the coordinate system can be moved to the point P_1, keeping the directions of the coordinate axes. Therefore,

$$\vec{P_1P_2} = \langle x_2 - x_1,\ y_2 - y_1,\ z_2 - z_1\rangle,$$

according to the coordinate transformation law (11.1), where $P_0 = P_1$. Thus, in order to find the components of the vector $\vec{P_1P_2}$ from the coordinates of its points, one has to subtract the coordinates of the initial point P_1 from the corresponding components of the final point P_2. □

DEFINITION 11.3. (Norm of a Vector). *The number*

$$\|\mathbf{a}\| = \sqrt{a_1^2 + a_2^2 + a_3^2}$$

is called the norm *of a vector* $\mathbf{a}$*.*

By Example 11.1 and the distance formula (11.2), the norm of a vector is the length of any oriented segment representing the vector. The norm of a vector is also called the *magnitude* or *length* of a vector.

DEFINITION 11.4. (Zero Vector).
A vector with vanishing components, $\mathbf{0} = \langle 0, 0, 0\rangle$*, is called a* zero vector*.*

A vector $\mathbf{a}$ is a zero vector if and only if its norm vanishes, $\|\mathbf{a}\| = 0$. Indeed, if $\mathbf{a} = \mathbf{0}$, then $a_1 = a_2 = a_3 = 0$ and hence $\|\mathbf{a}\| = 0$. For the converse, it follows from the condition $\|\mathbf{a}\| = 0$ that $a_1^2 + a_2^2 + a_3^2 = 0$, which is only possible if $a_1 = a_2 = a_3 = 0$, or $\mathbf{a} = \mathbf{0}$. Recall that an "if and only if" statement actually implies two statements. First, if $\mathbf{a} = \mathbf{0}$, then $\|\mathbf{a}\| = 0$ (the direct statement). Second, if $\|\mathbf{a}\| = 0$, then $\mathbf{a} = \mathbf{0}$ (the converse statement).

72.3. Vector Algebra. Continuing the analogy between the vectors and velocities of a moving object, consider two objects moving parallel but with different rates (speeds). Their velocities as vectors are parallel, but they have different magnitudes. What is the relation between the components of such vectors? Take a vector $\mathbf{a} = \langle a_1, a_2, a_3\rangle$. It can be viewed as the largest diagonal of a rectangle with one vertex at the origin and the opposite vertex at coordinates (a_1, a_2, a_3). The adjacent sides of the rectangle have lengths given by the corresponding components of $\mathbf{a}$ (modulo the signs if they happen to be negative). The direction of the diagonal does not change if the sides of the rectangle are scaled by the same factor, while the length of the diagonal is scaled

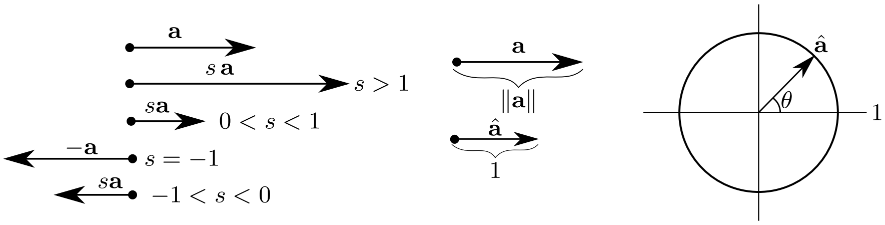

FIGURE 11.5. **Left**: Multiplication of a vector $\mathbf{a}$ by a number s. If $s > 0$, the result of the multiplication is a vector parallel to $\mathbf{a}$ whose length is scaled by the factor s. If $s < 0$, then $s\mathbf{a}$ is a vector whose direction is the opposite to that of $\mathbf{a}$ and whose length is scaled by $|s|$.
Middle: Construction of a unit vector parallel to $\mathbf{a}$. The unit vector $\hat{\mathbf{a}}$ is a vector parallel to $\mathbf{a}$ whose length is 1. Therefore, it is obtained from $\mathbf{a}$ by dividing the latter by its length $\|\mathbf{a}\|$, i.e., $\hat{\mathbf{a}} = s\mathbf{a}$, where $s = 1/\|\mathbf{a}\|$.
Right: A unit vector in a plane can always be viewed as an oriented segment whose initial point is at the origin of a coordinate system and whose terminal point lies on the circle of unit radius centered at the origin. If θ is the polar angle in the plane, then $\hat{\mathbf{a}} = \langle\cos\theta, \sin\theta, 0\rangle$.

by this factor. This geometrical observation leads to the following algebraic rule.

DEFINITION 11.5. (Multiplication of a Vector by a Number). *A vector* $\mathbf{a}$ *multiplied by a number* s *is a vector whose components are multiplied by* s*:*

$$s\mathbf{a} = \langle sa_1,\ sa_2,\ sa_3\rangle.$$

If $s > 0$, then the vector $s\mathbf{a}$ has the same direction as $\mathbf{a}$. If $s < 0$, then the vector $s\mathbf{a}$ has the direction opposite to $\mathbf{a}$. For example, the vector $-\mathbf{a}$ has the same magnitude as $\mathbf{a}$ but points in the direction opposite to $\mathbf{a}$. The magnitude of $s\mathbf{a}$ reads:

$$\|s\mathbf{a}\| = \sqrt{(sa_1)^2 + (sa_2)^2 + (sa_3)^2} = \sqrt{s^2}\,\sqrt{a_1^2 + a_2^2 + a_3^2} = |s|\,\|\mathbf{a}\|\,;$$

that is, when a vector is multiplied by a number, its magnitude changes by the factor $|s|$. The geometrical analysis of the multiplication of a vector by a number leads to the following simple algebraic criterion for two vectors being parallel.

THEOREM 11.1. *Two nonzero vectors are parallel if they are proportional:*

$$\mathbf{a} \parallel \mathbf{b} \quad \Longleftrightarrow \quad \mathbf{a} = s\mathbf{b}$$

for some real s*.*

If all the components of the vectors in question do not vanish, then this criterion may also be written as

$$\mathbf{a} \parallel \mathbf{b} \quad \Longleftrightarrow \quad s = \frac{a_1}{b_1} = \frac{a_2}{b_2} = \frac{a_3}{b_3}\,,$$

which is easy to verify. If, say, $b_1 = 0$, then $\mathbf{b}$ is parallel to $\mathbf{a}$ when $a_1 = b_1 = 0$ and $a_2/b_2 = a_3/b_3$.

DEFINITION 11.6. (Unit Vector). *A vector* $\hat{\mathbf{a}}$ *is called a* unit vector *if its norm equals 1,* $\|\hat{\mathbf{a}}\| = 1$.

Any nonzero vector $\mathbf{a}$ can be turned into a unit vector $\hat{\mathbf{a}}$ that is parallel to $\mathbf{a}$. The norm (length) of the vector $s\mathbf{a}$ reads $\|s\mathbf{a}\| = |s|\|\mathbf{a}\| = s\|\mathbf{a}\|$ if $s > 0$. So, by choosing $s = 1/\|\mathbf{a}\|$, the unit vector parallel to $\mathbf{a}$ is obtained:

$$\hat{\mathbf{a}} = \frac{1}{\|\mathbf{a}\|}\,\mathbf{a} = \left\langle \frac{a_1}{\|\mathbf{a}\|},\ \frac{a_2}{\|\mathbf{a}\|},\ \frac{a_3}{\|\mathbf{a}\|} \right\rangle\,.$$

For example, owing to the trigonometric identity, $\cos^2\theta + \sin^2\theta = 1$, any unit vector in the xy plane can always be written in the form $\hat{\mathbf{a}} = \langle\cos\theta, \sin\theta, 0\rangle$, where θ is the angle counted from the positive x axis toward the vector $\mathbf{a}$ counterclockwise. Note that, in many practical applications, the components of a vector often have dimensions. For instance, the components of a displacement vector are measured in units of length (meters, inches, etc.), the components of a velocity vector are measured in, for example, meters per second, and so on. The magnitude of a vector **a** has the same dimension as its components. Therefore, the corresponding unit vector $\hat{\mathbf{a}}$ is dimensionless. It specifies only the direction of a vector **a**.

72.3.1. The Parallelogram Rule. Suppose a person is walking on the deck of a ship with speed v m/s. In 1 second, the person goes a distance v from point A to B of the deck. The velocity vector relative to the deck is $\mathbf{v} = \vec{AB}$ and $\|\mathbf{v}\| = |AB| = v$ (the speed). The ship moves relative to the water so that in 1 second it comes to a point D from a point C on the surface of the water. The ship's velocity vector relative to the water is then $\mathbf{u} = \vec{CD}$ with magnitude $u = \|\mathbf{u}\| = |CD|$. What is the velocity vector of the person relative to the water? Suppose the point A on the deck coincides with the point C on the surface of the water. Then the velocity vector is the displacement vector of the person relative to the water in 1 second. As the person walks on the deck along the segment AB, this segment travels the distance u parallel to itself along the vector **u** relative to the water. In 1 second, the point B of the deck is moved to a point B' on the surface of the water so that the displacement vector of the person relative to the water will be $\vec{AB'}$. Apparently, the displacement vector $\vec{BB'}$ coincides with the ship's velocity **u** because B travels the distance u parallel to **u**. This suggests a simple geometrical rule for finding $\vec{AB'}$ as shown in Figure 11.6. Take the vector $\vec{AB} = \mathbf{v}$, place the vector **u** so that its initial point coincides with B, and make the oriented segment with the initial point of **v** and the final point of **u** in this diagram. The resulting vector is the displacement vector of the person relative to the surface of the water in 1 second and hence defines the velocity of the person relative to the water. This geometrical procedure is called *addition of vectors*.

Consider a parallelogram whose adjacent sides, the vectors **a** and **b**, extend from the vertex of the parallelogram. The sum of the vectors **a** and **b** is a vector, denoted $\mathbf{a} + \mathbf{b}$, that is the diagonal of the parallelogram extended from the same vertex. Note that the parallel sides of the parallelogram represent the same vector (they are parallel and have the same length). This geometrical rule for adding vectors is called the *parallelogram rule*. It follows from the parallelogram rule that the addition of vectors is *commutative*:

$$\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a};$$

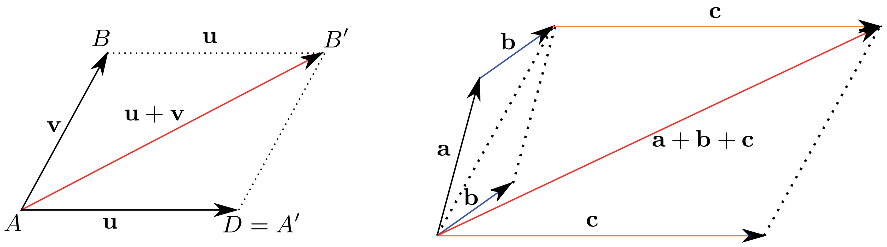

FIGURE 11.6. **Left**: Parallelogram rule for adding two vectors. If two vectors form adjacent sides of a parallelogram at a vertex A, then the sum of the vectors is a vector that coincides with the diagonal of the parallelogram and originates at the vertex A.
Right: Adding several vectors by using the parallelogram rule. Given the first vector in the sum, all other vectors are transported parallel so that the initial point of the next vector in the sum coincides with the terminal point of the previous one. The sum is the vector that originates from the initial point of the first vector and terminates at the terminal point of the last vector. It does not depend on the order of vectors in the sum.

that is, the order in which the vectors are added does not matter. To add several vectors (e.g., $\mathbf{a}+\mathbf{b}+\mathbf{c}$), one can first find $\mathbf{a}+\mathbf{b}$ by the parallelogram rule and then add $\mathbf{c}$ to the vector $\mathbf{a}+\mathbf{b}$. Alternatively, the vectors $\mathbf{b}$ and $\mathbf{c}$ can be added first, and then the vector $\mathbf{a}$ can be added to $\mathbf{b}+\mathbf{c}$. According to the parallelogram rule, the resulting vector is the same:

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

This means that the addition of vectors is *associative.* So several vectors can be added in any order. Take the first vector, then move the second vector parallel to itself so that its initial point coincides with the final point of the first vector. The third vector is moved parallel so that its initial point coincides with the final point of the second vector, and so on. Finally, make a vector whose initial point coincides with the initial point of the first vector and whose final point coincides with the final point of the last vector in the sum. To visualize this process, imagine a man walking along the first vector, then going parallel to the second vector, then parallel to the third vector, and so on. The endpoint of his walk is independent of the order in which he chooses the vectors.

72.3.2. Algebraic Addition of Vectors.

Definition 11.7. *The sum of two vectors* $\mathbf{a} = \langle a_1, a_2, a_3\rangle$ *and* $\mathbf{b} = \langle b_1, b_2, b_3\rangle$ *is a vector whose components are the sums of the corresponding components of* $\mathbf{a}$ *and* $\mathbf{b}$*:*

$$\mathbf{a} + \mathbf{b} = \langle a_1 + b_1,\ a_2 + b_2,\ a_3 + b_3\rangle.$$

This definition is equivalent to the geometrical definition of adding vectors, that is, the parallelogram rule that has been motivated by studying the velocity of a combined motion. Indeed, put $\mathbf{a} = \vec{OA}$, where the endpoint A has the coordinates (a_1, a_2, a_3). A vector $\mathbf{b}$ represents all parallel segments of the same length $\|\mathbf{b}\|$. In particular, $\mathbf{b}$ is one such oriented segment whose initial point coincides with A. Suppose that $\mathbf{a}+\mathbf{b} = \vec{OC} = \langle c_1, c_2, c_3\rangle$, where C has coordinates (c_1, c_2, c_3). By the parallelogram rule, $\mathbf{b} = \vec{AC} = \langle c_1 - a_1, c_2 - a_2,\, c_3 - a_3\rangle$, where the relation between the components of a vector and the coordinates of its endpoints has been used. The equality of two vectors means the equality of the corresponding components, that is, $b_1 = c_1 - a_1$, $b_2 = c_2 - a_2$, and $b_3 = c_3 - a_3$, or $c_1 = a_1 + b_1$, $c_2 = a_2 + b_2$, and $c_3 = a_3 + b_3$ as required by the algebraic addition of vectors.

72.3.3. Rules of Vector Algebra. Combining addition of vectors with multiplication by real numbers, the following simple rule can be established by either geometrical or algebraic means:

$$s(\mathbf{a} + \mathbf{b}) = s\mathbf{a} + s\mathbf{b}\,, \qquad (s + t)\mathbf{a} = s\mathbf{a} + t\mathbf{a}\,.$$

The difference of two vectors can be defined as $\mathbf{a} - \mathbf{b} = \mathbf{a} + (-1)\mathbf{b}$. In the parallelogram with adjacent sides $\mathbf{a}$ and $\mathbf{b}$, the sum of vectors $\mathbf{a}$ and $(-1)\mathbf{b}$ represents the vector that originates from the endpoint of $\mathbf{b}$ and ends at the endpoint of $\mathbf{a}$ because $\mathbf{b} + [\mathbf{a} + (-1)\mathbf{b}] = \mathbf{a}$ in accordance with the geometrical rule for adding vectors; that is $\mathbf{a} \pm \mathbf{b}$ are two diagonals of the parallelogram. The procedure is illustrated in Figure 11.7 (left panel).

72.4. Study Problems.

Problem 11.6. *Consider two nonparallel vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *in a plane. Show that any vector* $\mathbf{c}$ *in this plane can be written as a linear combination* $\mathbf{c} = t\mathbf{a} + s\mathbf{b}$ *for some real* t *and* s*.*

Solution: By parallel transport, the vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ can be moved so that their initial points coincide. The vectors $t\mathbf{a}$ and $s\mathbf{b}$ are parallel to $\mathbf{a}$ and $\mathbf{b}$, respectively, for all values of s and t. Consider the lines $\mathcal{L}_a$ and $\mathcal{L}_b$ that contain the vectors $\mathbf{a}$ and $\mathbf{b}$, respectively. Construct

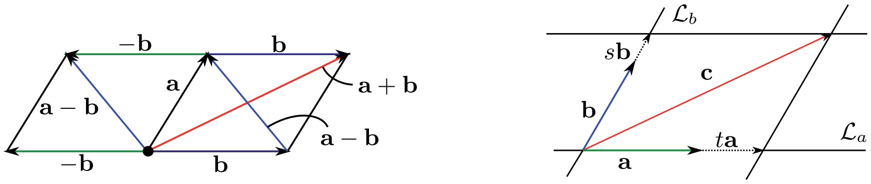

FIGURE 11.7. **Left**: Subtraction of two vectors. The difference $\mathbf{a}-\mathbf{b}$ is viewed as the sum of $\mathbf{a}$ and $-\mathbf{b}$, the vector that has the direction opposite to $\mathbf{b}$ and the same length as $\mathbf{b}$. The parallelogram rule for adding $\mathbf{a}$ and $-\mathbf{b}$ shows that the difference $\mathbf{a}-\mathbf{b}=\mathbf{a}+(-\mathbf{b})$ is the vector that originates from the terminal point of $\mathbf{b}$ and ends at the terminal of $\mathbf{a}$ if $\mathbf{a}$ and $\mathbf{b}$ are adjacent sides of a parallelogram; that is, the sum $\mathbf{a}+\mathbf{b}$ and the difference $\mathbf{a}-\mathbf{b}$ are the two diagonals of the parallelogram.
Right: Illustration to Study Problem 11.6. Any vector in a plane can always be represented as a linear combination of two nonparallel vectors.

two lines through the end point of $\mathbf{c}$; one is parallel to $\mathcal{L}_a$ and the other to $\mathcal{L}_b$ as shown in Figure 11.7 (right panel). The intersection points of these lines with $\mathcal{L}_a$ and $\mathcal{L}_b$ and the initial and final points of $\mathbf{c}$ form the vertices of the parallelogram whose diagonal is $\mathbf{c}$ and whose adjacent sides are parallel to $\mathbf{a}$ and $\mathbf{b}$. Therefore, $\mathbf{a}$ and $\mathbf{b}$ can always be scaled so that $t\mathbf{a}$ and $s\mathbf{b}$ become the adjacent sides of the constructed parallelogram. For a given $\mathbf{c}$, the reals t and s are uniquely defined by the proposed geometrical construction. By the parallelogram rule, $\mathbf{c}=t\mathbf{a}+s\mathbf{b}$. □

Problem 11.7. *Find the coordinates of a point B that is at a distance of 6 units of length from the point $A(1,-1,2)$ in the direction of the vector $\mathbf{v}=\langle 2,1,-2\rangle$.*

SOLUTION: The position vector of the point A is $\mathbf{a}=\vec{OA}=\langle 1,-1,2\rangle$. The position vector of the point B is $\mathbf{b}=\mathbf{a}+s\mathbf{v}$, where s is a positive number to be chosen such that the length $|AB|=s\|\mathbf{v}\|$ equals 6. Since $\|\mathbf{v}\|=3$, one finds $s=2$. Therefore, $\mathbf{b}=\langle 1,-1,2\rangle+2\langle 2,1,-2\rangle=\langle 5,1,-2\rangle$. □

Problem 11.8. *Consider a straight line segment with the endpoints $A(1,2,3)$ and $B(-2,-1,0)$. Find the coordinates of the point C on the segment such that it is twice as far from A as it is from B.*

Solution: Let $\mathbf{a} = \langle 1, 2, 3\rangle$, $\mathbf{b} = \langle -1, 0, 1\rangle$, and $\mathbf{c}$ be position vectors of A, B, and C, respectively. The question is to express $\mathbf{c}$ via $\mathbf{a}$ and $\mathbf{b}$. One has $\mathbf{c} = \mathbf{a} + \vec{AC}$. The vector $\vec{AC}$ is parallel to $\vec{AB} = \langle -3, -3, -3\rangle$ and hence $\vec{AC} = s\vec{AB}$. Since $|AC| = 2|CB|$, $|AC| = \frac{2}{3}|AB|$ and therefore $s = \frac{2}{3}$. Thus, $\mathbf{c} = \mathbf{a} + \frac{2}{3}\vec{AB} = \mathbf{a} + \frac{2}{3}(\mathbf{b} - \mathbf{a}) = \langle -1, 0, 1\rangle$. □

Problem 11.9. *In Study Problem 11.6, let $\|\mathbf{a}\| = 1$, $\|\mathbf{b}\| = 2$, and the angle between $\mathbf{a}$ and $\mathbf{b}$ be $2\pi/3$. Find the coefficients s and t if the vector $\mathbf{c}$ has a norm of 6 and bisects the angle between $\mathbf{a}$ and $\mathbf{b}$.*

Solution: It follows from the solution of Study Problem 11.6 that the numbers s and t do not depend on the coordinate system relative to which the components of all the vectors are defined. So choose the coordinate system so that $\mathbf{a}$ is parallel to the x axis and $\mathbf{b}$ lies in the xy plane. With this choice, $\mathbf{a} = \langle 1, 0, 0\rangle$ and $\mathbf{b} = \langle \|\mathbf{b}\|\cos(2\pi/3), \|\mathbf{b}\|\sin(2\pi/3), 0\rangle = \langle -1, \sqrt{3}, 0\rangle$. Similarly, $\mathbf{c}$ is the vector of length $\|\mathbf{c}\| = 6$ that makes the angle $\pi/3$ with the x axis, and therefore $\mathbf{c} = \langle 3, 3\sqrt{3}, 0\rangle$. Equating the corresponding components in the relation $\mathbf{c} = t\mathbf{a} + s\mathbf{b}$, one finds $3 = t - s$ and $3\sqrt{3} = s\sqrt{3}$, or $s = 3$ and $t = 6$. Hence, $\mathbf{c} = 6\mathbf{a} + 3\mathbf{b}$. □

Problem 11.10. *Suppose the three coordinate planes are all mirrored. A light ray strikes the mirrors. Determine the direction in which the reflected ray will go.*

Solution: Let $\mathbf{u}$ be a vector parallel to the incident ray. Under a reflection from a plane mirror, the component of $\mathbf{u}$ perpendicular to the plane changes its sign. Therefore, after three consecutive reflections from each coordinate plane, all three components of $\mathbf{u}$ change their signs, and the reflected ray will go parallel to the incident ray but in the exact opposite direction. For example, suppose the ray is reflected first by the xz plane, then by the yz plane, and finally by the xy plane. In this case, $\mathbf{u} = \langle u_1, u_2, u_3\rangle \to \langle u_1, -u_2, u_3\rangle \to \langle -u_1, -u_2, u_3\rangle \to \langle -u_1, -u_2, -u_3\rangle = -\mathbf{u}$. □

Remark. This principle is used to design reflectors like the cat's-eyes on bicycles and those that mark the border lines of a road. No matter from which direction such a reflector is illuminated (e.g., by the headlights of a car), it reflects the light in the opposite direction (so that it will always be seen by the driver).

72.5. Exercises. **(1)** Find the components of each of the following vectors and their norms:

(i) The vector has endpoints $A(1, 2, 3)$ and $B(-1, 5, 1)$ and is directed from A to B.

(ii) The vector has endpoints $A(1,2,3)$ and $B(-1,5,1)$ and is directed from B to A.

(iii) The vector has the initial point $A(1,2,3)$ and the final point C that is the midpoint of the line segment AB, where $B = (-1,5,1)$.

(iv) The position vector is of a point P obtained from the point $A(-1,2,-1)$ by transporting the latter along the vector $\mathbf{u} = \langle 2,2,1\rangle$ 3 units of length and then along the vector $\mathbf{w} = \langle -3,0,-4\rangle$ 10 units of length.

(v) The position vector of the vertex C of a triangle ABC in the xy plane if A is at the origin, $B = (a,0,0)$, the angle at the vertex B is $\pi/3$, and $|BC| = 3a$.

(2) Consider a triangle ABC. Let $\mathbf{a}$ be a vector from the vertex A to the midpoint of the side BC, let $\mathbf{b}$ be a vector from B to the midpoint of AC, and let $\mathbf{c}$ be a vector from C to the midpoint of AB. Use vector algebra to find $\mathbf{a}+\mathbf{b}+\mathbf{c}$.

(3) Let $\mathbf{u}_k$, $k = 1,2,...,n$, be unit vectors in the plane such that the smallest angle between the two vectors $\mathbf{u}_k$ and $\mathbf{u}_{k+1}$ is $2\pi/n$. What can be said about the sum $\mathbf{u}_1+\mathbf{u}_2+\cdots+\mathbf{u}_n$? What happens when $n \to \infty$?

(4) A plane flies at a speed of v mi/h relative to the air. There is a wind blowing at a speed of u mi/h in the direction that makes the angle θ with the direction in which the plane moves. What is the speed of the plane relative to the ground?

(5) Let pointlike massive objects be positioned at P_i, $i = 1,2,...,n$, and let m_i be the mass at P_i. The point P_0 is called the *center of mass* if

$$m_1\mathbf{r}_1 + m_2\mathbf{r}_2 + \cdots + m_n\mathbf{r}_n = \mathbf{0},$$

where $\mathbf{r}_i$ is the vector from P_0 to P_i. Express the position vector of the center of mass via the position vectors of the point masses. In particular, find the center of mass of three point masses, $m_1 = m_2 = m_3 = m$, located at the vertices of a triangle ABC for $A(1,2,3)$, $B(-1,0,1)$, and $C(1,1,-1)$.

73. The Dot Product

Definition 11.8. (Dot Product).
The dot product $\mathbf{a}\cdot\mathbf{b}$ *of two vectors* $\mathbf{a} = \langle a_1,a_2,a_3\rangle$ *and* $\mathbf{b} = \langle b_1,b_2,b_3\rangle$ *is a number:*

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

It follows from this definition that the dot product has the following properties:

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

which hold for any vectors **a**, **b**, and **c** and a number s. The first property states that the order in which two vectors are multiplied in the dot product does not matter; that is, the dot product is *commutative.* The second property means that the result of the dot product does not depend on whether the vector **a** is scaled first and then multiplied by **b** or the dot product $\mathbf{a}\cdot\mathbf{b}$ is computed first and the result multiplied by s. The third relation shows that the dot product is *distributive.*

73.1. Geometrical Significance of the Dot Product. As it stands, the dot product is an algebraic rule for calculating a number out of six given numbers that are components of the two vectors involved. The components of a vector depend on the choice of the coordinate system. Naturally, one should ask whether the numerical value of the dot product depends on the coordinate system relative to which the components of the vectors are determined. It turns out that it does not. Therefore, it represents an intrinsic geometrical quantity associated with two vectors involved in the product. To elucidate the geometrical significance of the dot product, note first the relation between the dot product and the norm (length) of a vector:

$$\mathbf{a}\cdot\mathbf{a} = a_1^2 + a_2^2 + a_3^2 = \|\mathbf{a}\|^2 \qquad \text{or} \qquad \|\mathbf{a}\| = \sqrt{\mathbf{a}\cdot\mathbf{a}}.$$

Thus, if $\mathbf{a} = \mathbf{b}$ in the dot product, then the latter does not depend on the coordinate system with respect to which the components of **a** are defined. Next, consider the triangle whose adjacent sides are the vectors **a** and **b** as depicted in Figure 11.8 (left panel).

Then the other side of the triangle can be represented by the difference $\mathbf{c} = \mathbf{b} - \mathbf{a}$. The squared length of this latter side is

$$\mathbf{c}\cdot\mathbf{c} = (\mathbf{b}-\mathbf{a})\cdot(\mathbf{b}-\mathbf{a}) = \mathbf{b}\cdot\mathbf{b} + \mathbf{a}\cdot\mathbf{a} - 2\mathbf{a}\cdot\mathbf{b}, \tag{11.4}$$

where the algebraic properties of the dot product have been used. Therefore, the dot product can be expressed via the geometrical invariants, namely, the lengths of the sides of the triangle:

$$\mathbf{a}\cdot\mathbf{b} = \frac{1}{2}\left(\|\mathbf{c}\|^2 - \|\mathbf{b}\|^2 - \|\mathbf{a}\|^2\right). \tag{11.5}$$

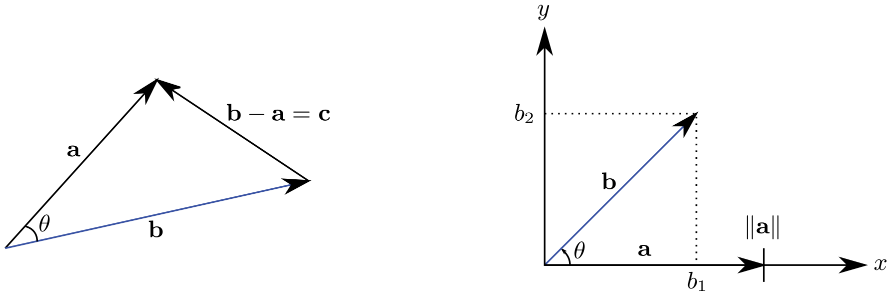

FIGURE 11.8. **Left**: Independence of the dot product from the choice of a coordinate system. The dot product of two vectors that are adjacent sides of a triangle can be expressed via the lengths of the triangle sides as shown in (11.5).
Right: Geometrical significance of the dot product. It determines the angle between two vectors as stated in (11.6). Two nonzero vectors are perpendicular if and only if their dot product vanishes. This follows from (11.5) and the Pythagorean theorem: $\|\mathbf{a}\|^2+\|\mathbf{b}\|^2=\|\mathbf{c}\|^2$ for a right-angled triangle.

This means that the numerical value of the dot product is independent of the choice of coordinate system. Thus, it can be computed in any coordinate system. In particular, let us take the coordinate system in which the vector $\mathbf{a}$ is parallel to the x axis and the vector $\mathbf{b}$ lies in the xy plane as shown in Figure 11.8 (right panel). Let the angle between $\mathbf{a}$ and $\mathbf{b}$ be θ. By definition, this angle lies in the interval $[0,\pi]$. When $\theta=0$, the vectors $\mathbf{a}$ and $\mathbf{b}$ point in the same direction. When $\theta=\pi/2$, they are perpendicular, and they point in the opposite directions if $\theta=\pi$. In the chosen coordinate system, $\mathbf{a}=\langle\|\mathbf{a}\|,0,0\rangle$ and $\mathbf{b}=\langle\|\mathbf{b}\|\cos\theta,\|\mathbf{b}\|\sin\theta,0\rangle$. Hence,

$$(11.6)\qquad \mathbf{a}\cdot\mathbf{b}=\|\mathbf{a}\|\|\mathbf{b}\|\cos\theta \qquad \text{or} \qquad \cos\theta=\frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}\,.$$

Equation (11.6) reveals the geometrical significance of the dot product. It determines the angle between two oriented segments in space. It provides a simple algebraic method to establish a mutual orientation of two straight line segments in space. The following theorem is useful in practical applications.

THEOREM 11.2. (Geometrical Significance of the Dot Product).
Two nonzero vectors are perpendicular if and only if their dot product vanishes:

$$\mathbf{a} \perp \mathbf{b} \quad \Longleftrightarrow \quad \mathbf{a} \cdot \mathbf{b} = 0.$$

In particular, for a triangle with sides a, b, and c and an angle θ between sides a and b, it follows from the relation (11.4) that

$$c^2 = a^2 + b^2 - 2ab\cos\theta.$$

For a right-angled triangle, the Pythagorean theorem is recovered: $c^2 = a^2 + b^2$.

EXAMPLE 11.2. *Consider a triangle whose vertices are* $A(1,1,1)$*,* $B(-1,2,3)$*, and* $C(1,4,-3)$*. Find all the angles of the triangle.*

SOLUTION: Let the angles at the vertices A, B, and C be α, β, and γ, respectively. Then $\alpha + \beta + \gamma = 180°$. So it is sufficient to find any two angles. To find the angle α, define the vectors $\mathbf{a} = \vec{AB} = \langle -2, 1, 2\rangle$ and $\mathbf{b} = \vec{AC} = \langle 0, 3, -4\rangle$. The initial point of these vectors is A, and hence the angle between the vectors coincides with α. Since $\|\mathbf{a}\| = 3$ and $\|\mathbf{b}\| = 5$, by the geometrical property of the dot product,

$$\cos\alpha = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \frac{0+3-8}{15} = -\frac{1}{3} \quad \Longrightarrow$$
$$\alpha = \cos^{-1}(-1/3) \approx 109.5° \,.$$

To find the angle β, define the vectors $\mathbf{a} = \vec{BA} = \langle 2, -1, -2\rangle$ and $\mathbf{b} = \vec{BC} = \langle 2, 2, -6\rangle$ with the initial point at the vertex B. Then the angle between these vectors coincides with β. Since $\|\mathbf{a}\| = 3$, $\|\mathbf{b}\| = 2\sqrt{11}$, and $\mathbf{a}\cdot\mathbf{b} = 4 - 2 + 12 = 14$, one finds $\cos\beta = 14/(6\sqrt{11})$ and $\beta = \cos^{-1}(7/(3\sqrt{11})) \approx 45.3°$. Therefore, $\gamma \approx 180° - 109.5° - 45.3° = 25.2°$. Note that the range of the function $\cos^{-1}$ must be taken from $0°$ to $180°$ in accordance with the definition of the angle between two vectors. □

THEOREM 11.3. (Cauchy-Schwarz Inequality).
For any two vectors $\mathbf{a}$ *and* $\mathbf{b}$*,*

$$|\mathbf{a}\cdot\mathbf{b}| \le \|\mathbf{a}\|\,\|\mathbf{b}\|,$$

where the equality is reached only if the vectors are parallel.

This inequality is a direct consequence of the first relation in (11.6) and the inequality $|\cos\theta| \le 1$. The equality is reached only when $\theta = 0$ or $\theta = \pi$, that is, when $\mathbf{a}$ and $\mathbf{b}$ are parallel.

THEOREM 11.4. (Triangle Inequality).
For any two vectors **a** *and* **b**,

$$\|\mathbf{a}+\mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|.$$

PROOF. Put $\|\mathbf{a}\| = a$ and $\|\mathbf{b}\| = b$ so that $\mathbf{a}\cdot\mathbf{a} = \|\mathbf{a}\|^2 = a^2$ and similarly $\mathbf{b}\cdot\mathbf{b} = b^2$. Using the algebraic rules for the dot product,

$$\|\mathbf{a}+\mathbf{b}\|^2 = (\mathbf{a}+\mathbf{b})\cdot(\mathbf{a}+\mathbf{b}) = a^2+b^2+2\mathbf{a}\cdot\mathbf{b} \leq a^2+b^2+2ab = (a+b)^2,$$

where the Cauchy-Schwarz inequality has been used. By taking the square root of both sides, the triangle inequality is obtained. □

The triangle inequality has a simple geometrical meaning. Consider a triangle with sides **a**, **b**, and **c**. The directions of the vectors are chosen so that $\mathbf{c} = \mathbf{a}+\mathbf{b}$. The triangle inequality states that the length $\|\mathbf{c}\|$ cannot exceed the total length of the other two sides. It is also clear that the maximal length $\|\mathbf{c}\| = \|\mathbf{a}\|+\|\mathbf{b}\|$ is attained only if **a** and **b** are parallel and point in the same direction. If they are parallel but point in the opposite direction, then the length $\|\mathbf{c}\|$ becomes minimal and coincides with the difference of $\|\mathbf{a}\|$ and $\|\mathbf{b}\|$. This observation can be stated in the following algebraic form:

$$\Big|\|\mathbf{a}\| - \|\mathbf{b}\|\Big| \leq \|\mathbf{a}+\mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|. \tag{11.7}$$

73.2. Direction Angles. Consider three unit vectors $\hat{\mathbf{e}}_1 = \langle 1,0,0\rangle$, $\hat{\mathbf{e}}_2 = \langle 0,1,0\rangle$, and $\hat{\mathbf{e}}_3 = \langle 0,0,1\rangle$ that are parallel to the coordinate axes x, y, and z, respectively. By the rules of vector algebra, any vector can be written as the sum of three mutually perpendicular vectors:

$$\mathbf{a} = \langle a_1, a_2, a_3\rangle = a_1\hat{\mathbf{e}}_1 + a_2\hat{\mathbf{e}}_2 + a_3\hat{\mathbf{e}}_3\,.$$

The vectors $a_1\hat{\mathbf{e}}_1$, $a_2\hat{\mathbf{e}}_2$, and $a_3\hat{\mathbf{e}}_3$ are adjacent sides of the rectangle whose largest diagonal coincides with the vector **a** as shown in Figure 11.9 (right panel).

Define the angle α that is counted from the positive direction of the x axis toward the vector **a**. In other words, the angle α is the angle between $\hat{\mathbf{e}}_1$ and **a**. Similarly, the angles β and γ are, by definition, the angles between **a** and the unit vectors $\hat{\mathbf{e}}_2$ and $\hat{\mathbf{e}}_3$, respectively. Then

$$\cos\alpha = \frac{\hat{\mathbf{e}}_1\cdot\mathbf{a}}{\|\hat{\mathbf{e}}_1\|\|\mathbf{a}\|} = \frac{a_1}{\|\mathbf{a}\|}\,,\quad \cos\beta = \frac{\hat{\mathbf{e}}_2\cdot\mathbf{a}}{\|\hat{\mathbf{e}}_2\|\|\mathbf{a}\|} = \frac{a_2}{\|\mathbf{a}\|}\,,$$
$$\cos\gamma = \frac{\hat{\mathbf{e}}_3\cdot\mathbf{a}}{\|\hat{\mathbf{e}}_3\|\|\mathbf{a}\|} = \frac{a_3}{\|\mathbf{a}\|}\,.$$

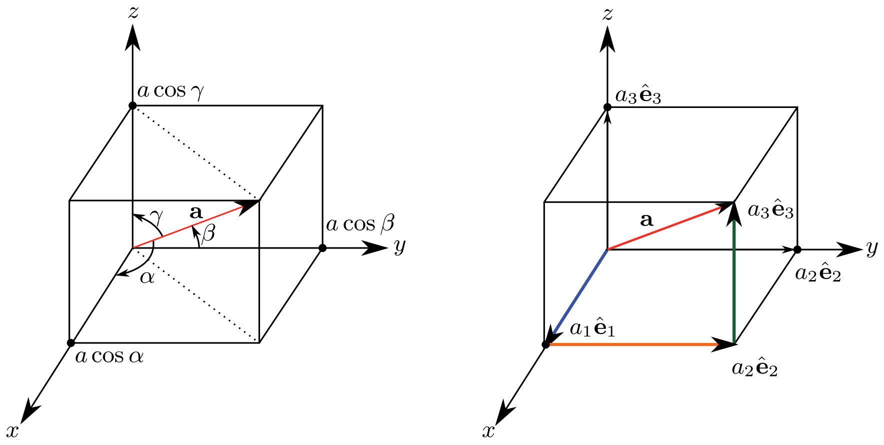

FIGURE 11.9. **Left**: Direction angles of a vector are defined as the angles between the vector and three coordinates axes. Each angle ranges between 0 and π and is counted from the corresponding positive coordinate semiaxis toward the vector. The cosines of the direction angles of a vector are components of the unit vector parallel to that vector.
Right: Decomposition of a vector into the sum of three mutually perpendicular vectors that are parallel to the coordinate axes of a rectangular coordinate system. The vector is the diagonal of the rectangle, whereas the vectors in the sum form the edges of the rectangle.

These cosines are nothing but the components of the unit vector parallel to $\mathbf{a}$:

$$\hat{\mathbf{a}} = \frac{1}{\|\mathbf{a}\|}\,\mathbf{a} = \langle \cos\alpha, \cos\beta, \cos\gamma \rangle \,.$$

Thus, the angles α, β, and γ uniquely determine the direction of a vector. For this reason, they are called *direction angles*. Note that they cannot be set independently because they always satisfy the condition $\|\hat{\mathbf{a}}\| = 1$ or

$$\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1 \,.$$

In practice (physics, mechanics, etc.), vectors are often specified by their magnitude $\|\mathbf{a}\| = a$ and direction angles. The components are then found by $a_1 = a\cos\alpha$, $a_2 = a\cos\beta$, and $a_3 = a\cos\gamma$.

73.3. Practical Applications.

73.3.1. Static Problems. According to Newton's mechanics, a pointlike object that was at rest remains at rest if the vector sum of all forces applied to it vanishes. This is the fundamental law of statics:

$$\mathbf{F}_1 + \mathbf{F}_2 + \cdots + \mathbf{F}_n = \mathbf{0}.$$

This vector equation implies three scalar equations that require vanishing each of the three components of the total force. If there is a system of pointlike objects, then the system is at rest if each object is at rest, and hence the sum of all forces applied to each object vanishes. This gives a system of vector equations, each of which is the above equilibrium condition for a particular object. A typical static problem is to determine either the magnitudes of some forces or the values of some geometrical parameters at which the system in question is at rest.

EXAMPLE 11.3. *Let a ball of mass m be attached to the ceiling by two ropes so that the smallest angle between the first rope and the ceiling is θ_1 and the angle θ_2 is defined similarly for the second rope. Find the magnitudes of the tension forces in the ropes.*

SOLUTION: Set the coordinate system so that the x axis is horizontal and oriented from the first rope to the second ropes as depicted in Figure 11.10 (left panel). The ropes are in the xy plane, while the gravitational force is in the direction opposite to the y axis. Let T_1 and T_2 be the magnitudes of the tension forces. Then in this coordinate system the forces acting on the ball are

$$\begin{aligned} \mathbf{T}_1 &= \langle -T_1\cos\theta_1, T_1\sin\theta_1, 0\rangle\,, \\ \mathbf{T}_2 &= \langle T_2\cos\theta_2, T_2\sin\theta_2, 0\rangle\,, \quad \mathbf{G} = \langle 0, -mg, 0\rangle\,, \end{aligned}$$

where $\mathbf{G}$ is the gravitational force and g is the acceleration of the free fall ($g \approx 9.8$ m/s^2); that is, mg is the weight of the ball. The equilibrium condition

$$\mathbf{T}_1 + \mathbf{T}_2 + \mathbf{G} = \mathbf{0}$$

leads to two equations for the components (the third components of all vectors are identically 0):

$$-T_1\cos\theta_1 + T_2\cos\theta_2 = 0\,, \quad T_1\sin\theta_1 + T_2\sin\theta_2 - mg = 0,$$

which can be solved for T_1 and T_2. By multiplying the first equation by $\sin\theta_1$ and the second by $\cos\theta_1$ and then adding them, one gets $T_2 = mg\cos\theta_1/\sin(\theta_1+\theta_2)$. Substituting T_2 into the first equation, the tension T_1 is obtained. □

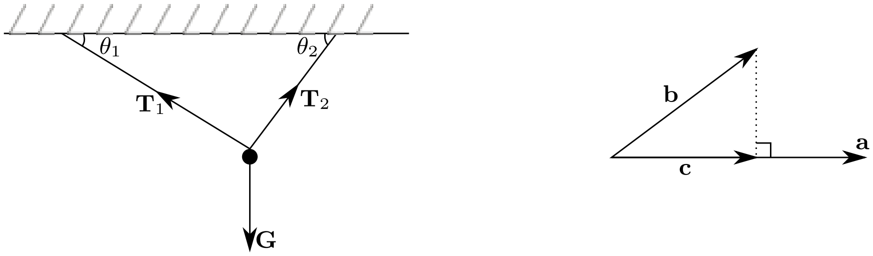

FIGURE 11.10. **Left**: Illustration to Example 11.3. At equilibrium, the vector sum of all forces acting on the ball vanishes. The components of the forces are easy to find in the coordinate system in which the x axis is horizontal and the y axis is vertical.
Right: Illustration to Study Problem 11.11. The vector $\mathbf{c}$ is the projection of a vector $\mathbf{b}$ onto $\mathbf{a}$. It is a vector parallel to $\mathbf{a}$. The initial points of $\mathbf{b}$ and $\mathbf{c}$ coincide. The line through the terminal points of $\mathbf{b}$ and $\mathbf{c}$ is perpendicular to $\mathbf{a}$.

73.3.2. Work Done by a Force. Suppose that an object of mass m moves with speed v. The quantity $K = mv^2/2$ is called the *kinetic energy* of the object. Suppose that the object has moved along a straight line segment from a point P_1 to a point P_2 under the action of a constant force $\mathbf{F}$. A law of physics states that a change in an object's kinetic energy is equal to the work W done by this force:

$$K_2 - K_1 = \mathbf{F} \cdot \overrightarrow{P_1P_2} = W\,,$$

where K_1 and K_2 are the kinetic energies at the initial and final points of the motion, respectively.

EXAMPLE 11.4. *Let an object slide on an inclined plane without friction under the gravitational force. Find the final speed v of the object if the relative height of the initial and final points is h and the object was initially at rest.*

SOLUTION: Choose the coordinate system so that the displacement vector $\overrightarrow{P_1P_2}$ and the gravitational force are in the xy plane. Let the y axis be vertical so that the gravitational force is $\mathbf{F} = \langle 0, -mg, 0\rangle$, where m is the mass and g is the acceleration of the free fall. The initial point is chosen to have the coordinates $(0, h, 0)$ while the final point is $(L, 0, 0)$, where L is the distance the object travels in the horizontal direction while sliding. The displacement vector is $\vec{P_1P_2} = \langle L, -h, 0\rangle$. Since $K_1 = 0$, one has

$$\frac{mv^2}{2} = W = \mathbf{F}\cdot \vec{P_1P_2} = mgh \quad\Longrightarrow\quad v = \sqrt{2gh}\,.$$

Note that the speed is independent of the mass of the object and the inclination angle of the plane (its tangent is h/L); it is fully determined by the relative height only. □

73.4. Study Problems.

Problem 11.11. (Projection of $\mathbf{b}$ onto $\mathbf{a}$).
Consider two vectors $\mathbf{a}$ *and* $\mathbf{b}$ *with a common initial point* O. *Consider the line through the endpoint of* $\mathbf{b}$ *that is perpendicular to* $\mathbf{a}$. *Let* C *be the point intersection of this line with the line containing the vector* $\mathbf{a}$. *Find the vector* $\mathbf{c} = \vec{OC}$. *This vector is called a* projection *of* $\mathbf{b}$ *onto* $\mathbf{a}$.

Solution: (See the right panel of Fig. 11.10). By construction, $\mathbf{c}$ is parallel to $\mathbf{a}$ and hence proportional to it; $\mathbf{c} = s\mathbf{a}$ for some real s. Let the angle between $\mathbf{b}$ and $\mathbf{a}$ be θ. Then, by construction, $s > 0$ if $\theta < 90^\circ$ ($\mathbf{c}$ and $\mathbf{a}$ point in the same direction) and $s < 0$ if $\theta > 90^\circ$ ($\mathbf{c}$ and $\mathbf{a}$ point in the opposite directions). Also, from the right-angled triangle, $\|\mathbf{c}\| = \|\mathbf{b}\|\cos\theta$ if $\theta < 90^\circ$ and $\|\mathbf{c}\| = -\|\mathbf{b}\|\cos\theta$ if $\theta > 90^\circ$. Therefore,

$$\mathbf{c} = s\mathbf{a}\,, \qquad s = \frac{\|\mathbf{b}\|\cos\theta}{\|\mathbf{a}\|} = \frac{\|\mathbf{b}\|\|\mathbf{a}\|\cos\theta}{\|\mathbf{a}\|^2} = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|^2}\,.$$

□

Problem 11.12. *Find all values of* t *for which the vectors* $\mathbf{a} = \langle 2t, 3-t, -1\rangle$ *and* $\mathbf{b} = \langle t, t, 3+t\rangle$ *are orthogonal.*

Solution: By the geometrical property of the dot product, two vectors are orthogonal if and only if their dot product vanishes. Therefore, $\mathbf{a}\cdot\mathbf{b} = 2t^2 + t(3-t) - (3+t) = (t+1)^2 - 4 = 0$. The solutions of this equation are $t = 1$ and $t = -3$. □

Problem 11.13. *Describe the set of points in space whose position vector* $\mathbf{r}$ *satisfies the condition* $(\mathbf{r}-\mathbf{a})\cdot(\mathbf{r}-\mathbf{b}) = 0$. *Hint:* Note that the position vector satisfying the condition $\|\mathbf{r}-\mathbf{c}\| = R$ describes a sphere of radius R whose center has the position vector $\mathbf{c}$.

Solution: The equation of a sphere can also be written in the form $\|\mathbf{r}-\mathbf{c}\|^2 = (\mathbf{r}-\mathbf{c})\cdot(\mathbf{r}-\mathbf{c}) = R^2$. The equation $(\mathbf{r}-\mathbf{a})\cdot(\mathbf{r}-\mathbf{b}) = 0$ can be transformed into the sphere equation by completing the squares. Using the algebraic properties of the dot product,

$$\begin{aligned}(\mathbf{r}-\mathbf{a})\cdot(\mathbf{r}-\mathbf{b}) &= \mathbf{r}\cdot\mathbf{r}-\mathbf{r}\cdot(\mathbf{a}+\mathbf{b})+\mathbf{a}\cdot\mathbf{b}\\ &=(\mathbf{r}-\mathbf{c})\cdot(\mathbf{r}-\mathbf{c})-\mathbf{c}\cdot\mathbf{c}+\mathbf{a}\cdot\mathbf{b},\\ \mathbf{c} &= \tfrac{1}{2}(\mathbf{a}+\mathbf{b}),\\ \mathbf{c}\cdot\mathbf{c}-\mathbf{a}\cdot\mathbf{b} = \mathbf{R}\cdot\mathbf{R}\,, &\qquad \mathbf{R}=\tfrac{1}{2}(\mathbf{a}-\mathbf{b})\,.\end{aligned}$$

Hence, the set is a sphere of radius $R = \|\mathbf{R}\|$, and its center is positioned at $\mathbf{c}$. □

73.5. Exercises.

(1) Find the dot product $\mathbf{a}\cdot\mathbf{b}$ if
(i) $\mathbf{a} = \langle 1, 2, 3\rangle$ and $\mathbf{b} = \langle -1, 2, 0\rangle$
(ii) $\mathbf{a} = \hat{\mathbf{e}}_1 + 3\hat{\mathbf{e}}_2 - \hat{\mathbf{e}}_3$ and $\mathbf{b} = 3\hat{\mathbf{e}}_1 - 2\hat{\mathbf{e}}_2 + \hat{\mathbf{e}}_3$

(2) For what values of b are the vectors $\langle -6\,, b\,, 2\rangle$ and $\langle b\,, b^2\,, b\rangle$ orthogonal?

(3) Find the angle at the vertex A of a triangle ABC for $A(1, 0, 1)$, $B(1, 2, 3)$, and $C(0, 1, 1)$.

(4) Find the cosines of the angles of a triangle ABC for $A(0, 1, 1)$, $B(-2, 4, 3)$, and $C(1, 2, -1)$.

(5) Find the unit vector parallel to $\mathbf{a} = \langle 2, -1, -2\rangle$ and the unit vector whose direction is opposite to $\mathbf{a}$.

(6) Consider a triangle whose any two adjacent sides are unit vectors. What are possible values of the dot products of any two such unit vectors?

(7) Consider a cube whose edges have length a. Find the angle between its largest diagonal and any edge adjacent to the diagonal.

(8) A vector $\mathbf{a}$ makes the angle $\pi/3$ with the positive x axis, the angle $\pi/6$ with the negative y axis, and the angle $\pi/4$ with the positive z axis. Find the components of $\mathbf{a}$ if its length is 6.

(9) Find the components of all unit vectors $\hat{\mathbf{u}}$ that make the angle $\pi/6$ with the positive z axis.
Hint: Put $\hat{\mathbf{u}} = a\hat{\mathbf{v}} + b\hat{\mathbf{e}}_3$, where $\hat{\mathbf{v}}$ is a unit vector in the xy plane. Find a, b, and all $\hat{\mathbf{v}}$ using the polar angle in the xy plane.

(10) If $\mathbf{c} = \|\mathbf{a}\|\mathbf{b} + \|\mathbf{b}\|\mathbf{a}$, where $\mathbf{a}$ and $\mathbf{b}$ are non zero vectors, show that $\mathbf{c}$ bisects the angle between $\mathbf{a}$ and $\mathbf{b}$.

(11) Let the vectors $\mathbf{a}$ and $\mathbf{b}$ have the same length. Show that the vectors $\mathbf{a}+\mathbf{b}$ and $\mathbf{a}-\mathbf{b}$ are orthogonal.

(12) Consider a parallelogram with adjacent sides of length a and b. If d_1 and d_2 are the lengths of the diagonals, prove the parallelogram law: $d_1^2 + d_2^2 = 2(a^2 + b^2)$.

Hint: Consider the vectors **a** and **b** that are adjacent sides of the parallelogram and express the diagonals via **a** and **b**. Use the dot product to evaluate $d_1^2 + d_2^2$.

(13) Two balls of mass m and $3m$, respectively, are connected by a piece of rope of length h. Then the balls are attached to different points on a horizontal ceiling by a piece of rope with the same length h so that the distance L between the points is greater than h but less than $3h$. Find the equilibrium positions of the balls.

74. The Cross Product

74.1. Determinant of a Square Matrix.

DEFINITION 11.9. *The determinant of a* 2×2 *matrix is the number computed by the following rule:*

$$\det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21},$$

that is, the product of the diagonal elements minus the product of the off-diagonal elements.

DEFINITION 11.10. *The determinant of a* 3×3 *matrix* A *is the number obtained by the following rule:*

$$\det\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = a_{11}\det A_{11} - a_{12}\det A_{12} + a_{13}\det A_{13}$$
$$= \sum_{k=1}^{3} (-1)^{k+1} a_{1k} \det A_{1k},$$

$$A_{11} = \begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix}, \quad A_{12} = \begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix}, \quad A_{13} = \begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix},$$

where the matrices A_{1k}, $k = 1, 2, 3$, *are obtained from the original matrix* A *by removing the row and column containing the element* a_{1k}.

It is straightforward to verify that the determinant can be expanded over any row or column:

$$\det A = \sum_{k=1}^{3} (-1)^{k+m} a_{mk} \det A_{mk} \quad \text{for any } m = 1, 2, 3,$$

$$\det A = \sum_{m=1}^{3} (-1)^{k+m} a_{mk} \det A_{mk} \quad \text{for any } k = 1, 2, 3,$$

where the matrix A_{mk} is obtained from A by removing the row and column containing a_{mk}. This definition of the determinant is extended to $N \times N$ square matrices by letting k and m range over $1, 2, ..., N$.

In particular, the determinant of a triangular matrix (i.e., the matrix all of whose elements either above or below the diagonal vanish) is the product of its diagonal elements:

$$\det \begin{pmatrix} a_1 & b & c \\ 0 & a_2 & d \\ 0 & 0 & a_3 \end{pmatrix} = \det \begin{pmatrix} a_1 & 0 & 0 \\ b & a_2 & 0 \\ c & d & a_3 \end{pmatrix} = a_1 a_2 a_3$$

for any numbers b, c, and d. Also, it follows from the expansion of the determinant over any column or row that, if any two rows or any two columns are swapped in the matrix, its determinant changes sign.

Example 11.5. *Calculate* $\det A$, *where*

$$A = \begin{pmatrix} 1\ 2\ 3 \\ 0\ 1\ 3 \\ -1\ 2\ 1 \end{pmatrix}.$$

Solution: Expanding the determinant over the first row yields

$$\det A = 1(1 \cdot 1 - 2 \cdot 3) - 2(0 \cdot 1 - (-1) \cdot 3) + 3(0 \cdot 2 - (-1) \cdot 1) = -8\,.$$

Alternatively, expanding the determinant over the second row yields the same result:

$$\det A = -0(2 \cdot 1 - 3 \cdot 2) + 1(1 \cdot 1 - (-1) \cdot 3) - 3(1 \cdot 2 - (-1) \cdot 2) = -8\,.$$

One can check that the same result can be obtained by expanding the determinant over any row or column. □

74.2. The Cross Product of Two Vectors.

Definition 11.11. (Cross Product).
The cross product of two vectors $\mathbf{a} = \langle a_1, a_2, a_3 \rangle$ *and* $\mathbf{b} = \langle b_1, b_2, b_3 \rangle$ *is a vector that is the determinant of the formal matrix expanded over the first row:*

$$\begin{aligned} \mathbf{a} \times \mathbf{b} &= \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} \\ &= \hat{\mathbf{e}}_1 \det \begin{pmatrix} a_2 & a_3 \\ b_2 & b_3 \end{pmatrix} - \hat{\mathbf{e}}_2 \det \begin{pmatrix} a_1 & a_3 \\ b_1 & b_3 \end{pmatrix} + \hat{\mathbf{e}}_3 \det \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix} \\ &= \langle a_2 b_3 - a_3 b_2,\ a_3 b_1 - a_1 b_3,\ a_1 b_2 - a_2 b_1 \rangle. \qquad (11.8) \end{aligned}$$

Note that the first row of the matrix consists of the unit vectors parallel to the coordinate axes rather than numbers. For this reason, it is referred as to a *formal* matrix. The use of the determinant is merely a compact way to write the algebraic rule to compute the components of the cross product.

The cross product has the following properties that follow from its definition:

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

The first property is obtained by swapping the components of $\mathbf{b}$ and $\mathbf{a}$ in (11.8). It states that the cross product is skew-symmetric (i.e., it is *not commutative* and the order in which the vectors are multiplied is essential); changing the order leads to the opposite vector. The cross product is *distributive* according to the second property. To prove it, change a_i to $a_i + c_i$, $i = 1, 2, 3$, in (11.8). If a vector $\mathbf{a}$ is scaled by a number s and the resulting vector is multiplied by $\mathbf{b}$, the result is the same as the cross product $\mathbf{a}\times\mathbf{b}$ computed first and then scaled by s (change a_i to sa_i in (11.8) and then factor out s).

74.3. Geometrical Significance of the Cross Product. The above algebraic definition of the cross product uses a particular coordinate system relative to which the components of the vectors are defined. Does the cross product depend on the choice of the coordinate system? To answer this question, one should investigate whether both its *direction* and its *magnitude* depend on the choice of the coordinate system. Let us first investigate the mutual orientation of the oriented segments $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{a}\times\mathbf{b}$. A simple algebraic calculation leads to the following result:

$$\mathbf{a}\cdot(\mathbf{a}\times\mathbf{b}) = a_1(a_2b_3 - a_3b_2) + a_2(a_3b_1 - a_1b_3) + a_3(a_1b_2 - a_2b_1) = 0.$$

By the skew symmetry of the cross product, it is also concluded that $\mathbf{b}\cdot(\mathbf{a}\times\mathbf{b}) = -\mathbf{b}\cdot(\mathbf{b}\times\mathbf{a}) = 0$. By the geometrical property of the dot product, the cross product must be perpendicular to both vectors $\mathbf{a}$ and $\mathbf{b}$:

$$(11.9)\quad \mathbf{a}\cdot(\mathbf{a}\times\mathbf{b}) = \mathbf{b}\cdot(\mathbf{a}\times\mathbf{b}) = 0 \quad\Longleftrightarrow\quad \mathbf{a}\times\mathbf{b}\perp\mathbf{a} \text{ and } \mathbf{a}\times\mathbf{b}\perp\mathbf{b}.$$

This shows that the direction of the cross product does not depend on the choice of the coordinate system modulo the reflection $\mathbf{a}\times\mathbf{b}\to\mathbf{a}\times\mathbf{b}$. So, by a suitable rotation, the coordinate system can be oriented so that the cross product is parallel to the z axis. Then the vectors $\mathbf{a}$ and $\mathbf{b}$

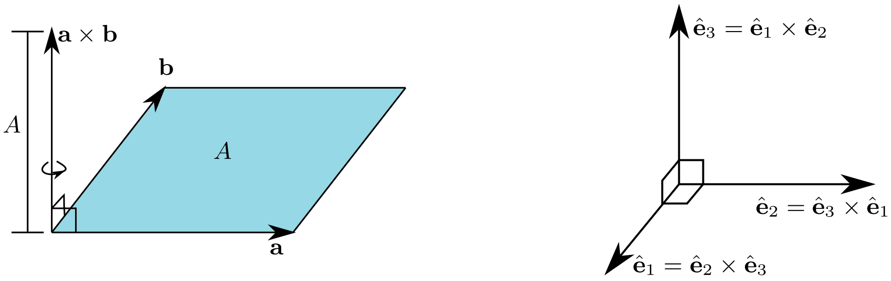

FIGURE 11.11. **Left**: Geometrical interpretation of the cross product of two vectors. The cross product is a vector that is perpendicular to both vectors in the product. Its length equals the area of the parallelogram whose adjacent sides are the vectors in the product. If the fingers of the right hand curl in the direction of a rotation from the first to second vector through the smallest angle between them, then the thumb points in the direction of the cross product of the vectors.
Right: Illustration to Study Problem 11.15.

are in the xy plane. Let θ_a and θ_b be angles counted from the positive x axis counterclockwise toward the vectors **a** and **b**, respectively. The components of these vectors are $\mathbf{a} = \langle \|\mathbf{a}\| \cos\theta_a, \|\mathbf{a}\| \sin\theta_a, 0\rangle$ and $\mathbf{b} = \langle \|\mathbf{b}\| \cos\theta_b, \|\mathbf{b}\| \sin\theta_b, 0\rangle$ (compare with the polar coordinates of two points in a plane with the position vectors **a** and **b**). Then the cross product reads:

$$\mathbf{a} \times \mathbf{b} = \hat{\mathbf{e}}_3 \|\mathbf{a}\| \|\mathbf{b}\| (\cos\theta_a \sin\theta_b - \sin\theta_a \cos\theta_b) = \hat{\mathbf{e}}_3 \|\mathbf{a}\| \|\mathbf{b}\| \sin(\theta_b - \theta_a)\,.$$

Two important conclusions can be deduced from this expression.

74.3.1. The Right-Hand Rule (the Cross-Product Direction). By definition, the angles θ_a and θ_b range over the interval $[0, 2\pi)$. Let $0 \leq \theta \leq \pi$ be the (smallest) angle between the vectors **a** and **b** (just as defined by their dot product). The lengths $\|\mathbf{a}\|$ and $\|\mathbf{b}\|$ and the angle difference $\theta_b - \theta_a$ are independent of the orientation of the coordinate axes in the plane and so must be the cross product. In particular, one can choose the x axis parallel to the vector **a** (or $\theta_a = 0$). Then $\theta_b = \theta$ if $\theta_b \leq \pi$ and $\theta_b = 2\pi - \theta$ if $\pi < \theta_b < 2\pi$. In the former case, $\sin(\theta_b - \theta_a) = \sin\theta$ and the cross product points in the same direction as the z axis, while in the latter case, $\sin(\theta_b - \theta_a) = -\sin\theta$ and the cross product points in the direction opposite to the z axis. This leads to the coordinate independent rule that determines the direction of the cross product known as the *right-hand rule*: *If the fingers of the right hand curl in the direction of a rotation from* $\mathbf{a}$ *toward* $\mathbf{b}$ *through the smallest angle between them, then the thumb points in the direction of* $\mathbf{a}\times\mathbf{b}$.

Thus, the cross product is always perpendicular to the plane containing $\mathbf{a}$ and $\mathbf{b}$ and oriented ("up" or "down" relative to the plane) according to the right-hand rule.

Remark. The transformation in which the coordinate axes change their direction to the opposite is called the *parity transformation*. Evidently, under the parity transformation, coordinates of every point change their sign, and hence every vector (defined as an ordered triple of numbers) changes its direction, $\mathbf{a} = \langle a_1, a_2, a_3\rangle \to \langle -a_1, -a_2, -a_3\rangle = -\mathbf{a}$. However, the cross product of two vectors does not change under the parity transformation: $\mathbf{a}\times\mathbf{b} \to (-\mathbf{a})\times(-\mathbf{b}) = \mathbf{a}\times\mathbf{b}$. For this reason, the cross product is sometimes referred to as a *pseudovector* or an *axial vector*. The coordinate systems related by the parity transformation cannot be obtained from one another by rotations, just like "left" and "right" are swapped in a mirror reflection. There are forces in nature that are axial vectors. So the world and its mirror image can be distinguished by studying the results of the actions of such forces. Physical experiments reveal that the parity symmetry is indeed broken in our Universe!

74.3.2. The Area of a Parallelogram (the Cross-Product Magnitude). By the definition of the angle θ, $\sin\theta \geq 0$. Therefore, the magnitude of the cross product is expressed via the geometrical invariants—the length of the vectors and the angle between them:

$$\|\mathbf{a}\times\mathbf{b}\| = \|\mathbf{a}\|\|\mathbf{b}\|\sin\theta\,.$$

Now consider the parallelogram with adjacent sides $\mathbf{a}$ and $\mathbf{b}$. If $\|\mathbf{a}\|$ is the length of its base, then $h = \|\mathbf{b}\|\sin\theta$ is its height. Then the magnitude of the cross product, $\|\mathbf{a}\times\mathbf{b}\| = \|\mathbf{a}\|h$, must be the area of the parallelogram. This completes a proof of the following theorem.

THEOREM 11.5. (Geometrical Significance of the Cross Product). *The cross product* $\mathbf{a}\times\mathbf{b}$ *of vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *is the vector that is perpendicular to both vectors,* $\mathbf{a}\times\mathbf{b}\perp\mathbf{a}$ *and* $\mathbf{a}\times\mathbf{b}\perp\mathbf{b}$, *has a magnitude equal to the area of the parallelogram with adjacent sides* $\mathbf{a}$ *and* $\mathbf{b}$, *and is directed according to the right-hand rule.*

It should be emphasized that no coordinate system is required to determine the cross product of two vectors. The geometrical properties of the cross product can be used to obtain another algebraic criterion for two vectors that are parallel.

Corollary 11.1. *Two nonzero vectors are parallel if and only if their cross product vanishes:*

$$\mathbf{a}\times\mathbf{b}=\mathbf{0} \quad\Longleftrightarrow\quad \mathbf{a}\,\|\,\mathbf{b}\,.$$

When two vectors are parallel, the area of the corresponding parallelogram vanishes, $\|\mathbf{a}\times\mathbf{b}\| = 0$. The latter is true if and only if $\mathbf{a}\times\mathbf{b}=\mathbf{0}$. Conversely, for two parallel vectors, there is a number s such that $\mathbf{a}=s\mathbf{b}$. Hence, $\mathbf{a}\times\mathbf{b}=(s\mathbf{a})\times\mathbf{b}=s(\mathbf{b}\times\mathbf{b})=\mathbf{0}$.

One of the most important applications of the cross product is in calculations of the areas of planar figures in space.

Corollary 11.2. (Area of a Triangle).
Consider a triangle with two adjacent sides represented by the vectors $\mathbf{a}$ *and* $\mathbf{b}$ *such that the vectors have the same initial point at a vertex of the triangle. Then the area of the triangle is*

$$\text{Area }\triangle=\frac{1}{2}\|\mathbf{a}\times\mathbf{b}\|.$$

Indeed, by the geometrical construction, the area of the triangle is half of the area of a parallelogram with adjacent sides $\mathbf{a}$ and $\mathbf{b}$.

Example 11.6. *Let* $A=(1,1,1)$*,* $B=(2,-1,3)$*, and* $C=(-1,3,1)$*. Find the area of the triangle* ABC *and a vector normal to the plane that contains the triangle.*

Solution: According to the geometrical properties of the cross product, in order to find a vector normal to a plane, one should take the cross product of any two nonparallel vectors in the plane. For example, $\mathbf{a}=\vec{AB}=\langle 1,-2,2\rangle$ and $\mathbf{b}=\vec{AC}=\langle -2,2,0\rangle$. Then $\mathbf{a}\times\mathbf{b}=\langle -4,-4,-6\rangle$ is normal to the plane. Note that the cross product of any other pair of vectors corresponding to the sides of the triangle can only be a scaled vector $s\langle -4,-4,6\rangle$ because any two normal vectors of a given plane must be parallel and hence proportional. Since $\|\langle -4,-4,-6\rangle\| = 2\|\langle 2,2,3\rangle\| = 2\sqrt{17}$, the area of the triangle ABC is $\sqrt{17}$ by Corollary 11.2. The units here are squared units of length used to measure the coordinates of the triangle vertices (e.g., m^2 if the coordinates are measured in meters). □

74.4. Study Problems.

Problem 11.14. *Find the most general vector* $\mathbf{r}$ *that satisfies the equations* $\mathbf{a}\cdot\mathbf{r}=0$ *and* $\mathbf{b}\cdot\mathbf{r}=0$*, where* $\mathbf{a}$ *and* $\mathbf{b}$ *are nonzero, nonparallel vectors.*

SOLUTION: The conditions imposed on $\mathbf{r}$ hold if and only if the vector $\mathbf{r}$ is orthogonal to both vectors $\mathbf{a}$ and $\mathbf{b}$. Therefore, it must be parallel to their cross product. Thus, $\mathbf{r} = t(\mathbf{a} \times \mathbf{b})$ for any real t. □

Problem 11.15. *Use geometrical means to find the cross products of the unit vectors parallel to the coordinate axes.*

SOLUTION: Consider $\hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_2$. Since $\hat{\mathbf{e}}_1 \perp \hat{\mathbf{e}}_2$ and $\|\hat{\mathbf{e}}_1\| = \|\hat{\mathbf{e}}_2\| = 1$, their cross product must be a unit vector perpendicular to both $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$. There are only two such vectors, $\pm\hat{\mathbf{e}}_3$. By the right-hand rule, it follows that

$$\hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_2 = \hat{\mathbf{e}}_3 \,.$$

Similarly, the other cross products are shown to be obtained by cyclic permutations of the indices 1, 2, and 3 in the above relation. A permutation of any two indices leads to a change in sign (e.g., $\hat{\mathbf{e}}_2 \times \hat{\mathbf{e}}_1 = -\hat{\mathbf{e}}_3$). Since a cyclic permutation of three indices $\{ijk\} \to \{kij\}$ (and so on) consists of two permutations of any two indices, the relation between the unit vectors can be cast in the form

$$\hat{\mathbf{e}}_i = \hat{\mathbf{e}}_j \times \hat{\mathbf{e}}_k \;, \quad \{ijk\} = \{123\} \text{ and cyclic permutations.}$$

□

Problem 11.16. *Prove the "bac − cab" rule:*

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

SOLUTION: If $\mathbf{c}$ and $\mathbf{b}$ are parallel, then $\mathbf{d} = \mathbf{0}$. If $\mathbf{c}$ and $\mathbf{b}$ are not parallel, then $\mathbf{d}$ must be perpendicular to both $\mathbf{a}$ and $\mathbf{b} \times \mathbf{c}$. From the condition $\mathbf{d} \perp \mathbf{b} \times \mathbf{c}$, it follows that $\mathbf{d}$ lies in the plane containing $\mathbf{b}$ and $\mathbf{c}$ and hence is a linear combination of them, $\mathbf{d} = s\mathbf{b} + t\mathbf{c}$. From the condition $\mathbf{d} \perp \mathbf{a}$ or $\mathbf{a} \cdot \mathbf{d} = 0$, it follows that $s = p(\mathbf{a} \cdot \mathbf{c})$ and $t = -p(\mathbf{a} \cdot \mathbf{b})$ for some real p. Since the magnitude of the cross product is independent of the choice of the coordinate system, the number p can be fixed by computing $\mathbf{d}$ in any convenient coordinate system. By rotating the coordinate system, one can always direct the x axis along the vector $\mathbf{c}$ so that $\mathbf{c} = \|\mathbf{c}\|\hat{\mathbf{e}}_1$, while the vector $\mathbf{b}$ lies in the xy plane so that $\mathbf{b} = b_1\hat{\mathbf{e}}_1 + b_2\hat{\mathbf{e}}_2$. Then $\mathbf{b} \times \mathbf{c} = -\hat{\mathbf{e}}_3 b_2\|\mathbf{c}\|$ and therefore, for a generic $\mathbf{a} = \langle a_1, a_2, a_3\rangle$,

$$\begin{aligned}\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) &= -\hat{\mathbf{e}}_1\|\mathbf{c}\|a_2b_2 + \hat{\mathbf{e}}_2 b_2\|\mathbf{c}\|a_1 = -\mathbf{c}a_2b_2 + (\mathbf{b} - b_1\hat{\mathbf{e}}_1)\|\mathbf{c}\|a_1 \\ &= \mathbf{b}\|\mathbf{c}\|a_1 - \mathbf{c}(a_1b_1 + a_2b_2) = \mathbf{b}(\mathbf{c} \cdot \mathbf{a}) - \mathbf{c}(\mathbf{a} \cdot \mathbf{b})\,,\end{aligned}$$

that is, $p = 1$. Of course, the statement can also be proved by a direct use of the algebraic definition of the cross product (a brute-force method). □

Problem 11.17. *Prove the Jacobi identity*

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) + \mathbf{b} \times (\mathbf{c} \times \mathbf{a}) + \mathbf{c} \times (\mathbf{a} \times \mathbf{b}) = \mathbf{0}.$$

Solution: Note that the second and third terms on the left side are obtained from the first by cyclic permutations of the vectors. Making use of the *bac – cab* rule for the first term and then adding to it its two cyclic permutations, one can convince oneself that the coefficients at each of the vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are added up to make 0. □

Remark. Note that the Jacobi identity implies in particular that

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) \neq (\mathbf{a} \times \mathbf{b}) \times \mathbf{c};$$

that is, the multiplication law defined by the cross product does not generally obey the associative law for multiplication of numbers.

Problem 11.18. *Consider all vectors in a plane. Any such vector* $\mathbf{a}$ *can be uniquely determined by specifying its length* $a = \|\mathbf{a}\|$ *and the angle* θ_a *that is counted from the positive* x *axis toward the vector* $\mathbf{a}$ *(i.e.,* $0 \leq \theta_a < 2\pi$*). The relation* $\langle a_1, a_2 \rangle = \langle a\cos\theta_a, a\sin\theta_a \rangle$ *establishes a one-to-one correspondence between ordered pairs* (a_1, a_2) *and* (a, θ_a)*. Define the vector product of two vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *as the vector* $\mathbf{c}$ *for which* $c = ab$ *and* $\theta_c = \theta_a + \theta_b$*. Show that this product is associative and commutative, that is, that* $\mathbf{c}$ *does not depend on the order of vectors in the product.*

Solution: Let us denote the vector product by a small circle to distinguish it from the dot and cross products, $\mathbf{a} \circ \mathbf{b} = \mathbf{c}$. Since $\mathbf{c} = \langle ab\cos(\theta_a + \theta_b), ab\sin(\theta_a + \theta_b) \rangle$, the commutativity of the vector product $\mathbf{a} \circ \mathbf{b} = \mathbf{b} \circ \mathbf{a}$ follows from the commutativity of the product and addition of numbers: $ab = ba$ and $\theta_a + \theta_b = \theta_b + \theta_a$. Similarly, the associativity of the vector product $(\mathbf{a} \circ \mathbf{b}) \circ \mathbf{c} = \mathbf{a} \circ (\mathbf{b} \circ \mathbf{c})$ follows from the associativity of the product and addition of ordinary numbers: $(ab)c = a(bc)$ and $(\theta_a + \theta_b) + \theta_c = \theta_a + (\theta_b + \theta_c)$. □

Remark. The vector product introduced for vectors in a plane is known as the *product of complex numbers*. It is interesting to note that no commutative and associative vector product (i.e., "vector times vector = vector") can be defined in a Euclidean space of more than two dimensions.

Problem 11.19. *Let* $\mathbf{u}$ *be a vector rotating in the* xy *plane about the* z *axis. Given a vector* $\mathbf{v}$*, find the position of* $\mathbf{u}$ *such that the magnitude of the cross product* $\mathbf{v} \times \mathbf{u}$ *is maximal.*

Solution: For any two vectors, $\|\mathbf{v} \times \mathbf{u}\| = \|\mathbf{v}\|\|\mathbf{u}\| \sin\theta$, where θ is the angle between $\mathbf{v}$ and $\mathbf{u}$. The magnitude of $\mathbf{v}$ is fixed, while the magnitude of $\mathbf{u}$ does not change when rotating. Therefore, the absolute maximum of the cross-product magnitude is reached when $\sin\theta = 1$ or $\cos\theta = 0$ (i.e., when the vectors are orthogonal). The corresponding algebraic condition is $\mathbf{v}\cdot\mathbf{u} = 0$. Since $\mathbf{u}$ is rotating in the xy plane, its components are $\mathbf{u} = \langle \|\mathbf{u}\| \cos\phi, \|\mathbf{u}\| \sin\phi, 0\rangle$, where $0 \le \phi < 2\pi$ is the angle counted counterclockwise from the x axis toward the current position of $\mathbf{u}$. Put $\mathbf{v} = \langle v_1, v_2, v_3\rangle$. Then the direction of $\mathbf{u}$ is determined by the equation $\mathbf{v} \cdot \mathbf{u} = \|\mathbf{u}\|(v_1 \cos\phi + v_2 \sin\phi) = 0$, and hence $\tan\phi = -v_1/v_2$. This equation has two solutions in the range $0 \le \phi < 2\pi$: $\phi = -\tan^{-1}(v_1/v_2)$ and $\phi = -\tan^{-1}(v_1/v_2) + \pi$. Geometrically, these solutions correspond to the case when $\mathbf{u}$ is parallel to the line $y = -(v_1/v_2)x$ in the xy plane. □

74.5. Exercises. **(1)** Find the cross product $\mathbf{a} \times \mathbf{b}$ if
(i) $\mathbf{a} = \langle 1, 2, 3\rangle$ and $\mathbf{b} = \langle -1, 0, 1\rangle$
(ii) $\mathbf{a} = \hat{\mathbf{e}}_1 + 3\hat{\mathbf{e}}_2 - \hat{\mathbf{e}}_3$ and $\mathbf{b} = 3\hat{\mathbf{e}}_1 - 2\hat{\mathbf{e}}_2 + \hat{\mathbf{e}}_3$

(2) Find the area of a triangle ABC for $A(1, 0, 1)$, $B(1, 2, 3)$, and $C(0, 1, 1)$ and a nonzero vector perpendicular to the plane containing the triangle.

(3) Suppose $\mathbf{a}$ lies in the xy plane, its initial point is at the origin, and its terminal point is in first quadrant of the xy plane. Let $\mathbf{b}$ be parallel to $\hat{\mathbf{e}}_3$. Use the right-hand rule to determine whether the angle between $\mathbf{a} \times \mathbf{b}$ and the unit vectors parallel to the coordinate axes lies in the interval $(0, \pi/2)$ or $(\pi/2, \pi)$ or equals $\pi/2$.

(4) If vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ have the initial point at the origin and lie, respectively, in the positive quadrants of the xy, yz, and xz planes, find the octants in which the pairwise cross products of these vectors lie.

(5) Let $A = (1, 2, 1)$ and $B = (-1, 0, 2)$ be vertices of a parallelogram. If the other two vertices are obtained by moving A and B by 3 units of length along the vector $\mathbf{a} = \langle 2, 1, -2\rangle$, find the area of the parallelogram.

(6) Consider four points in space. Suppose that the coordinates of the points are known. Describe a procedure based on vector algebra to determine whether the points are in one plane. In particular, are the points $(1, 2, 3)$, $(-1, 0, 1)$, $(1, 3, -1)$, and $(0, 1, 2)$ in one plane?

(7) Let the sides of a triangle have lengths a, b, and c and let the angles at the vertices opposite to the sides a, b, and c be, respectively,

α, β, and γ. Prove that

$$\frac{\sin\alpha}{a} = \frac{\sin\beta}{b} = \frac{\sin\gamma}{c}.$$

Hint: Define the sides as vectors and express the area of the triangle via the vectors at each vertex of the triangle.

(8) Consider a polygon with four vertices A, B, C, and D. If the coordinates of the vertices are specified, describe the procedure based on vector algebra to calculate the area of the polygon. In particular, put $A = (0, 0)$ $B = (x_1, y_1)$, $C = (x_2, y_2)$, and $D = (x_3, y_3)$, and express the area via x_i and y_i, $i = 1, 2, 3$.

(9) Consider a parallelogram. Construct another parallelogram whose adjacent sides are diagonals of the first parallelogram. Find the relation between the areas of the parallelograms.

(10) Given two nonparallel vectors $\mathbf{a}$ and $\mathbf{b}$, show that any vector $\mathbf{r}$ in space can be written as a linear combination $\mathbf{r} = x\mathbf{a} + y\mathbf{b} + z\mathbf{a} \times \mathbf{b}$ and that the numbers x, y, and z are unique for every $\mathbf{r}$.
Hint: See Study Problems 11.14 and 11.6.

(11) A tetrahedron is a solid with four vertices and four triangular faces. Let $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$, and $\mathbf{v}_4$ be vectors with lengths equal to the areas of the faces and directions perpendicular to the faces and pointing outward. Show that $\mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3 + \mathbf{v}_4 = \mathbf{0}$.

(12) If $\mathbf{a} \cdot \mathbf{b} = \mathbf{a} \cdot \mathbf{c}$ and $\mathbf{a} \times \mathbf{b} = \mathbf{a} \times \mathbf{c}$, does it follow that $\mathbf{b} = \mathbf{c}$?

75. The Triple Product

Definition 11.12. *The triple product of three vectors* $\mathbf{a}$, $\mathbf{b}$, *and* $\mathbf{c}$ *is a number obtained by the rule:* $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$.

It follows from the algebraic definition of the cross product and the definition of the determinant of a 3×3 matrix that

$$\begin{aligned}\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) &= a_1 \det\begin{pmatrix} b_2 & b_3 \\ c_2 & c_3 \end{pmatrix} - a_2 \det\begin{pmatrix} b_1 & b_3 \\ c_1 & c_3 \end{pmatrix} + a_3 \det\begin{pmatrix} b_1 & b_2 \\ c_1 & c_2 \end{pmatrix} \\ &= \det\begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix}.\end{aligned}$$

This provides a convenient way to calculate the numerical value of the triple product. If two rows of a matrix are swapped, then its determinant changes sign. Therefore,

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

This means, in particular, that the absolute value of the triple product is independent of the order of the vectors in the triple product.

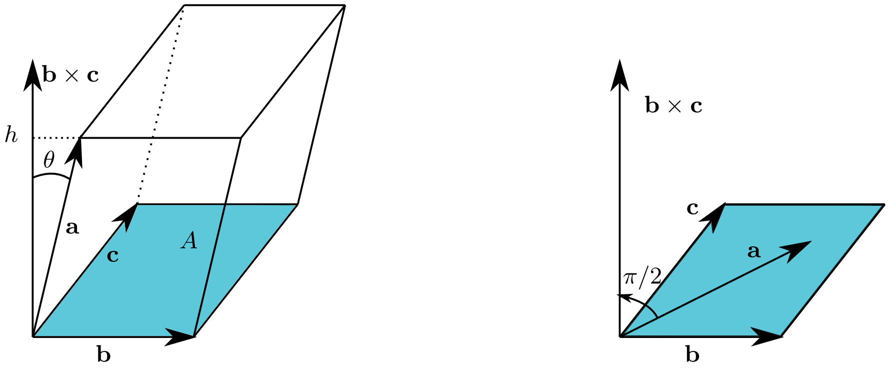

FIGURE 11.12. **Left**: Geometrical interpretation of the triple product as the volume of the parallelepiped whose adjacent sides are the vectors in the product: $h = \|\mathbf{a}\| \cos\theta$, $A = \|\mathbf{b} \times \mathbf{c}\|$, $V = hA = \|\mathbf{a}\|\,\|\mathbf{b} \times \mathbf{c}\| \cos\theta = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$.
Right: Test for the coplanarity of three vectors. Three vectors are coplanar if and only if their triple product vanishes: $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = 0$.

75.1. Geometrical Significance of the Triple Product. Suppose that $\mathbf{b}$ and $\mathbf{c}$ are not parallel (otherwise, $\mathbf{b}\times\mathbf{c} = \mathbf{0}$). Let θ be the angle between $\mathbf{a}$ and $\mathbf{b} \times \mathbf{c}$ as shown in Figure 11.12 (left panel). If $\mathbf{a} \perp \mathbf{b} \times \mathbf{c}$ (i.e., $\theta = \pi/2$), then the triple product vanishes. Let $\theta \neq \pi/2$. Consider the parallelepiped whose adjacent sides being are the vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$. The faces of the parallelepiped are the parallelograms whose adjacent sides are pairs of the vectors. In particular, the cross product $\mathbf{b} \times \mathbf{c}$ is perpendicular to the face containing the vectors $\mathbf{b}$ and $\mathbf{c}$, whereas $A = \|\mathbf{b} \times \mathbf{c}\|$ is the area of this face of the parallelepiped (the area of the parallelogram with adjacent sides $\mathbf{b}$ and $\mathbf{c}$). By the geometrical property of the dot product, $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = A\|\mathbf{a}\| \cos\theta$. On the other hand, the distance between the two faces parallel to both $\mathbf{b}$ and $\mathbf{c}$ (or the height of the parallelepiped) is $h = \|\mathbf{a}\| \cos\theta$ if $\theta < \pi/2$ and $h = -\|\mathbf{a}\| \cos\theta$ if $\theta > \pi/2$ or, $h = \|\mathbf{a}\| |\cos\theta|$. The volume of the parallelepiped is $V = Ah$. This leads to the following theorem.

THEOREM 11.6. (Geometrical Significance of the Triple Product). *The volume V of a parallelepiped whose adjacent sides are the vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ is the absolute value of their triple product:*

$$V = |\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})| .$$

Thus, the triple product is a convenient algebraic tool for calculating volumes. There is a useful consequence of this theorem.

DEFINITION 11.13. (Coplanar Vectors).
Vectors are said to be coplanar if they are in one plane.

Clearly, any two vectors are always coplanar. What is an algebraic condition for three vectors being coplanar?

COROLLARY 11.3. (Criterion for Three Vectors to Be Coplanar).
Three vectors are coplanar if and only if their triple product vanishes:

$$\mathbf{a},\ \mathbf{b},\ \mathbf{c} \text{ are coplanar} \quad \Longleftrightarrow \quad \mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}) = 0.$$

Indeed, if the vectors are coplanar (Figure 11.12, right panel), then the cross product of any two vectors must be perpendicular to the plane where the vectors are and therefore the triple product vanishes. If, conversely, the triple product vanishes, then either $\mathbf{b}\times\mathbf{c} = \mathbf{0}$ or $\mathbf{a}\perp\mathbf{b}\times\mathbf{c}$. In the former case, $\mathbf{b}$ is parallel to $\mathbf{c}$, or $\mathbf{c} = t\mathbf{b}$, and hence $\mathbf{a}$ always lies in a plane with $\mathbf{b}$ and $\mathbf{c}$. In the latter case, all three vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are perpendicular to $\mathbf{b}\times\mathbf{c}$ and therefore must be in one plane (perpendicular to $\mathbf{b}\times\mathbf{c}$).

EXAMPLE 11.7. *Determine whether the points $A(1,1,1)$, $B(2,0,2)$, $C(3,1,-1)$, and $D(0,2,3)$ are in the same plane.*

SOLUTION: Consider the vectors $\mathbf{a} = \vec{AB} = \langle 1,-1,1\rangle$, $\mathbf{b} = \vec{AC} = \langle 2,0,2\rangle$, and $\mathbf{c} = \vec{AD} = \langle -1,1,2\rangle$. The points in question are in the same plane if and only if the vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are coplanar, or $\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}) = 0$. One finds.

$$\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}) = \det\begin{pmatrix} 1 & -1 & 1 \\ 2 & 0 & 2 \\ -1 & 1 & 2 \end{pmatrix} = 1(0-2)+1(4+2)+1(2-0) = 6 \neq 0\,.$$

Therefore, the points are not in the same plane. □

The triple product can be used to find the distances between two sets of points in space. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two sets of points in space. Let a point A_1 belong to $\mathcal{S}_1$, let a point A_2 belong to $\mathcal{S}_2$, and let $|A_1A_2|$ be the distance between them.

75.2. Distances Between Lines and Planes. If the lines or planes in space are not intersecting, then how can one find the distance between them? This question can be answered using the geometrical properties of the triple and cross products.

Definition 11.14. (Distance Between Sets in Space).
The distance D between two sets of points in space, $\mathcal{S}_1$ and $\mathcal{S}_2$, is the largest number that is less than or equal to all the numbers $|A_1A_2|$ when the point A_1 ranges over $\mathcal{S}_1$ and the point A_2 ranges over $\mathcal{S}_2$.

Naturally, if the sets have at least one common point, the distance between them vanishes. The distance between sets may vanish even if the sets have no common points. For example, let $\mathcal{S}_1$ be an open interval $(0, 1)$ on, say, the x axis, while $\mathcal{S}_2$ is the interval $(1, 2)$ on the same axis. Apparently, the sets have no common points (the point $x = 1$ does not belongs to either of them). The distance is the largest number D such that $D \leq |x_1 - x_2|$, where $0 < x_1 < 1$ and $1 < x_2 < 2$. The value of $|x_1 - x_2| > 0$ can be made smaller than any preassigned positive number by taking x_1 and x_2 close enough to 1. Since the distance $D \geq 0$, the only possible value is $D = 0$. Intuitively, the sets are separated by a single point that is not an "extended" object, and hence the distance between them should vanish. In other words, there are situations in which the minimum of $|A_1A_2|$ is not attained for some $A_1 \in \mathcal{S}_1$, or some $A_2 \in \mathcal{S}_2$, or both. Nevertheless, the distance between the sets is still well defined as the largest number that is less than or equal to all numbers $|A_1A_2|$. Such a number is called the *infimum* of the set of numbers $|A_1A_2|$ and denoted $\inf |A_1A_2|$. Thus,

$$D = \inf |A_1A_2| \,, \quad A_1 \in \mathcal{S}_1 \,, \quad A_2 \in \mathcal{S}_2 \,.$$

The notation $A_1 \in \mathcal{S}_1$ stands for "a point A_1 belongs to the set $\mathcal{S}_1$," or simply "A_1 is an element of $\mathcal{S}_1$." The definition is illustrated in Figure 11.13 (left panel).

Theorem 11.7. (Distance Between Parallel Planes).
The distance between parallel planes $\mathcal{P}_1$ and $\mathcal{P}_2$ is given by

$$D = \frac{|\vec{AP} \cdot (\vec{AB} \times \vec{AC})|}{\|\vec{AB} \times \vec{AC}\|},$$

where A, B, and C are any three points in the plane $\mathcal{P}_1$ that are not on the same line, and P is any point in the plane $\mathcal{P}_2$.

Proof. Since the points A, B, and C are not on the same line, the vectors $\mathbf{b} = \vec{AB}$ and $\mathbf{c} = \vec{AC}$ are not parallel and their cross product is a vector perpendicular to the planes (see Figure 11.13, right panel). Consider the parallelepiped with adjacent sides $\mathbf{a} = \vec{AP}$, $\mathbf{b}$, and $\mathbf{c}$. Two of its faces lie in the parallel planes, one in $\mathcal{P}_1$ and the other in $\mathcal{P}_2$ (i.e., the parallelograms with adjacent sides $\mathbf{b}$ and $\mathbf{c}$). The distance between the planes is, by construction, the parallelepiped height, which is equal

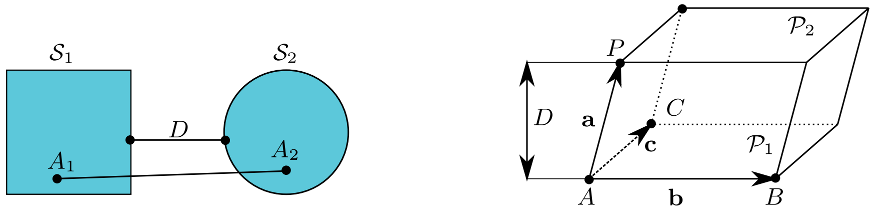

FIGURE 11.13. **Left**: Distance between two point sets $\mathcal{S}_1$ and $\mathcal{S}_2$ defined as the largest distance that is less than or equal to all distances $|A_1A_2|$, where A_1 ranges over all points in $\mathcal{S}_1$ and A_2 ranges over all points in $\mathcal{S}_2$.
Right: Distance between two parallel planes (Theorem 11.7). Consider a parallelepiped whose opposite faces lie in the planes $\mathcal{P}_1$ and $\mathcal{P}_2$. Then the distance D between the planes is the height of the parallelepiped, which can be computed as the ratio $D = V/A$, where $V = |\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})|$ is the volume of the parallelepiped and $A = \|\mathbf{b} \times \mathbf{c}\|$ is the area of the face.

to V/A, where V and A are the parallelepiped volume and area of the face parallel to $\mathbf{b}$ and $\mathbf{c}$. The conclusion follows from the geometrical properties of the triple and cross products: $V = |\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})|$ and $A = \|\mathbf{b} \times \mathbf{c}\|$. □

Similarly, the distance between two parallel lines $\mathcal{L}_1$ and $\mathcal{L}_2$ can be determined. Recall that lines are parallel if they are not intersecting and lie in the same plane. Let A and B be any two points on the line $\mathcal{L}_1$ and let C be any point on the line $\mathcal{L}_2$. Consider the parallelogram with adjacent sides $\mathbf{a} = \vec{AB}$ and $\mathbf{b} = \vec{AC}$ as depicted in Figure 11.14 (left panel). The distance between the lines is the height of this parallelogram, which is $A/\|\mathbf{a}\|$, where $A = \|\mathbf{a} \times \mathbf{b}\|$, is the parallelogram area and $\|\mathbf{a}\|$ is the length of its base.

COROLLARY 11.4. (Distance Between Parallel Lines).
The distance between two parallel lines $\mathcal{L}_1$ and $\mathcal{L}_2$ is

$$D = \frac{\|\vec{AB} \times \vec{AC}\|}{\|\vec{AB}\|},$$

where A and B are any two distinct points on the line $\mathcal{L}_1$ and C is any point on the line $\mathcal{L}_2$.

By construction, D is the height of the parallelogram whose adjacent sides are the vectors $\vec{AB}$ and $\vec{AC}$. Therefore, D is its area divided

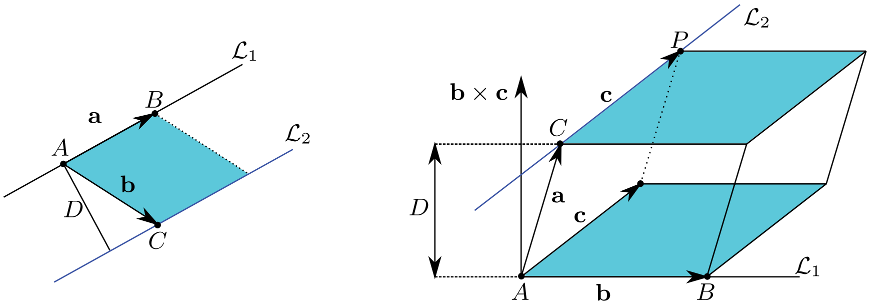

FIGURE 11.14. **Left**: Distance between two parallel lines. Consider a parallelogram whose two parallel sides lie in the lines. Then the distance between the lines is the height of the parallelogram (Corollary 11.4).
Right: Distance between skew lines. Consider a parallelepiped whose two non parallel edges AB and CP in the opposite faces lie in the skew lines $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively. Then the distance between the lines is the height of the parallelepiped, which can be computed as the ratio of the volume and the area of the face (Corollary 11.5).

by the length of the base $\vec{AB}$. By the geometrical properties of the cross product, $\|\vec{AB} \times \vec{AC}\|$ is the area of the parallelogram.

DEFINITION 11.15. (Skew Lines).
Two lines that are not intersecting and not parallel are called skew lines.

To determine the distance between skew lines $\mathcal{L}_1$ and $\mathcal{L}_2$, consider any two points A and B on $\mathcal{L}_1$ and any two points C and P on $\mathcal{L}_2$. Define the vectors $\mathbf{b} = \vec{AB}$ and $\mathbf{c} = \vec{CP}$ that are parallel to lines $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively. Since the lines are not parallel, the cross product $\mathbf{b} \times \mathbf{c}$ does not vanish. The lines $\mathcal{L}_1$ and $\mathcal{L}_2$ lie in the parallel planes perpendicular to $\mathbf{b} \times \mathbf{c}$ (by the geometrical properties of the cross product, $\mathbf{b} \times \mathbf{c}$ is perpendicular to $\mathbf{b}$ and $\mathbf{c}$). The distance between the lines coincides with the distance between these parallel planes. Consider the parallelepiped with adjacent sides $\mathbf{a} = \vec{AC}$, $\mathbf{b}$, and $\mathbf{c}$ as shown in Figure 11.14 (right panel). The lines lie in the parallel planes that contain the faces of the parallelepiped parallel to the vectors $\mathbf{b}$ and $\mathbf{c}$. Thus, the distance between skew lines can be found from the distance between the parallel planes containing them, $D = V/A$, where V and A are the parallelepiped volume and the area of the base $A = \|\mathbf{b} \times \mathbf{c}\|$.

Corollary 11.5. (Distance Between Skew Lines). *The distance between two skew lines $\mathcal{L}_1$ and $\mathcal{L}_2$ is*

$$D = \frac{|\vec{AC} \cdot (\vec{AB} \times \vec{CP})|}{\|\vec{AB} \times \vec{CP}\|},$$

where A and B are any two distinct points on $\mathcal{L}_1$, while C and P are any two distinct points on $\mathcal{L}_2$.

Note that, given any two lines, one can calculate D, provided, of course, that the vectors $\vec{AB}$ and $\vec{CP}$ are not parallel; that is, the lines are not parallel. If $D = 0$, then the lines must intersect. This gives a simple algebraic criterion for two lines being skew or intersecting.

75.3. Study Problems.

Problem 11.20. *Find the most general vector* $\mathbf{r}$ *that satisfies the equation* $\mathbf{a} \cdot (\mathbf{r} \times \mathbf{b}) = 0$*, where* $\mathbf{a}$ *and* $\mathbf{b}$ *are nonzero, nonparallel vectors.*

Solution: By the algebraic property of the triple product, $\mathbf{a}\cdot(\mathbf{r}\times\mathbf{b}) = \mathbf{r}\cdot(\mathbf{b}\times\mathbf{a}) = 0$. Hence, $\mathbf{r} \perp \mathbf{a}\times\mathbf{b}$. The vector $\mathbf{r}$ lies in the plane parallel to both $\mathbf{a}$ and $\mathbf{b}$ because $\mathbf{a} \times \mathbf{b}$ is orthogonal to these vectors. Any vector in the plane is a linear combination of any two nonparallel vectors in it: $\mathbf{r} = t\mathbf{a} + s\mathbf{b}$ for any real t and s (see Study Problem 11.6). □

Problem 11.21. (Volume of a Tetrahedron). *A tetrahedron is a solid with four vertices and four triangular faces. Its volume $V = \frac{1}{3}Ah$, where h is the distance from a vertex to the opposite face and A is the area of that face. Given coordinates of the vertices B, C, D, and P, express the volume of the tetrahedron through them.*

Solution: Put $\mathbf{b} = \vec{BC}$, $\mathbf{c} = \vec{BD}$, and $\mathbf{a} = \vec{AP}$. The area of the triangle BCD is $A = \frac{1}{2}\|\mathbf{b} \times \mathbf{c}\|$. The distance from P to the plane $\mathcal{P}_1$ containing the face BCD is the distance between $\mathcal{P}_1$ and the parallel plane $\mathcal{P}_2$ through the vertex P. Hence,

$$V = \frac{1}{3}\,A\,\frac{|\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})|}{\|\mathbf{b}\times\mathbf{c}\|} = \frac{1}{6}\,|\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})|\,.$$

So the volume of a tetrahedron with adjacent sides $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ is one-sixth the volume of the parallelepiped with the same adjacent sides. Note the result does not depend on the choice of a vertex. Any vertex could have been chosen instead of B in the above solution. □

75.4. Exercises. **(1)** Determine whether the points $A = (1, 2, 3)$, $B = (1, 0, 1)$, $C = (-1, 1, 2)$, and $D = (-2, 1, 0)$ are in one plane and, if not, find the volume of the parallelepiped with adjacent edges AB, AC, and AD.

(2) Find

(i) all values of s at which the points $A(s, 0, s)$, $B(1, 0, 1)$, $C(s, s, 1)$, and $D(0, 1, 0)$ are in the same plane

(ii) all values of s at which the volume of the parallelepiped with adjacent sides AB, AC, and AD is 9 units

(3) Verify whether the vectors $\mathbf{a} = \hat{\mathbf{e}}_1 + 2\hat{\mathbf{e}}_2 - \hat{\mathbf{e}}_3$, $\mathbf{b} = 2\hat{\mathbf{e}}_1 - \hat{\mathbf{e}}_2 + \hat{\mathbf{e}}_3$, and $\mathbf{c} = 3\hat{\mathbf{e}}_1 + \hat{\mathbf{e}}_2 - 2\hat{\mathbf{e}}_3$ are coplanar.

(4) Let the numbers u, v, and w be such that $uvw = 1$ and $u^3 + v^3 + w^3 = 1$. Are the vectors $\mathbf{a} = u\hat{\mathbf{e}}_1 + v\hat{\mathbf{e}}_2 + w\hat{\mathbf{e}}_3$, $\mathbf{b} = v\hat{\mathbf{e}}_1 + w\hat{\mathbf{e}}_2 + u\hat{\mathbf{e}}_3$, and $\mathbf{c} = w\hat{\mathbf{e}}_1 + u\hat{\mathbf{e}}_2 + v\hat{\mathbf{e}}_3$ coplanar? If not, what is the volume of the parallelepiped with adjacent edges $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$?

(5) Prove that

$$(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d}) = \det \begin{pmatrix} \mathbf{a} \cdot \mathbf{c} & \mathbf{b} \cdot \mathbf{c} \\ \mathbf{a} \cdot \mathbf{d} & \mathbf{b} \cdot \mathbf{d} \end{pmatrix}.$$

Hint: Use the invariance of the triple product under a cyclic permutation of vectors in it and Study Problem 11.16.

(6) Let a set $\mathcal{S}_1$ be the circle $x^2 + y^2 = 1$ and let a set $\mathcal{S}_2$ be the line through the points $(0, 2)$ and $(2, 0)$. What is the distance between the sets $\mathcal{S}_1$ and $\mathcal{S}_2$?

(7) Consider a plane through three points $A = (1, 2, 3)$, $B = (2, 3, 1)$, and $C = (3, 1, 2)$. Find the distance between the plane and a point P obtained from A by moving the latter 3 units of length along the vector $\mathbf{a} = \langle -1, 2, 2 \rangle$.

(8) Consider two lines. The first line passes through the points $(1, 2, 3)$ and $(2, -1, 1)$, while the other passes through the points $(-1, 3, 1)$ and $(1, 1, 3)$. Find the distance between the lines.

(9) Find the distance between the line through the points $(1, 2, 3)$ and $(2, 1, 4)$ and the plane through the points $(1, 1, 1)$, $(3, 1, 2)$, and $(1, 2, -1)$.

Hint: If the line is not parallel to the plane, then they intersect and the distance is 0. So check first whether the line is parallel to the plane. How can this be done?

(10) Consider the line through the points $(1, 2, 3)$ and $(2, 1, 2)$. If a second line passes through the points $(1, 1, s)$ and $(2, -1, 0)$, find all values of s, if any, at which the distance between the lines is $9/2$ units.

76. Planes in Space

76.1. A Geometrical Description of a Plane in Space. Let a plane $\mathcal{P}$ go through a point P_0. Clearly, there are many planes that contain a particular point in space. All such planes can be obtained from a particular plane by a general rotation about the point P_0. To eliminate this freedom and define the plane uniquely, one can demand that every line in the plane be perpendicular to a given vector $\mathbf{n}$. This vector is called a *normal* of the plane $\mathcal{P}$. Thus, the geometrical description of a plane $\mathcal{P}$ in space entails specifying a point P_0 that belongs to $\mathcal{P}$ and a normal $\mathbf{n}$ of $\mathcal{P}$.

76.2. An Algebraic Description of a Plane in Space. Let a plane $\mathcal{P}$ be defined by a point P_0 that belongs to it and a normal $\mathbf{n}$. In some coordinate system, the point P_0 has coordinates (x_0, y_0, z_0) and the vector $\mathbf{n}$ is specified by its components $\mathbf{n} = \langle n_1, n_2, n_3 \rangle$. A generic point in space P has coordinates (x, y, z). An algebraic description of a plane amounts to specifying conditions on the variables (x, y, z) such that the point $P(x, y, z)$ belongs to the plane $\mathcal{P}$. Let $\mathbf{r}_0 = \langle x_0, y_0, z_0 \rangle$ and $\mathbf{r} = \langle x, y, z \rangle$ be the position vectors of a particular point P_0 in the plane and a generic point P in space, respectively. Consider the vector $\vec{P_0P} = \mathbf{r} - \mathbf{r}_0 = \langle x - x_0, y - y_0, z - z_0 \rangle$. This vector lies in the plane $\mathcal{P}$ if and only if it is orthogonal to the normal $\mathbf{n}$, according to the geometrical description of a plane (see Figure 11.15, left panel). The algebraic condition equivalent to the geometrical one, $\mathbf{n} \perp \vec{P_0P}$, reads $\mathbf{n} \cdot \vec{P_0P} = 0$. Thus, the following theorem has just been proved.

THEOREM 11.8. (Equation of a Plane).
A point with coordinates (x, y, z) belongs to a plane through a point $P_0(x_0, y_0, z_0)$ and normal to a vector $\mathbf{n} = \langle n_1, n_2, n_3 \rangle$ if and only if

$$n_1(x - x_0) + n_2(y - y_0) + n_3(z - z_0) = 0 \qquad \text{or} \qquad \mathbf{n} \cdot \mathbf{r} = \mathbf{n} \cdot \mathbf{r}_0,$$

where $\mathbf{r}$ and $\mathbf{r}_0$ are position vectors of a generic point and a particular point P_0 in the plane.

So a general solution of the equation $\mathbf{n} \cdot \mathbf{r} = d$, where $\mathbf{n}$ is a given vector and d is a given number, is a set of position vectors of all points of the plane that is perpendicular to $\mathbf{n}$. The number d determines the position of the plane in space in the following way. If $\mathbf{r}_0$ is the position vector of a particular point in the plane, then $d = \mathbf{n} \cdot \mathbf{r}_0$. The position vector of another point in the very same plane is $\mathbf{r}_0 + \mathbf{a}$, where the vector $\mathbf{a}$ is in the plane (a particular point P_0 has just been displaced in the plane along the vector $\mathbf{a}$). The number d is independent of the

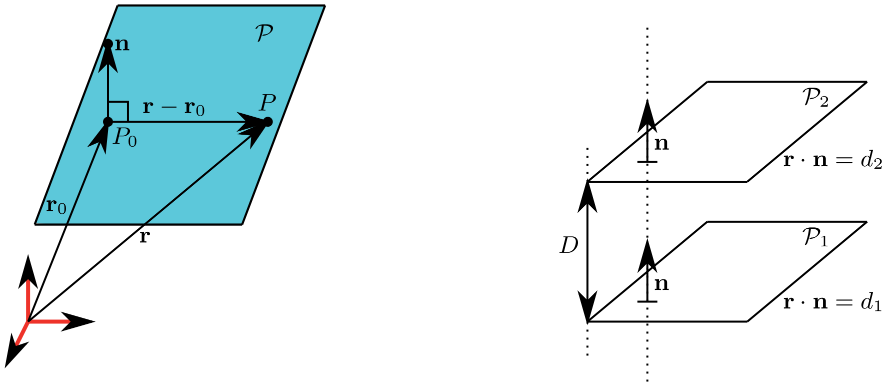

FIGURE 11.15. **Left**: Algebraic description of a plane. If $\mathbf{r}_0$ is a position vector of a particular point in the plane and $\mathbf{r}$ is the position vector of a generic point in the plane, then the vector $\mathbf{r}-\mathbf{r}_0$ lies in the plane and is perpendicular to its normal, that is, $\mathbf{n}\cdot(\mathbf{r}-\mathbf{r}_0)=0$. **Right**: Equations of parallel planes differ only by their constant terms. The difference of the constant terms determines the distance between the planes as stated in (11.12).

choice of a particular point in the plane because $d=\mathbf{n}\cdot(\mathbf{r}_0+\mathbf{a})=\mathbf{n}\cdot\mathbf{r}_0$ and the vectors $\mathbf{n}$ and $\mathbf{a}$ are orthogonal, $\mathbf{n}\perp\mathbf{a}$. The number d changes if the point P_0 is moved along the normal $\mathbf{n}$, but the result of such a displacement of P_0 is a point that is not in the original plane. Thus, the equations $\mathbf{n}\cdot\mathbf{r}=d_1$ and $\mathbf{n}\cdot\mathbf{r}=d_2$ describe two *parallel* planes if $d_1\neq d_2$; that is, variations of d correspond to shifts of the plane parallel to itself along its normal (see Figure 11.15, right panel).

Note also that the normal vector of a given plane is not uniquely defined because its magnitude is irrelevant for the geometrical description of the plane. If $\mathbf{n}$ is a normal, then $s\mathbf{n}$ is also a normal of the same plane for any nonzero real s. In the algebraic approach, the scaling of $\mathbf{n}$ does not change the equation of the plane, $(s\mathbf{n})\cdot\mathbf{r}=(s\mathbf{n})\cdot\mathbf{r}_0$, or, by cancelling the scaling factor s in this equation, $\mathbf{n}\cdot\mathbf{r}=\mathbf{n}\cdot\mathbf{r}_0$. Thus, two planes are parallel if their normals are parallel. From the algebraic point of view, two planes are parallel if their normals are proportional:

$$\mathcal{P}_1\parallel\mathcal{P}_1 \quad\longleftrightarrow\quad \mathbf{n}_1\parallel\mathbf{n}_2 \quad\longleftrightarrow\quad \mathbf{n}_1=s\mathbf{n}_2$$

for some real s.

Definition 11.16. (Angle Between Two Planes).
The angle between the normals of two planes is called the angle between the planes.

If $\mathbf{n}_1$ and $\mathbf{n}_2$ are the normals, then the angle θ between them is determined by

$$\cos\theta = \frac{\mathbf{n}_1 \cdot \mathbf{n}_2}{\|\mathbf{n}_1\|\,\|\mathbf{n}_2\|} = \hat{\mathbf{n}}_1 \cdot \hat{\mathbf{n}}_2 .$$

Note that a plane as a geometrical set of points in space is not changed if the direction of its normal is reversed (i.e., $\mathbf{n} \to -\mathbf{n}$). So the range of θ can always be restricted to the interval $[0, \pi/2]$. Indeed, if θ happens to be in the interval $[\pi/2, \pi]$ (i.e., $\cos\theta \leq 0$), then the angle $\theta - \pi/2$ can also be viewed as the angle between the planes because one can always reverse the direction of one of the normals $\mathbf{n}_1 \to -\mathbf{n}_1$ or $\mathbf{n}_2 \to -\mathbf{n}_2$ so that $\cos\theta \to -\cos\theta$.

The planes are perpendicular if their normals are perpendicular. For example, the planes $x + y + z = 1$ and $x + 2y - 3z = 4$ are perpendicular because their normals $\mathbf{n}_1 = \langle 1, 1, 1\rangle$ and $\mathbf{n}_2 = \langle 1, 2, -3\rangle$ are perpendicular: $\mathbf{n}_1 \cdot \mathbf{n}_2 = 1 + 2 - 3 = 0$ (i.e., $\mathbf{n}_1 \perp \mathbf{n}_2$).

Example 11.8. *Find an equation of the plane through three given points $A(1, 1, 1)$, $B(2, 3, 0)$, and $C(-1, 0, 3)$.*

Solution: A plane is specified by a particular point P_0 in it and by a vector $\mathbf{n}$ normal to it. Three points on the plane are given, so any of them can be taken as P_0, for example, $P_0 = A$ or $(x_0, y_0, z_0) = (1, 1, 1)$. A vector normal to a plane can be found as the cross product of any two nonparallel vectors in that plane (see Figure 11.16, left panel). So put $\mathbf{a} = \vec{AB} = \langle 1, 2, -1\rangle$ and $\mathbf{b} = \vec{AC} = \langle -2, -1, 2\rangle$. Then one can take $\mathbf{n} = \mathbf{a} \times \mathbf{b} = \langle 3, 0, -3\rangle$. An equation of the plane is $3(x - 1) + 0(y - 1) + (-3)(z - 1) = 0$, or $x - z = 0$. Since the equation does not contain the variable y, the plane is parallel to the y axis. Note that if the y component of $\mathbf{n}$ vanishes (i.e., there is no y in the equation), then $\mathbf{n}$ is orthogonal to $\hat{\mathbf{e}}_2$ because $\mathbf{n} \cdot \hat{\mathbf{e}}_2 = 0$; that is, the y axis is perpendicular to $\mathbf{n}$ and hence parallel to the plane. □

76.3. The Distance Between a Point and a Plane. Consider the plane through a point P_0 and normal to a vector $\mathbf{n}$. Let P_1 be a point in space. What is the distance between P_1 and the plane? Let the angle between $\mathbf{n}$ and the vector $\vec{P_0P_1}$ be θ (see Figure 11.16, right panel). Then the distance in question is $D = \|\vec{P_0P_1}\| \cos\theta$ if $\theta \leq \pi/2$ (the length of the straight line segment connecting P_1 and the plane along

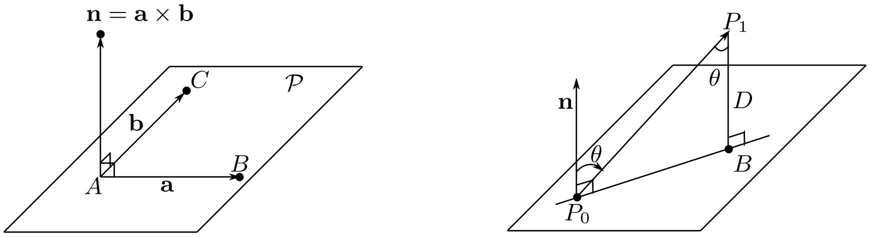

FIGURE 11.16. **Left**: Illustration to Example 11.8. The cross product of two non parallel vectors in a plane is a normal of the plane.
Right: Distance between a point P_1 and a plane. An illustration to the derivation of the distance formula (11.10). The segment P_1B is parallel to the normal $\mathbf{n}$ so that the triangle P_0P_1B is right-angled. Therefore, $D = |P_1B| = |P_0P_1|\cos\theta$.

the normal $\mathbf{n}$). For $\theta > \pi/2$, $\cos\theta$ must be replaced by $-\cos\theta$ because $D \geq 0$. So

$$(11.10) \qquad D = \|\vec{P_0P_1}\||\cos\theta| = \frac{\|\mathbf{n}\|\|\vec{P_0P_1}\||\cos\theta|}{\|\mathbf{n}\|} = \frac{|\mathbf{n}\cdot\vec{P_0P_1}|}{\|\mathbf{n}\|}\,.$$

Note that this distance formula can be obtained from the distance between two parallel planes (Corollary 11.7). Indeed, the vector $\vec{AB} \times \vec{AC}$ is the cross product of two vectors in the plane and hence can be used as the normal $\mathbf{n}$, whereas the vector $\vec{AP}$ can be used as $\vec{P_0P_1}$.

Let $\mathbf{r}_0$ and $\mathbf{r}_1$ be position vectors of P_0 and P_1, respectively. Then $\vec{P_0P_1} = \mathbf{r}_1 - \mathbf{r}_0$, and

$$(11.11) \qquad D = \frac{|\mathbf{n}\cdot(\mathbf{r}_1 - \mathbf{r}_0)|}{\|\mathbf{n}\|} = \frac{|\mathbf{n}\cdot\mathbf{r}_1 - d|}{\|\mathbf{n}\|},$$

which is a bit more convenient for calculating the distance if the plane is defined algebraically by an equation $\mathbf{n}\cdot\mathbf{r} = d$.

76.3.1. Distance Between Parallel Planes. Equation (11.11) allows us to obtain a simple formula for the distance between two parallel planes defined by the equations $\mathbf{n}\cdot\mathbf{r} = d_1$ and $\mathbf{n}\cdot\mathbf{r} = d_2$ (see Figure 11.15, right panel):

$$(11.12) \qquad D = \frac{|d_1 - d_2|}{\|\mathbf{n}\|}\,.$$

Indeed, the distance between two parallel planes is the distance between the first plane and any point $\mathbf{r}_0$ in the second plane. By (11.11), this distance is $D = |\mathbf{n}\cdot\mathbf{r}_1 - d_1|/\|\mathbf{n}\| = |d_2 - d_1|/\|\mathbf{n}\|$ because $\mathbf{n}\cdot\mathbf{r}_0 = d_2$ for any point in the second plane.

EXAMPLE 11.9. *Find an equation of a plane that is parallel to the plane $2x - y + 2z = 2$ and at a distance of 3 units from it.*

SOLUTION: There are a few ways to solve this problem. From the geometrical point of view, a plane is defined by a particular point in it and its normal. Since the planes are parallel, they must have the same normal $\mathbf{n} = \langle 2, -1, 2\rangle$. Note that the coefficients at the variables in the plane equation define the components of the normal vector. Therefore, the problem is reduced to finding a particular point. Let P_0 be a particular point on the given plane. Then a point on a parallel plane can be obtained from it by shifting P_0 by a distance of 3 units along the normal $\mathbf{n}$. If $\mathbf{r}_0$ is the position vector of P_0, then a point on a parallel plane has a position vector $\mathbf{r}_0 + s\mathbf{n}$, where the displacement vector $s\mathbf{n}$ must have a length of 3, or $\|s\mathbf{n}\| = |s|\|\mathbf{n}\| = 3|s| = 3$ and therefore $s = \pm 1$. Naturally, there should be two planes parallel to the given one and at the same distance from it. To find a particular point on the given plane, one can set two coordinates to 0 and find the value of the third coordinate from the equation of the plane. Take, for instance, $P_0(1, 0, 0)$. Particular points on the parallel planes are $\mathbf{r}_0 + \mathbf{n} = \langle 1, 0, 0\rangle + \langle 2, -1, 2\rangle = \langle 3, -1, 2\rangle$ and, similarly, $\mathbf{r}_0 - \mathbf{n} = \langle -1, 1, -2\rangle$. Using these points in the standard equation of a plane, the equations of two parallel planes are obtained:

$$2x - y + 2z = 11 \quad \text{and} \quad 2x - y + 2z = -7.$$

An alternative algebraic solution is based on the distance formula (11.12) for parallel planes. An equation of a plane parallel to the given one should have the form $2x - y + 2z = d$. The number d is determined by the condition that $|d - 2|/\|\mathbf{n}\| = 3$ or $|d - 2| = 9$, or $d = \pm 9 + 2$. □

76.4. Study Problems.

Problem 11.22. *Find an equation of the plane that is normal to a straight line segment AB and bisects it if $A = (1, 1, 1)$ and $B = (-1, 3, 5)$.*

SOLUTION: One has to find a particular point in the plane and its normal. Since AB is perpendicular to the plane, $\mathbf{n} = \vec{AB} = \langle -2, 2, 4\rangle$. The midpoint of the segment lies in the plane. Hence, $P_0(0, 2, 3)$ (the coordinates of the midpoints are the half sums of the corresponding coordinates of the endpoints). The equation reads $-2x + 2(y - 2) + 4(z - 3) = 0$ or $-x + y + 2z = 8$. □

Problem 11.23. *Find an equation of the plane through the point $P_0(1,2,3)$ that is perpendicular to the planes $x+y+z=1$ and $x-y+2z=1$.*

Solution: One has to find a particular point in the plane and any vector perpendicular to it. The first part of the problem is easy to solve: P_0 is given. Let $\mathbf{n}$ be a normal of the plane in question. Then, from the geometrical description of a plane, it follows that $\mathbf{n}\perp\mathbf{n}_1=\langle 1,1,1\rangle$ and $\mathbf{n}\perp\mathbf{n}_2=\langle 1,-1,2\rangle$, where $\mathbf{n}_1$ and $\mathbf{n}_2$ are normals of the given planes. So $\mathbf{n}$ is a vector perpendicular to two given vectors. By the geometrical property of the cross product, such a vector can be constructed as $\mathbf{n}=\mathbf{n}_1\times\mathbf{n}_2=\langle 3,-1,-2\rangle$. Hence, the equation reads $3(x-1)-(y-2)-2(z-3)=0$ or $3x-y-2z=-5$. □

Problem 11.24. *Determine whether two planes $x+2y-2z=1$ and $2x+4y+4z=10$ are parallel and, if not, find the angle between them.*

Solution: The normals are $\mathbf{n}_1=\langle 1,2,-2\rangle$ and $\mathbf{n}_2=\langle 2,4,4\rangle=2\langle 1,2,2\rangle$ (i.e., they are not proportional). Hence, the planes are not parallel. Since $\|\mathbf{n}_1\|=3$, $\|\mathbf{n}_1\|=6$, and $\mathbf{n}_1\cdot\mathbf{n}_2=2$, the angle is determined by $\cos\theta=2/18=1/9$ or $\theta=\cos^{-1}(1/9)$. □

76.5. Exercises. **(1)** Find an equation of the plane through the origin and parallel to the plane $2x-2y+z=4$. What is the distance between the two planes?

(2) Do the planes $2x+y-z=1$ and $4x+2y-2z=10$ intersect?

(3) Consider a parallelepiped with one vertex at the origin at which the adjacent sides are the vectors $\mathbf{a}=\langle 1,2,3\rangle$, $\mathbf{b}=\langle 2,1,1\rangle$, and $\mathbf{c}=\langle -1,0,1\rangle$. Find equations of the planes that contain the faces of the parallelepiped.

(4) Determine whether the planes $2x+y-z=3$ and $x+y+z=1$ are intersecting. If they are, find the angle between them.

(5) Find an equation of the plane with x intercept a, y intercept b, and z intercept c. What is the distance between the origin and the plane?

(6) Find an equation for the set of points that are equidistant from the points $(1,2,3)$ and $(-1,2,1)$. Give a geometrical description of the set.

(7) Find an equation of the plane that is perpendicular to the plane $x+y+z=1$ and contains the line through the points $(1,2,3)$ and $(-1,1,0)$.

(8) To which of the planes $x+y+z=1$ and $x+2y-z=2$ is the point $(1,2,3)$ the closest?

(9) Give a geometrical description of the following families of planes:
(i) $x + y + z = c$
(ii) $x + y + cz = 1$
(iii) $x \sin c + y \cos c + z = 1$
where c is a parameter.

(10) Consider three planes with normals $\mathbf{n}_1$, $\mathbf{n}_2$, and $\mathbf{n}_3$ such that each pair of the planes is intersecting. Under what condition on the normals are the three lines of intersection parallel or even coincide?

77. Lines in Space

77.1. A Geometrical Description of a Line in Space. Consider two points in space. They can be connected by a path. Among all the continuous paths that connect the two points, there is a distinct one, namely, the one that has the smallest length. This path is called a *straight line segment.*

Definition 11.17. (Geometrical Description of a Line).
A line $\mathcal{L}$ is a set of points in space such that the shortest path connecting any pair of points of $\mathcal{L}$ belongs to $\mathcal{L}$.

Given a point P_0, there are infinitely many lines through P_0, all of which are related by rigid rotations about the point P_0. Therefore, to fix a line uniquely, one should specify a direction to which the line is parallel, in addition to its position P_0. The direction can be determined by a vector $\mathbf{v}$. It follows from the geometrical description of a line that $\mathbf{v}$ is a vector connecting any two points of the line. The length or norm of $\mathbf{v}$ is irrelevant for specifying the direction; that is, any parallel vector (or the oriented segment between another pair of points of the line) is just as good as $\mathbf{v}$. Thus, a line $\mathcal{L}$ is uniquely specified by a particular point P_0 of $\mathcal{L}$ and any vector $\mathbf{v}$ to which the line is parallel, $\mathbf{v} \parallel \mathcal{L}$.

Remark. The very notion of a line, defined as the shortest path between two points in space, is deeply rooted in the very structure of space itself. How can a line be realized in the space in which we live? One can use a piece of rope, as in the ancient world, or the "line of sight" (i.e., the path traveled by light from one point to another). Einstein's theory of gravity states that "straight lines" defined as trajectories traversed by light are not exactly the same as "straight lines" in a Euclidean space. So a Euclidean space may only be viewed as a mathematical approximation (or model) of our space. A good analogy would be to compare the shortest paths in a plane and on the surface of a sphere; they are not the same as the latter are "curved." The concept of curvature of a path is discussed in the next chapter. The shortest path between two points in a space is called a *geodesic* (by analogy with the shortest path on the surface of the Earth). The geodesics of a Euclidean space (straight lines) do not have curvature, whereas the geodesics of our space (i.e., the paths traversed by light) do have curvature that is determined by the distribution of gravitating masses (planets, stars, etc.). A deviation of the geodesics from straight lines near the surface of the Earth is very hard to notice. However, a deviation of the trajectory of light from a straight line has been observed for the light coming from a distant star to the Earth and passing near the Sun. Einstein's theory of general relativity asserts that a better model of our space is a *Riemann space*. A sufficiently small neighborhood in a Riemann space looks like a portion of a Euclidean space.

77.2. An Algebraic Description of a Line. In some coordinate system, a particular point of a line $\mathcal{L}$ has coordinates $P_0(x_0, y_0, z_0)$, and a vector parallel to $\mathcal{L}$ is defined by its components, $\mathbf{v} = \langle v_1, v_2, v_3\rangle$. Let $\mathbf{r} = \langle x, y, z\rangle$ be a position vector of a generic point of $\mathcal{L}$ and let $\mathbf{r}_0 = \langle x_0, y_0, z_0\rangle$ be the position vector of P_0. The vector $\mathbf{r} - \mathbf{r}_0$ must be parallel to $\mathcal{L}$ because it is the oriented straight line segment connecting two points of $\mathcal{L}$ (see Figure 11.17, left panel). Hence, a point (x, y, z) belongs to $\mathcal{L}$ if and only if $\mathbf{r} - \mathbf{r}_0 \parallel \mathbf{v}$. The equivalent algebraic condition reads $\mathbf{r} - \mathbf{r}_0 = t\mathbf{v}$ for some real t. To obtain all points of $\mathcal{L}$, one should let t range over all real numbers.

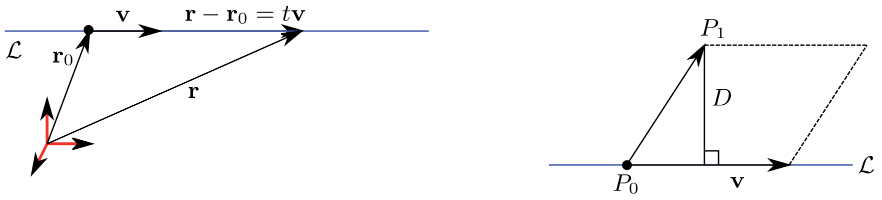

FIGURE 11.17. **Left**: Algebraic description of a line $\mathcal{L}$ through $\mathbf{r}_0$ and parallel to a vector $\mathbf{v}$. If $\mathbf{r}_0$ and $\mathbf{r}$ are position vectors of particular and generic points of the line, then the vector $\mathbf{r}-\mathbf{r}_0$ is parallel to the line and hence must be proportional to a vector $\mathbf{v}$, that is, $\mathbf{r} - \mathbf{r}_0 = t\mathbf{v}$ for some real number t.
Right: Distance between a point P_1 and a line $\mathcal{L}$ through a point P_0 and parallel to a vector $\mathbf{v}$. It is the height of the parallelogram whose adjacent sides are the vectors $\overrightarrow{P_0P_1}$ and $\mathbf{v}$.

THEOREM 11.9. (Equations of a Line).
The coordinates of the points of the line $\mathcal{L}$ through a point P_0 and parallel to a vector $\mathbf{v}$ *satisfy the vector equation*

$$\mathbf{r} = \mathbf{r}_0 + t\mathbf{v}\,, \quad -\infty < t < \infty. \tag{11.13}$$

or the parametric equations

$$x = x_0 + tv_1\,, \quad y = y_0 + tv_2\,, \quad x = z_0 + tv_3\,, \quad -\infty < t < \infty. \tag{11.14}$$

The parametric equations of the line can be solved for t. As a result, one infers equations for the coordinates x, y, and z:

$$t = \frac{x - x_0}{v_1} = \frac{y - y_0}{v_2} = \frac{z - z_0}{v_3}\,,$$

provided none of the components of $\mathbf{v}$ vanish. These equations are called *symmetric* equations of a line. Note that these equations make sense only if all the components of $\mathbf{v}$ do not vanish. If, say, $v_1 = 0$, then the first equation in (11.14) does not contain the parameter t at all. So the symmetric equations are written in the form

$$x = x_0\,, \quad \frac{y - y_0}{v_2} = \frac{z - z_0}{v_3}.$$

EXAMPLE 11.10. *Find the vector, parametric, and symmetric equations of the line through the points* $A(1, 1, 1)$ *and* $B(1, 2, 3)$.

SOLUTION: Take $\mathbf{v} = \vec{AB} = \langle 0, 1, 2\rangle$ and $P_0 = A$. Then

$$\begin{aligned} &\mathbf{r} = \langle 1, 1, 1\rangle + t\langle 0, 1, 2\rangle, \\ &x = 1\,, \quad y = 1 + t\,, \quad z = 1 + 2t\,, \\ &x = 1\,, \quad y - 1 = \frac{z - 1}{2} \end{aligned}$$

are the vector, parametric, and symmetric equations of the line, respectively. □

77.2.1. Distance Between a Point and a Line. Consider a line $\mathcal{L}$ through P_0 parallel to $\mathbf{v}$. What is the distance between a given point P_1 and the line $\mathcal{L}$? Consider a parallelogram with vertex P_0 and whose adjacent sides are the vectors $\mathbf{v}$ and $\vec{P_0P_1}$ as depicted in Figure 11.17 (right panel). The distance in question is the height of this parallelogram, which is therefore its area divided by the length of the base $\|\mathbf{v}\|$. If $\mathbf{r}_0$ and $\mathbf{r}_1$ are position vectors of P_0 and P_1, then $\vec{P_0P_1} = \mathbf{r}_1 - \mathbf{r}_0$ and hence

$$D = \frac{\|\mathbf{v} \times \vec{P_0P_1}\|}{\|\mathbf{v}\|} = \frac{\|\mathbf{v} \times (\mathbf{r}_1 - \mathbf{r}_0)\|}{\|\mathbf{v}\|}.$$

77.3. Relative Positions of Lines in Space. Two lines in space can be intersecting, parallel, or skew. Given an algebraic description of the lines, how can one find out which of the above three cases occurs? If the lines are intersecting, how can one find the coordinates of the point of intersection? Suppose the line $\mathcal{L}_1$ contains a point P_1 and is parallel to $\mathbf{v}_1$, while $\mathcal{L}_2$ contains a point P_2 and is parallel to $\mathbf{v}_2$.

COROLLARY 11.6. (Criterion for Two Lines to Parallel).
Two lines are parallel if and only if their direction vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are parallel:

$$\mathcal{L}_1 \parallel \mathcal{L}_2 \quad \Longleftrightarrow \quad \mathbf{v}_1 \parallel \mathbf{v}_2 \quad \Longleftrightarrow \quad \mathbf{v}_1 = s\mathbf{v}_2 \ \text{ for some real } s.$$

Suppose that the lines are not parallel. Then they are either skew or intersecting. In the latter case, the distance between the lines is 0 as they have a common point (see Section 75), whereas in the former case the distance cannot be 0. Since the lines are not parallel, $\mathbf{v}_1 \times \mathbf{v}_2 \neq \mathbf{0}$. Making use of the distance formula between skew lines (see Corollary 11.5), one proves the following.

COROLLARY 11.7. (Criteria for Two Non-parallel Lines to Be Skew or Intersecting).
Let P_1 be a point of $\mathcal{L}_1$ and let P_2 be a point of $\mathcal{L}_2$. Let $\mathbf{v}_1\|\mathcal{L}_1$ and $\mathbf{v}_2\|\mathcal{L}_2$ and let the lines $\mathcal{L}_1$ and $\mathcal{L}_2$ be nonparallel. Then

$$\begin{aligned} \vec{P_1P_2}\cdot(\mathbf{v}_1\times\mathbf{v}_2) \neq 0 \quad &\Longleftrightarrow \quad \mathcal{L}_1\,,\mathcal{L}_2 \text{ are skew,}\\ \vec{P_1P_2}\cdot(\mathbf{v}_1\times\mathbf{v}_2) = 0 \quad &\Longleftrightarrow \quad \mathcal{L}_1\,,\mathcal{L}_2 \text{ are intersecting.} \end{aligned}$$

Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be intersecting. How can one find the point of intersection? To solve this problem, consider the vector equations for the lines $\mathbf{r}_t = \mathbf{r}_1 + t\mathbf{v}_1$ and $\mathbf{r}_s = \mathbf{r}_2 + s\mathbf{v}_2$. When changing the parameter t, the terminal point of $\mathbf{r}_t$ slides along the line $\mathcal{L}_1$, while the terminal point of $\mathbf{r}_s$ slides along the line $\mathcal{L}_2$ when changing the parameter s as depicted in Figure 11.18 (left panel). Note that the parameters of both lines are not related in any way according to the geometrical description of the lines. If two lines are intersecting, then there should exist a pair of numbers $(t,s) = (t_0,s_0)$ at which the terminal points of vectors $\mathbf{r}_t$ and $\mathbf{r}_s$ coincide, $\mathbf{r}_t = \mathbf{r}_s$. Let $\mathbf{v}_i = \langle a_i, b_i, c_i\rangle$, $i = 1,2$. Writing this vector equation in componentform, the following system of equations is obtained:

$$\begin{aligned} x_1 + ta_1 &= x_2 + sa_2,\\ y_1 + tb_1 &= x_2 + sb_2,\\ z_1 + tc_1 &= x_2 + sc_2. \end{aligned}$$

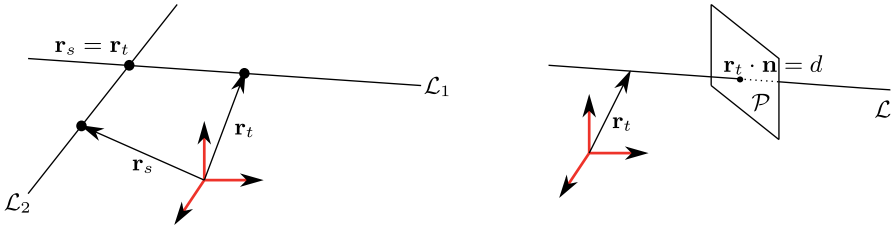

FIGURE 11.18. **Left**: Intersection point of two lines $\mathcal{L}_1$ and $\mathcal{L}_2$. The terminal point of the vector $\mathbf{r}_t$ traverses $\mathcal{L}_1$ as t ranges over all real numbers, while the terminal point of the vector $\mathbf{r}_s$ traverses $\mathcal{L}_2$ as s ranges over all real numbers independently of t. If the lines are intersecting, then there should exist a pair of numbers $(t, s) = (t_0, s_0)$ such that the vectors $\mathbf{r}_t$ and $\mathbf{r}_s$ coincide, which means that their components must be the same. This defines three equations on two variables t and s.
Right: Intersection point of a line $\mathcal{L}$ and a plane $\mathcal{P}$. The terminal point of the vector $\mathbf{r}_t$ traverses $\mathcal{L}$ as t ranges over all real numbers. If the line intersects the plane defined by the equation $\mathbf{r} \cdot \mathbf{n} = d$, then there should exist a particular value of t at which the vector $\mathbf{r}_t$ satisfies the equation of the plane: $\mathbf{r}_t \cdot \mathbf{n} = d$.

The system has three equations for only two variables. It is an *overdetermined* system, which may or may not have a solution. From the above geometrical analysis, it follows that, if the lines are parallel (i.e., $\mathbf{v}_1 \times \mathbf{v}_2 = \mathbf{0}$), then the system has no solution (the lines are distinct), or it might have infinitely many solutions (the lines coincide). For example, put $P_1 = P_2$ and $\mathbf{v}_1 = 2\mathbf{v}_2$. Then the system is satisfied by any pair $(t, s = 2t)$, where t is any real. The system has no solution if the criterion for two lines to be skew is satisfied. Finally, the system must have the only solution if the criterion for two nonparallel lines to be intersecting is satisfied. Let $(t, s) = (t_0, s_0)$ be a solution. Then the position vector of the point of intersection is $\mathbf{r}_1 + t_0\mathbf{v}_1$ or $\mathbf{r}_2 + s_0\mathbf{v}_2$.

77.4. Relative Positions of Lines and Planes. Consider a line $\mathbf{L}$ and a plane $\mathcal{P}$. The question of interest is to determine whether they are intersecting or parallel. If the line does not intersect the plane, then they must be parallel. In the latter case, the line must be perpendicular to the normal of the plane.

Corollary 11.8. (Criterion for a Line and a Plane to Be Parallel). *Let* $\mathbf{v}$ *be a vector parallel to a line* $\mathbf{L}$ *and let* $\mathbf{n}$ *be a normal of a plane* $\mathcal{P}$. *Then*

$$\mathcal{L} \parallel \mathcal{P} \quad \Longleftrightarrow \quad \mathbf{v} \parallel \mathbf{n} \quad \Longleftrightarrow \quad \mathbf{v}\cdot\mathbf{n} = 0.$$

If the line intersects the plane, then there should exist a particular value of the parameter t for which the position vector $\mathbf{r}_t = \mathbf{r}_0 + t\mathbf{v}$ of a point of $\mathcal{L}$ also satisfies the plane equation $\mathbf{r}\cdot\mathbf{n} = d$ (see Figure 11.18, right panel). The value of the parameter t that corresponds to the point of intersection is determined by the equation

$$\mathbf{r}_t\cdot\mathbf{n} = d \quad \Rightarrow \quad \mathbf{r}_0\cdot\mathbf{n} + t\mathbf{v}\cdot\mathbf{n} = d \quad \Rightarrow \quad t = \frac{d - \mathbf{r}_0\cdot\mathbf{n}}{\mathbf{v}\cdot\mathbf{n}}.$$

The position vector of the point of intersection is found by substituting this value of t into the vector equation of the line $\mathbf{r}_t = \mathbf{r}_0 + t\mathbf{v}$.

Example 11.11. *Find an equation of the plane* $\mathcal{P}$ *that is perpendicular to the plane* $\mathcal{P}_1$, $x + y - z = 1$, *and contains the line* $x - 1 = y/2 = z + 1$.

Solution: The plane $\mathcal{P}$ must be parallel to the line ($\mathcal{P}$ contains it) and the normal $\mathbf{n}_1 = \langle 1, 1, -1\rangle$ of $\mathcal{P}_1$ (as $\mathcal{P}\perp\mathcal{P}_1$). So the normal $\mathbf{n}$ of $\mathcal{P}$ is perpendicular to both $\mathbf{n}_1$ and the vector $\mathbf{v} = \langle 1, 2, 1\rangle$ that is parallel to the line. Therefore, one can take $\mathbf{n} = \mathbf{n}_1 \times \mathbf{v} = \langle 3, -2, 1\rangle$. To find a particular point of $\mathcal{P}$, note that the point of intersection of $\mathcal{P}_1$ and the line belongs to the plane $\mathcal{P}$. The line contains the point $(1, 0, -1)$. Put $\mathbf{r}_t = \mathbf{r}_0 + t\mathbf{v} = \langle 1 + t, 2t, -1 + t\rangle$. The equation $\mathbf{r}_t\cdot\mathbf{n}_1 = 1$ or $2 + 2t = 1$ has the solution $t = -1/2$. Hence, the position vector of a particular point of $\mathcal{P}$ is $\mathbf{r}_{t=-1/2} = \langle 1/2, -1, -3/2\rangle$. An equation of $\mathcal{P}$ reads $3(x + 1/2) - 2(y + 1) + (z + 3/2) = 0$ or $3x - 2y + z = -1$. □

77.5. Study Problems.

Problem 11.25. *Let* $\mathcal{L}_1$ *be the line through* $P_1(1, 1, 1)$ *and parallel to* $\mathbf{v}_1 = \langle 1, 2, -1\rangle$ *and let* $\mathcal{L}_2$ *be the line through* $P_2(4, 0, -2)$ *and parallel to* $\mathbf{v}_2 = \langle 2, 1, 0\rangle$. *Determine whether the lines are parallel, intersecting, or skew and find the line* $\mathcal{L}$ *that is perpendicular to both* $\mathcal{L}_1$ *and* $\mathcal{L}_2$ *and intersects them.*

Solution: The vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are not proportional, and hence the lines are not parallel. One has $\vec{P_1P_2} = \langle 3, -1, -3\rangle$ and $\mathbf{v}_1 \times \mathbf{v}_2 = \langle 1, -2, -3\rangle$. Therefore, $\vec{P_1P_2}\cdot(\mathbf{v}_1 \times \mathbf{v}_2) = 14 \neq 0$, and the lines are skew by Corollary 11.7. To find the line $\mathcal{L}$, note that it has to contain one point of each line. Let $\mathbf{r}_t = \mathbf{r}_1 + t\mathbf{v}_1$ be a position vector of a point of $\mathcal{L}_1$ and let $\mathbf{r}_s = \mathbf{r}_2 + s\mathbf{v}_2$ be a position vector of a point of $\mathcal{L}_2$ as

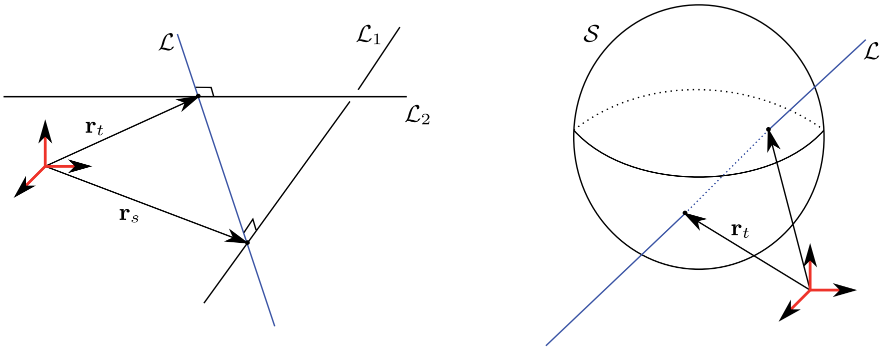

FIGURE 11.19. **Left**: Illustration to Study Problem 11.25. The vectors $\mathbf{r}_s$ and $\mathbf{r}_t$ trace out two given skewed lines $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively. There are particular values of t and s at which the distance $\|\mathbf{r}_t - \mathbf{r}_s\|$ becomes minimal. Therefore, the line $\mathcal{L}$ through such points $\mathbf{r}_t$ and $\mathbf{r}_s$ is perpendicular to both $\mathcal{L}_1$ and $\mathcal{L}_2$.
Right: Intersection of a line $\mathcal{L}$ and a sphere $\mathcal{S}$. An illustration to Study Problem 11.26. The terminal point of the vector $\mathbf{r}_t$ traverses the line as t ranges over all real numbers. If the line intersects the sphere, then there should exist a particular value of t at which the components of the vector $\mathbf{r}_t$ satisfy the equation of the sphere. This equation is quadratic in t, and hence it can have two distinct real roots, one multiple real root, or no real roots. These three cases correspond to two, one, or no points of intersection. One intersection point means that the line is tangent to the sphere.

shown in Figure 11.19 (left panel). As the line $\mathcal{L}$ should intersect both $\mathcal{L}_1$ and $\mathcal{L}_2$, there should exist a pair of values (t, s) of the parameters at which the vector $\mathbf{r}_s - \mathbf{r}_t$ is parallel to $\mathcal{L}$; that is, the vector $\mathbf{r}_s - \mathbf{r}_t$ becomes perpendicular to both vectors $\mathbf{v}_1$ and $\mathbf{v}_2$. The corresponding algebraic conditions are

$$\begin{aligned} \mathbf{r}_s - \mathbf{r}_t \perp \mathbf{v}_1 \quad &\Longleftrightarrow \quad (\mathbf{r}_s - \mathbf{r}_t) \cdot \mathbf{v}_1 = 4 + 4s - 6t = 0, \\ \mathbf{r}_s - \mathbf{r}_t \perp \mathbf{v}_1 \quad &\Longleftrightarrow \quad (\mathbf{r}_s - \mathbf{r}_t) \cdot \mathbf{v}_2 = 5 + 5s - 4t = 0. \end{aligned}$$

This system has the solution $t = 0$ and $s = -1$. Thus, the points with the position vectors $\mathbf{r}_{t=0} = \mathbf{r}_1$ and $\mathbf{r}_{s=-1} = \mathbf{r}_2 - \mathbf{v}_2 = \langle 2, -1, -2 \rangle$ belong to $\mathcal{L}$. So the vector $\mathbf{v} = \mathbf{r}_{s=-1} - \mathbf{r}_{t=0} = \langle 1, -3, -1 \rangle$ is parallel to $\mathcal{L}$.

Taking a particular point of $\mathcal{L}$ to be P_1, the parametric equations read $x = 1 + t$, $y = 1 - 3t$, $z = 1 - t$. □

Problem 11.26. *Consider a line through the origin that is parallel to the vector $\mathbf{v} = \langle 1, 1, 1\rangle$. Find the portion of this line that lies inside the sphere $x^2 + y^2 + z^2 - x - 2y - 3z = 9$.*

Solution: The parametric equations of the line are $x = t$, $y = t$, $z = t$. If the line intersects the sphere, then there should exist particular values of t at which the coordinates of a point of the line also satisfy the sphere equation (see Figure 11.19, right panel). In general, parametric equations of a line are linear in t, while a sphere equation is quadratic in the coordinates. Therefore, the equation that determines the values of t corresponding to the points of intersection is quadratic. A quadratic equation has two, one, or no real solutions. Accordingly, these cases correspond to two, one, and no points of intersection, respectively. In our case, $3t^2 - 6t = 9$ or $t^2 - 2t = 3$ and hence $t = -1$ and $t = 3$. The points of intersection are $(-1, -1, -1)$ and $(3, 3, 3)$. The line segment connecting them can be described by the parametric equations $x = t$, $y = t$, $z = t$, where $-1 \leq t \leq 3$. □

77.6. Exercises. **(1)** Find parametric equations of the line through the point $(1, 2, 3)$ and perpendicular to the plane $x + y + 2z = 1$. Find the point of intersection of the line and the plane.

(2) Find parametric and symmetric equations of the line of intersection of the planes $x + y + z = 1$ and $2x - 2y + z = 1$.

(3) Is the line through the points $(1, 2, 3)$ and $(2, -1, 1)$ perpendicular to the line through the points $(0, 1, -1)$ and $(1, 0, 2)$? Are the lines intersecting? If so, find the point of intersection.

(4) Determine whether the lines $x = 1 + 2t$, $y = 3t$, $z = 2 - t$ and $x + 1 = y - 4 = (z - 1)/3$ are parallel, skew, or intersecting. If they intersect, find the point of intersection.

(5) Find the vector equation of the straight line segment from the point $(1, 2, 3)$ to the point $(-1, 1, 2)$.

(6) Let $\mathbf{r}_1$ and $\mathbf{r}_2$ be position vectors of two points in space. Find the vector equation of the straight line segment from $\mathbf{r}_1$ to $\mathbf{r}_2$.

(7) Consider the plane $x + y + z = 0$ and a point $P = (1, 2, -3)$ in it. Find parametric equations of the lines through the origin that are at a distance of 1 unit from P.

(8) Find parametric, symmetric, and vector equations of the line through $(0, 1, 2)$ that is perpendicular to $\mathbf{v} = \langle 1, 2, 1\rangle$ and parallel to the plane $x + 2y + z = 3$.

(9) Find parametric equations of the line that is parallel to $\mathbf{v} = \langle 2, -1, 2\rangle$ and goes through the center of the sphere $x^2 + y^2 + z^2 = 2x + 6z - 6$. Restrict the range of the parameter to describe the portion of the line that is inside the sphere.

(10) Let the line $\mathcal{L}_1$ pass through the point $A(1, 1, 0)$ parallel to the vector $\mathbf{v} = \langle 1, -1, 2\rangle$ and let the line $\mathcal{L}_2$ pass through the point $B(2, 0, 2)$ parallel to the vector $\mathbf{w} = \langle -1, 1, 2\rangle$. Show that the lines are intersecting. Find the point C of intersection and parametric equations of the line $\mathcal{L}_3$ through C that is perpendicular to $\mathcal{L}_1$ and $\mathcal{L}_2$.

78. Quadric Surfaces

DEFINITION 11.18. (Quadric Surface).
The set of points whose coordinates in a rectangular coordinate system satisfy the equation

$$Ax^2 + By^2 + Cz^2 + pxy + qxz + vyz + \alpha x + \beta y + \gamma z + D = 0,$$

where A, B, C, p, q, v, α, β, γ, and D are real numbers, is called a quadric surface.

A sphere provides a simple example of a quadric surface: $x^2 + y^2 + z^2 - R^2 = 0$, that is, $A = B = C = 1$, $p = q = v = 0$, $\alpha = \beta = \gamma$, and $D = -R^2$, where R is the radius of the sphere. The equation that defines quadric surfaces is the most general equation *quadratic* in all the coordinates. This is why surfaces defined by it are called *quadric*. The task here is to classify all the shapes of quadric surfaces. The shape does not change under its rigid rotations and translations. On the other hand, the equation that describes the shape would change under rotations and translations of the coordinate system (see Section 71 and Example 11.2). The freedom in choosing the coordinate system can be used to simplify the equation for quadric surface and obtain a classification of different shapes described by it.

78.1. Quadric Cylinders. Consider first a simpler problem in which the equation of a quadric surface does not contain one of the coordinates, say, z (i.e., $C = q = v = \gamma = 0$). Then the set $\mathcal{S}$,

$$\mathcal{S} = \Big\{(x, y, z)\Big| Ax^2 + By^2 + pxy + \alpha x + \beta y + D = 0\Big\},$$

is the same curve in every horizontal plane $z = \text{const}$. For example, if $A = B = 1$, $p = 0$, and $D = -R^2$, the cross section of the surface $\mathcal{S}$ by any horizontal plane is a circle $x^2 + y^2 = R^2$. So the surface $\mathcal{S}$ is a cylinder of radius R that is swept by the circle when the latter is shifted up and down parallel to the z axis. Similarly, a general cylindrical shape is obtained by shifting a curve in the xy plane up and down parallel to the z axis. The task here is to classify all possible shapes of quadric cylinders.

In Example 11.2, it was established that, under a counterclockwise rotation of the coordinate system through an angle ϕ, $x \to x\cos\phi + y\sin\phi$ and $y \to y\cos\phi - x\sin\phi$. By substituting the transformed coordinates into the equation for $\mathcal{S}$, one obtains the equation for the *very same* shape in the new rotated coordinate system. The freedom of choosing the rotation angle ϕ can be used to simplify the equation. In particular, it is always possible to adjust ϕ so that in the new coordinate system the equation for $\mathcal{S}$ does not contain the "mixed" term xy. Indeed, after the substitution of the transformed coordinates into the equation, the coefficient at xy defines a new p:

$$p \to 2(A-B)\cos\phi\sin\phi + a(\cos^2\phi - \sin^2\phi) = (A-B)\sin(2\phi) \\ + a\cos(2\phi) = p'.$$

Therefore, the term xy disappears from the equation if the angle ϕ satisfies the condition $p' = 0$ or

$$\tan(2\phi) = \frac{q}{B-A}, \quad \text{and} \quad \phi = \frac{\pi}{4} \quad \text{if} \quad A = B. \tag{11.15}$$

The coefficients A and B (the factors at x^2 and y^2) and α and β (the factors at x and y) are transformed as

$$\begin{aligned} A &\to \tfrac{1}{2}[A+B+(A-B)\cos(2\phi) - a\sin(2\phi)] = A', \\ B &\to \tfrac{1}{2}[A+B-(A-B)\cos(2\phi) + a\sin(2\phi)] = B', \\ \alpha &\to \alpha\cos\phi - \beta\sin\phi = \alpha', \\ \beta &\to \beta\cos\phi + \alpha\sin\phi = \beta', \end{aligned}$$

where ϕ satisfies (11.15).

Depending on the values of A, B, and p, the following three cases can occur. First, $A' = B' = 0$, which is only possible if $A = B = p = 0$. In this case, $\mathcal{S}$ is defined by the equation $\alpha'x + \beta'y + D = 0$, which is a plane parallel to the z axis.

Second, only one of A' and B' vanishes, say, $B' = 0$ (note that for establishing the shape it is irrelevant how the horizontal and vertical coordinates in the xy plane are called). In this case, the equation for $\mathcal{S}$ assumes the form $A'x^2 + \alpha'x + \beta'y + D = 0$ or, by completing the squares,

$$\frac{A'}{\beta'}\Big(x - x_0\Big)^2 + (y - y_0) = 0\,, \quad x_0 = \frac{\alpha'}{2A'}\,,\ y_0 = \frac{1}{\beta'}\Big(Ax_0^2 - D\Big)\,;$$

here it is assumed that $\beta' \neq 0$ (otherwise, the equation does not define a curve in the xy plane and hence $\mathcal{S}$ is not a surface). This equation defines a parabola $y = A''x^2$, $A'' = A'/\beta'$, after the *translation* of the coordinate system: $x \to x + x_0$ and $y \to y + y_0$ (see Section 71).

Third, both A' and B' do not vanish. Then, after the completion of squares, the equation can be brought to the form

$$A'(x - x_0)^2 + B'(y - y_0)^2 = D',$$

where

$$x_0 = -\frac{\alpha'}{2A'}\,, \quad y_0 = -\frac{\beta'}{2B'}\,, \quad D' = -D + \frac{1}{2}(A'x_0^2 + B'y_0^2).$$

Finally, after the translation of the origin to the point (x_0, y_0), the equation becomes

$$A'x^2 + B'y^2 = D'.$$

If $D' = 0$, then this equation defines two straight lines $y = \pm mx$, where $m = (-A'/B')^{-1/2}$, provided A' and B' have opposite signs (otherwise, the equation has no solution). If $D' \neq 0$, then the equation can be written as $(A'/D')x^2 + (B'/D')y^2 = 1$. One can always assume that $A'/D' > 0$ (as the shape of the curve is independent of how the coordinate axes are called). Note also that the rotation of the coordinate system through the angle $\pi/2$ swaps the axes, $(x, y) \to (y, -x)$, which can be used to reverse the sign of A'/D'. Now put $A'/D' = 1/a^2$ and $B'/D' = \pm 1/b^2$ (depending on whether B'/D' is positive or negative) so that the equation becomes

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

When the plus is taken, this equation defines an ellipse. When the minus is taken, this equation defines a hyperbola.

The above results are summarized in the following theorem (see Figure 11.20).

Theorem 11.10. (Classification of Quadric Cylinders). *A general equation for quadric cylinders*

$$\mathcal{S} = \Big\{(x, y, z)\Big| Ax^2 + By^2 + pxy + \alpha x + \beta y + D = 0\Big\}$$

can be brought to one of the following standard forms by a suitable rotation and translation of the coordinate system:

$$y - ax^2 = 0 \qquad \text{(parabolic cylinder)},$$

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \qquad \text{(elliptic cylinder)},$$

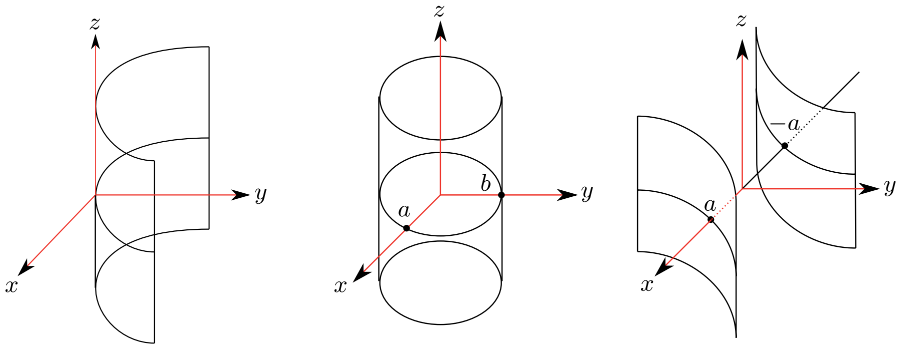

Figure 11.20. **Left**: Parabolic cylinder. The cross section by any horizontal plane $z = \text{const}$ is a parabola $y = ax^2$.
Middle: An elliptic cylinder. The cross section by any horizontal plane $z = \text{const}$ is an ellipse $x^2/a^2+y^2/b^2 = 1$.
Right: A hyperbolic cylinder. The cross section by any horizontal plane $z = \text{const}$ is a hyperbola $x^2/a^2-y^2/b^2 = 1$.

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \qquad \text{(hyperbolic cylinder)},$$

provided A, B, and p do not vanish simultaneously. If $A = B = p = 0$, then $\mathcal{S}$ is a plane.

78.2. Classification of General Quadric Surfaces. The classification of general quadric surfaces can be carried out in the same way (i.e., by simplifying the general quadratic equation by means of rotations and translations of the coordinate system). First, one can prove that there exists a rotation of the coordinate system such that in the new coordinate system the quadratic equation does not have "mixed" terms: $p \to p' = 0$, $q \to q' = 0$, and $v \to v' = 0$. Depending on how many of the coefficients A', B', and C' do not vanish, some of the linear terms or all of them can be eliminated by translations of the coordinate system. The corresponding technicalities require a substantial use of linear algebra methods, which goes beyond the scope of this course. So the final result is given without a proof.

Theorem 11.11. (Classification of Quadric Surfaces).
By rotating and translating a rectangular coordinate system, a general equation for quadric surfaces can be brought either to one of the

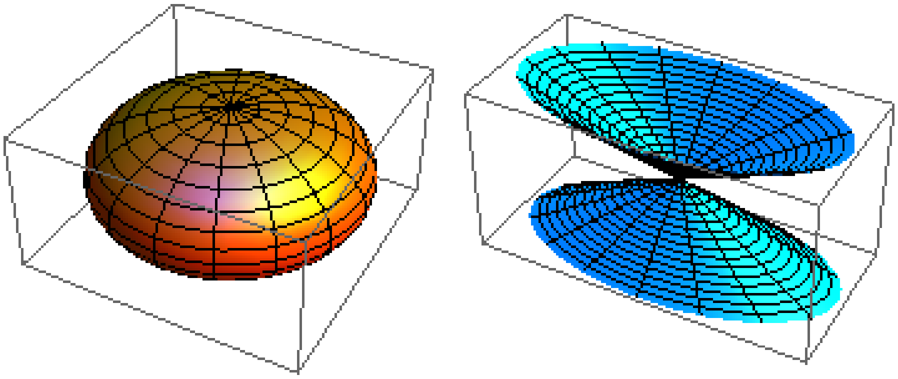

FIGURE 11.21. **Left**: An ellipsoid. A cross section by any coordinate plane is an ellipse.
Right: An elliptic double cone. A cross section by a horizontal plane $z =$ const is an ellipse. A cross section by any vertical plane through the z axis is two lines through the origin.

standard equations for quadric cylinders or to one of the following six forms:

$$\begin{aligned}
&\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1 &&\text{(ellipsoid)},\\
&\frac{z^2}{c^2}=\frac{x^2}{a^2}+\frac{y^2}{b^2} &&\text{(elliptic double cone)},\\
&\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=1 &&\text{(hyperboloid of one sheet)},\\
&-\frac{x^2}{a^2}-\frac{y^2}{b^2}+\frac{z^2}{c^2}=1 &&\text{(hyperboloid of two sheets)},\\
&\frac{z}{c}=\frac{x^2}{a^2}+\frac{y^2}{b^2} &&\text{(elliptic paraboloid)},\\
&\frac{z}{c}=\frac{x^2}{a^2}-\frac{y^2}{b^2} &&\text{(hyperbolic paraboloid)}.
\end{aligned}$$

It should be noted again that the shape of the surface does not depend on how the coordinate axes are called. So the shape does not change under any permutation of the coordinates (x, y, z) in the standard equations; only the orientation of the shape relative to the

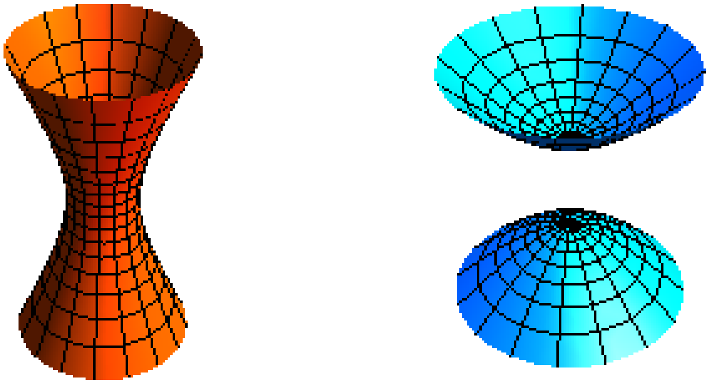

Figure 11.22. **Left**: A hyperboloid of one sheet. A cross section by a horizontal plane z =const is an ellipse. A cross section by a vertical plane x =const or y =const is a hyperbola.
Right: A hyperboloid of two sheets. A nonempty cross section by a horizontal plane is an ellipse. A cross section by a vertical plane is a hyperbola.

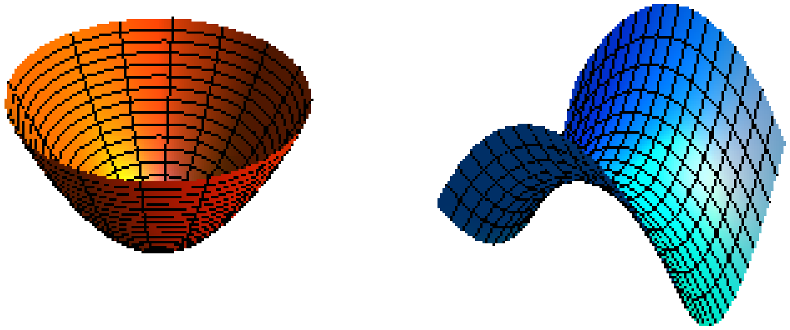

Figure 11.23. **Left**: An elliptic paraboloid. A nonempty cross section by a horizontal plane is an ellipse. A cross section by a vertical plane is a parabola.
Right: A hyperbolic paraboloid (a "saddle"). A cross section by a horizontal plane is a hyperbola. A cross section by a vertical plane is a parabola.

coordinate system changes. For example, the equations $x^2 + y^2 = R^2$ and $y^2 + z^2 = R^2$ describe a cylinder of radius R, but in the former case the cylinder axis coincides with the z axis, while the cylinder axis is the x axis in the latter case.

78.3. Visualization of Quadric Surfaces. The shape of a quadric surface can be understood by studying intersections of the surface with the coordinate planes $x = x_0$, $y = y_0$, and $z = z_0$. These intersections are also called *traces.*

An Ellipsoid. If $a^2 = b^2 = c^2 = R^2$, then the ellipsoid becomes a sphere of radius R. So, intuitively, an ellipsoid is a sphere "stretched" along the coordinate axes. Traces of an ellipsoid in the planes $x = x_0$, $|x_0| < a$, are ellipses $y^2/b^2 + z^2/c^2 = k$, where $k = 1 - x_0^2/a^2$. Apparently, the traces in the planes $x = \pm a$ consist of a single point $(\pm a, 0, 0)$, and there is no trace in any plane $x = x_0$ if $|x_0| > a$. Traces in the planes $y = y_0$ and $z = z_0$ are also ellipses and exist only if $|y_0| \leq b$ and $|z_0| \leq c$. Thus, *the characteristic geometrical property of an ellipsoid is that its traces are ellipses.*

A Paraboloid. Suppose $c > 0$. Then the paraboloid lies above the xy plane because it has no trace in all horizontal planes below the xy plane, $z = z_0 < 0$. In the xy plane, its trace contains just the origin. Similarly, a paraboloid with $c > 0$ lies below the xy plane. Horizontal traces (in the planes $z = z_0$) of the paraboloid are ellipses, $x^2/a^2 + y^2/b^2 = k$, where $k = z_0/c$. The ellipses become wider as z_0 gets larger ($c > 0$). For the sake of simplicity, put $c = 1$. Vertical traces (traces in the planes $x = x_0$ or $y = y_0$) are parabolas $z - k = y^2/b^2$, where $k = x_0^2/a^2$, or $z - k = x^2/a^2$, where $k = y_0^2/b^2$. So *the characteristic geometrical property of a paraboloid is that its horizontal traces are ellipses, while its vertical ones are parabolas.* If $a = b$, the ellipsoid is also called a *circular ellipsoid* because its horizontal traces are circles.

A Double Cone. Horizontal traces are ellipses $x^2/a^2 + y^2/b^2 = k$, where $k = z_0^2/c^2$. They become wider as $|z_0|$ grows, that is, as the horizontal plane moves away from the xy plane ($z = 0$). In the xy plane, the cone has a trace that consists of a single point (the origin). The vertical traces in the planes $x = 0$ or $y = 0$ are a pair of lines $z = \pm(c/b)y$ or $z = \pm(c/a)x$. Vertical traces in the planes $x = x_0 \neq 0$ or $y = y_0 \neq 0$ are hyperbolas $y^2/b^2 - z^2/c^2 = k$, where $k = -x_0^2/a^2$, or $x^2/a^2 - z^2/c^2 = k$, where $k = -y_0^2/b^2$. So *the characteristic geometrical property of a cone is that horizontal traces are ellipses; its vertical traces are either a pair of lines, if the plane contains the cone axis,* *or hyperbolas.* If $a = b$, the cone is called a *circular cone.* In this case, vertical traces in the planes containing the cone axis are a pair of lines with the same slope $c/a = c/b$. The angle ϕ between the axis of the cone and any of these lines defines the cone uniquely because $c/a = \cot\phi$, and the equation of the cone can be written as

$$z^2 = \cot^2(\phi)(x^2 + y^2)\,, \quad 0 < \phi < \pi/2\,.$$

The equation for an upper or lower cone of the double circular cone is

$$z = \pm\cot(\phi)\sqrt{x^2 + y^2}\,.$$

A Hyperbolic Paraboloid. Horizontal traces are hyperbolas $x^2/a^2 - y^2/b^2 = k$, where $k = z_0/c$. For simplicity, put $a = b$ and $c = -1$. Then, if $z_0 < 0$ (horizontal planes below the xy plane), then the hyperbolas are symmetric about the x axis, and their branches lie either in $x > 0$ or in $x < 0$ (i.e., they do not intersect the y axis). If $z_0 > 0$, then the hyperbolas are symmetric about the y axis, and their branches lie either in $y > 0$ or in $y < 0$ (i.e., they do not intersect the x axis). Vertical traces in the planes $x = x_0$ are upward parabolas, whereas in the planes $y = y_0$ they are downward parabolas. The hyperbolic paraboloid has the characteristic shape of a "saddle."

A Hyperboloid of One Sheet. *The characteristic geometrical property of a hyperboloid of one sheet is that its horizontal traces are ellipses and its vertical traces are hyperbolas.* Every horizontal plane has a trace of the hyperboloid, and the smallest one is in the xy plane (an ellipse with semi axes a and b). The semiaxes of the ellipses increase as the plane moves away from the xy plane.

A Hyperboloid of Two Sheets. A distinctive feature of this surface is that it consists of two sheets. Indeed, the hyperboloid has no trace in the horizontal planes $z = z_0$ if $-c < z_0 < c$. So one sheet lies above the plane $z = c$ and the other lies below the plane $z = -c$. Horizontal traces in the planes $z = z_0 > c$ or $z = z_0 < -c$ are ellipses. The upper sheet touches the plane $z = c$ at the point $(0, 0, c)$, while the lower sheet touches the plane $z = -c$ at the point $(0, 0, -c)$. Vertical traces are hyperbolas. So the characteristic geometrical properties of hyperboloids of one sheet and two sheets are similar, apart from the fact that the latter one consists of two sheets. Also, in the asymptotic region $|z| \gg c$, the hyperboloids approach the surface of the double cone. Indeed, in this case, $z^2/c^2 \gg 1$, and hence the equations $x^2/a^2+y^2/b^2 = \pm1 + z^2/c^2$ can be well approximated by the double-cone equation (±1 can be neglected on the right side of the equations). In the region $z > 0$, the hyperboloid of one sheet approaches the double cone from below, while the hyperboloid of two sheets approaches it from above.

For $z < 0$, the converse holds. In other words, the hyperboloid of two sheets lies "inside" the cone, while the hyperboloid of one sheet lies "outside" it.

78.4. Study Problems.

Problem 11.27. *Classify the quadric surface* $3x^2 + 3y^3 - 2xy = 4$.

Solution: The equation does not contain one variable (the z coordinate). The surface is a cylinder. To determine the type of cylinder, consider a rotation of the coordinate system in the xy plane and choose the rotation angle so that the coefficient at the "mixed" term vanishes. According to (11.15), $A = B = 3$ and hence $\phi = \pi/4$. Then $A' = (A + B - a)/2 = 4$ and $B' = (A + B + a)/2 = 2$. So, in the new coordinates, the equation becomes $x^2 + y^2/2 = 1$, which is an ellipse with semiaxes $a = 1$ and $b = \sqrt{2}$. The surface is an elliptic cylinder. □

Problem 11.28. *Classify the quadric surface* $x^2 - 2x + y + z = 0$.

Solution: By completing the squares, the equation can be transformed into the form $(x-1)^2+(y-1)+z = 0$. After shifting the origin to the point $(1, 1, 0)$, the equation becomes $x^2+y-z = 0$. Consider rotations of the coordinate system about the x axis: $y \to \cos\phi y + \sin\phi z$, $z \to \cos\phi z - \sin\phi y$. Under this rotation, $y - z \to (\cos\phi + \sin\phi)y + (\sin\phi - \cos\phi)z$. Therefore, for $\phi = \pi/4$, the equation assumes one of the standard forms $x^2 + \sqrt{2}\, y = 0$, which corresponds to a parabolic cylinder. □

Problem 11.29. *Classify the quadric surface* $x^2+z^2-2x+2z-y = 0$.

Solution: By completing the squares, the equation can be transformed into the form $(x - 1)^2 + (z + 1)^2 - (y + 2) = 0$. The latter can be brought into one of the standard forms by shifting the origin to the point $(1, -2, -1)$: $x^2 + z^2 = y$, which is a circular paraboloid. Its symmetry axis is parallel to the y axis (the line of intersection of the planes $x = 1$ and $z = -1$) and its vertex is $(1, -2, -1)$. □

Problem 11.30. *Sketch and/or describe the set of points in space formed by a family of lines through the point* $(1, 2, 0)$ *and parallel to* $\mathbf{v}_\theta = \langle\cos\theta, \sin\theta, 1\rangle$, *where* $\theta \in [0, 2\pi]$ *labels lines in the family.*

Solution: The parametric equations of each line are $x = 1 + t\cos\theta$, $y = 2 + t\sin\theta$, and $z = t$. Therefore, $(x - 1)^2 + (y - 2)^2 = z^2$ for all values of t and θ. Thus, the lines form a double cone whose axis is parallel to the z axis and whose vertex is $(1, 2, 0)$. Alternatively, one could notice that the vector $\mathbf{v}_\theta$ rotates about the z axis as θ changes.

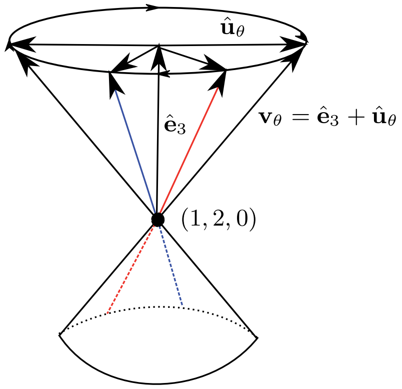
$\hat{\mathbf{u}}_\theta$
$\hat{\mathbf{e}}_3$
$\mathbf{v}_\theta = \hat{\mathbf{e}}_3 + \hat{\mathbf{u}}_\theta$
$(1, 2, 0)$

FIGURE 11.24. An illustration to Study Problem 11.30. The vector $\hat{\mathbf{u}}_\theta$ rotates about the vertical line so that the line through $(1, 2, 0)$ and parallel to $\mathbf{v}_\theta$ sweeps a double cone with the vertex at $(1, 2, 0)$.

Indeed, put $\mathbf{v}_\theta = \hat{\mathbf{u}} + \hat{\mathbf{e}}_z$, where $\hat{\mathbf{u}} = \langle \cos\theta, \sin\theta, 0\rangle$ is the unit vector in the xy plane as shown in Figure 11.24. It rotates as θ changes, making a full turn as θ increases from 0 to 2π. So the set in question can be obtained by rotating a particular line, say, the one corresponding to $\theta = 0$, about the vertical line through $(1, 2, 0)$. The line sweeps the double cone. □

78.5. Exercises. **(1)** Use traces to sketch and identify each of the following surfaces:
(i) $y^2 = x^2 + 9z^2$
(ii) $y = x^2 - z^2$
(iii) $4x^2 + 2y^2 + z^2 = 4$
(iv) $x^2 - y^2 + z^2 = -1$
(v). $y^2 + 4z^2 = 16$
(vi). $x^2 - y^2 + z^2 = 1$

(2) Reduce each of the following equations to one of the standard form, classify the surface, and sketch it:
(i) $x^2 + y^2 + 4z^2 - 2x + 4y = 0$
(ii) $x^2 - y^2 + z^2 + 2x - 2y + 4z + 2 = 0$
(iii) $x^2 + 4y^2 - 6x + z = 0$

(3) Find an equation for the surface obtained by rotating the line $y = 2x$ about the y axis.

(4) Find an equation for the surface consisting of all points that are equidistant from the point $(1, 1, 1)$ and the plane $z = 2$.

(5) Sketch the solid region bounded by the surface $z = \sqrt{x^2 + y^2}$ from below and by $x^2 + y^2 + z^2 - 2z = 0$ from above.

(6) Find an equation for the surface consisting of all points P for which the distance from P to the y axis is twice the distance from P to the zx plane. Identify the surface.

(7) Show that if the point (a, b, c) lies on the hyperbolic paraboloid $z = y^2 - x^2$, then the lines through (a, b, c) and parallel to $\mathbf{v} = \langle 1, 1, 2(b - a)\rangle$ and $\mathbf{u} = \langle 1, -1, -2(b - a)\rangle$ both lie entirely on this paraboloid. Deduce from this result that the hyperbolic paraboloid can be generated by the motion of a straight line. Show that hyperboloids of one sheet, cones, and cylinders can also be obtained by the motion of a straight line.

Remark. The fact that hyperboloids of one sheet are generated by the motion of a straight line is used to produce gear transmissions. The cogs of the gears are the generating lines of the hyperboloids.

(8) Find an equation for the cylinder of radius R whose axis goes through the origin and is parallel to a vector $\mathbf{v}$.

(9) Show that the curve of intersection of the surfaces $x^2 - 2y^2 + 3z^2 - 2x + y - z = 1$ and $2x^2 - 4y^2 + 6z^2 + x - y + 2z = 4$ lies in a plane.

CHAPTER 12

Vector Functions

79. Curves in Space and Vector Functions

To describe the motion of a pointlike object in space, its position vectors must be specified at every moment of time. A vector is defined by three components in a coordinate system. Therefore, the motion of the object can be described by an ordered triple of real-valued functions of time. This observation leads to the concept of vector-valued functions of a real variable.

Definition 12.1. (Vector Function).
Let $\mathcal{D}$ be a set of real numbers. A vector function $\mathbf{r}(t)$ of a real variable t is a rule that assigns a vector to every value of t from $\mathcal{D}$. The set $\mathcal{D}$ is called the domain *of the vector function.*

Most commonly used rules to define a vector function are algebraic rules that specify components of a vector function in a coordinate system as functions of a real variable: $\mathbf{r}(t) = \langle x(t), y(t), z(t)\rangle$. For example,

$$\mathbf{r}(t) = \langle \sqrt{1-t}\,,\ \ln(t)\,,\ t^2\rangle \quad \text{or} \quad x(t) = \sqrt{1-t}\,,\ y(t) = \ln(t)\,,\ z(t) = t^2\,.$$

Unless specified otherwise, the domain of the vector function is the set $\mathcal{D}$ of all values of t at which the algebraic rule makes sense; that is, all three components can be computed for any t from $\mathcal{D}$. In the above example, the domain of $x(t)$ is $-\infty < t \leq 1$, the domain of $y(t)$ is $0 < t < \infty$, and the domain of $z(t)$ is $-\infty < t < \infty$. The domain of the vector function is the intersection of the domains of its components: $\mathcal{D} = (0, 1]$.

Suppose that the components of a vector function $\mathbf{r}(t)$ are continuous functions on $\mathcal{D} = [a, b]$. Consider all vectors $\mathbf{r}(t)$, as t ranges over the domain $\mathcal{D}$, positioned so that their initial points are at a fixed point (e.g., the origin of a coordinate system). Then the terminal points of the vectors $\mathbf{r}(t)$ form a *curve* in space as depicted in Figure 12.1 (left panel). The simplest example is provided by the motion along a straight line, which is described by a linear vector function $\mathbf{r}(t) = \mathbf{r}_0 + t\mathbf{v}$. Thus, the *range* of a vector function defines a curve in

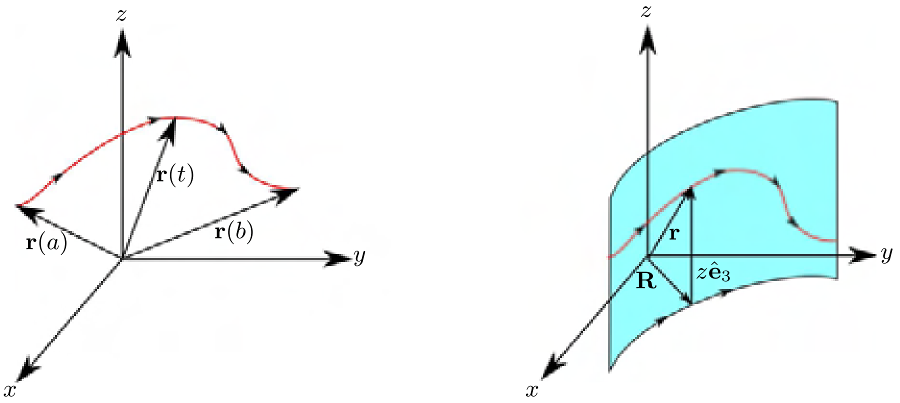

FIGURE 12.1. **Left**: The terminal point of a vector $\mathbf{r}(t)$ whose components are continuous functions of t traces out a curve in space.
Right: Graphing a space curve. Draw a curve in the xy plane defined by the parametric equations $x = x(t)$, $y = y(t)$. It is traced out by the vector $\mathbf{R}(t) = \langle x(t), y(t), 0\rangle$. This planar curve defines a cylindrical surface in space in which the space curve in question lies. The space curve is obtained by raising or lowering the points of the planar curve along the surface by the amount $z(t)$, that is, $\mathbf{r}(t) = \mathbf{R}(t) + \hat{\mathbf{e}}_3 z(t)$. In other words, the graph $z = z(t)$ is wrapped around the cylindrical surface.

space, and a graph of a vector function is a curve in space. There is a difference between graphs of ordinary functions and graphs of vector functions, though. The function of a real variable is uniquely defined by its graph. This is not so for vector functions. Suppose the shape of a curve is described geometrically, that is, as a point set in space (e.g., a line through two given points). One might ask the question: What is a vector function that traces out a given curve in space? The answer to this question is not unique. For example, a line $\mathcal{L}$ as a point set in space is uniquely defined by its particular point and a vector $\mathbf{v}$ parallel to it. If $\mathbf{r}_1$ and $\mathbf{r}_2$ are position vectors of two particular points of $\mathcal{L}$, then both vector functions $\mathbf{r}_1(t) = \mathbf{r}_1 + t\mathbf{v}$ and $\mathbf{r}_2(t) = \mathbf{r}_2 - 2t\mathbf{v}$ trace out $\mathcal{L}$ because the vector $-2\mathbf{v}$ is also parallel to $\mathcal{L}$.

The following, more sophisticated example is also of interest. Suppose one wants to find a vector function that traces out a semicircle of radius R. Let the semicircle be positioned in the upper part of the

xy plane ($y \geq 0$). The following three vector functions trace out the semicircle:

$$\begin{aligned}
\mathbf{r}_1(t) &= \langle t,\ \sqrt{R^2 - t^2},\ 0\rangle\,, \quad -R \leq t \leq R\,,\\
\mathbf{r}_2(t) &= \langle R\cos t,\ R\sin t,\ 0\rangle\,, \quad 0 \leq t \leq \pi\,,\\
\mathbf{r}_3(t) &= \langle -R\cos t,\ R\sin t,\ 0\rangle\,, \quad 0 \leq t \leq \pi\,.
\end{aligned}$$

This is easy to see by computing the norm of these vector functions: $\|\mathbf{r}_i(t)\|^2 = R^2$ or $x_i^2(t) + y_i^2(t) = R^2$, where $i = 1, 2, 3$, for any value of t; that is, the endpoints of the vectors $\mathbf{r}_i(t)$ always remain on the semi circle of radius R as t ranges over the specified interval. It can therefore be concluded that there are many vector functions that trace out the same curve in space defined as a point set in space.

Another observation is that there are vector functions that trace out the same curve in opposite directions. In the above example, the vector function $\mathbf{r}_2(t)$ traces out the semicircle counterclockwise, while the functions $\mathbf{r}_1(t)$ and $\mathbf{r}_3(t)$ do so clockwise. So a vector function defines the *orientation* of a spatial curve. However, this notion of the orientation of a curve should be regarded with caution. For example, the vector function $\mathbf{r}(t) = \langle R\cos t, R|\sin t|, 0\rangle$ traces out the semicircle twice, back and forth, when t ranges from 0 to 2π. In this case, the range of $\mathbf{r}(t)$ should be considered as two semicircles (one is oriented counterclockwise and the other clockwise), and these semicircles are superimposed one onto the other.

79.1. Graphing Space Curves. To visualize the shape of a curve C traced out by a vector function, it is convenient to think about $\mathbf{r}(t)$ as a trajectory of motion. The position of a particle in space may be determined by its position in a plane and its height relative to that plane. For example, this plane can be chosen to be the xy plane. Then

$$\mathbf{r}(t) = \langle x(t),\ y(t),\ z(t)\rangle = \langle x(t),\ y(t),\ 0\rangle + \langle 0,\ 0,\ z(t)\rangle = \mathbf{R}(t) + z(t)\hat{\mathbf{e}}_3.$$

Consider the curve defined by the parametric equations $x = x(t)$, $y = y(t)$ in the xy plane. One can mark a few points along the curve corresponding to particular values of t, say, P_n with coordinates $(x(t_n), y(t_n))$, $n = 1, 2, ..., N$. Then the corresponding points of the curve C are obtained from them by moving the points P_n along the direction normal to the plane (i.e., along the z axis in this case), by the amount $z(t_n)$; that is, P_n goes up if $z(t_n) > 0$ or down if $z(t_n) < 0$. In other words, as a particle moves along the curve $x = x(t)$, $y = y(t)$, it ascends or descends according to the corresponding value of $z(t)$.

The curve can also be visualized by using a piece of paper. Consider a general cylinder with the horizontal trace being the curve $x = x(t)$, $y = y(t)$, like a wall of the shape defined by this curve. Then make a graph of the function $z(t)$ on a piece of paper (wallpaper) and glue it to the wall so that the t axis of the graph is glued to the curve $x = x(t)$, $y = y(t)$ (i.e., each point t on the t axis is glued to the corresponding point $(x(t), y(t))$ of the curve). After such a procedure, the graph of $z(t)$ along the wall would coincide with the curve C traced out by $\mathbf{r}(t)$. The procedure is illustrated in Figure 12.1 (right panel).

Example 12.1. *Graph the vector function* $\mathbf{r} = \langle \cos t, \sin t, t\rangle$*, where* t *ranges over the real line.*

Solution: It is convenient to represent $\mathbf{r}(t)$ as the sum of a vector in the xy plane and a vector parallel to the z axis. In the xy plane, the curve $x = \cos t$, $y = \sin t$ is the circle of unit radius traced out counterclockwise so that the point $(1, 0, 0)$ corresponds to $t = 0$. The circular motion is periodic with period 2π. The height $z(t) = t$ rises linearly as the point moves along the circle. Starting from $(1, 0, 0)$, the curve makes one turn on the surface of the cylinder of unit radius climbing up by 2π in each turn. Think of a piece of paper with a straight line depicted on it that is wrapped around the cylinder. Thus, the curve traced by $\mathbf{r}(t)$ lies on the surface of a cylinder of unit radius and periodically winds about it climbing by 2π per turn. Such a curve is called a *helix*. The procedure is shown in Figure 12.2. □

79.2. Limits and Continuity of Vector Functions.

Definition 12.2. (Limit of a Vector Function).
A vector $\mathbf{r}_0$ *is called the* limit *of a vector function* $\mathbf{r}(t)$ *as* $t \to t_0$ *if*

$$\lim_{t\to t_0} \|\mathbf{r}(t) - \mathbf{r}_0\| = 0\,;$$

the limit is denoted as $\lim_{t\to t_0} \mathbf{r}(t) = \mathbf{r}_0$.

The left and right limits, $\lim_{t\to t_0^-} \mathbf{r}(t)$ and $\lim_{t\to t_0^+} \mathbf{r}(t)$, are defined similarly. This definition says that the length or norm of the vector $\mathbf{r}(t) - \mathbf{r}_0$ approaches 0 as t tends to t_0. The norm of a vector vanishes if and only if the vector is the zero vector. Therefore, the following theorem holds.

Theorem 12.1. (Limit of a Vector Function).
Let $\mathbf{r}(t) = \langle x(t), y(t), z(t)\rangle$ *and let* $\mathbf{r}_0 = \langle x_0, y_0, z_0\rangle$*. Then*

$$\lim_{t\to t_0} \mathbf{r}(t) = \mathbf{r}_0 \quad \Longleftrightarrow \quad \lim_{t\to t_0} x(t) = x_0\,,\ \ \lim_{t\to t_0} y(t) = y_0\,,\ \ \lim_{t\to t_0} z(t) = z_0.$$

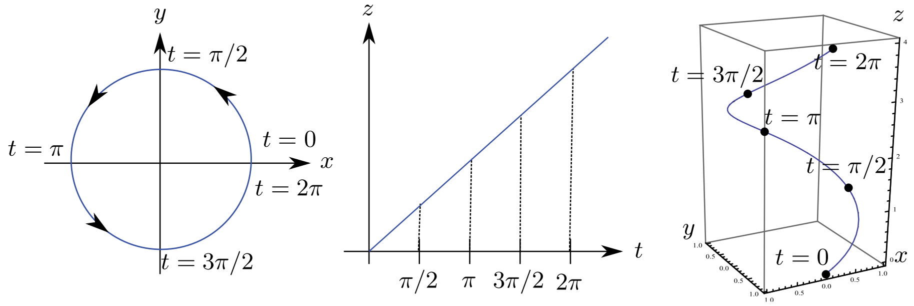

FIGURE 12.2. Graphing a helix. **Right**: The curve $\mathbf{R}(t) = \langle \cos t, \sin t, 0\rangle$ is a circle of unit radius, traced out counterclockwise. So the helix lies on the cylinder of unit radius whose symmetry axis is the z axis.
Middle: The graph $z = z(t) = t$ is a straight line that defines the height of helix points relative to the circle traced out by $\mathbf{R}(t)$.
Right: The graph of the helix $\mathbf{r}(t) = \mathbf{R}(t) + z(t)\hat{\mathbf{e}}_3$. As $\mathbf{R}(t)$ traverses the circle, the height $z(t) = t$ rises linearly. So the helix can be viewed as a straight line wrapped around the cylinder.

This theorem reduces the problem of finding the limit of a vector function to the problem of finding limits of three ordinary functions. It also says that the limit of a vector function exists if and only if the limits of its all components exist.

EXAMPLE 12.2. *Find the limit of* $\mathbf{r}(t) = \langle \sin(t)/t\,,\ t\ln t\,, (e^t - 1 - t)/t^2\rangle$ *as* $t \to 0^+$.

SOLUTION: By l'Hospital's rule,

$$\lim_{t\to 0^+} \frac{\sin t}{t} = \lim_{t\to 0^+} \frac{\cos t}{1} = 1,$$

$$\lim_{t\to 0^+} t\ln t = \lim_{t\to 0^+} \frac{\ln t}{t^{-1}} = \lim_{t\to 0^+} \frac{t^{-1}}{-t^{-2}} = -\lim_{t\to 0^+} t = 0,$$

$$\lim_{t\to 0^+} \frac{e^t - 1 - t}{t^2} = \lim_{t\to 0^+} \frac{e^t - 1}{2t} = \lim_{t\to 0^+} \frac{e^t}{2} = \frac{1}{2}.$$

Therefore, $\lim_{t\to 0^+} \mathbf{r}(t) = \langle 1, 0, 1/2\rangle$. □

Definition 12.3. (Continuity of a Vector Function).
A vector function $\mathbf{r}(t)$, $t \in [a,b]$, *is said to be continuous at* $t = t_0 \in [a,b]$ *if*

$$\lim_{t\to t_0} \mathbf{r}(t) = \mathbf{r}(t_0)\,.$$

A vector function $\mathbf{r}(t)$ *is continuous in the interval* $[a,b]$ *if it is continuous at every point of* $[a,b]$.

By Theorem 12.1, *a vector function is continuous if and only if all its components are continuous functions.*

Example 12.3. *Determine whether the vector function* $\mathbf{r}(t) = \langle \sin(2t)/t\,, t^2, e^t\rangle$ *for all* $t \neq 0$ *and* $\mathbf{r}(0) = \langle 1, 0, 1\rangle$ *is continuous.*

Solution: The components $y(t) = t^2$ and $z(t) = e^t$ are continuous for all real t and $y(0) = 0$ and $z(0) = 1$. The component $x(t) = \sin(2t)/t$ is continuous for all $t \neq 0$ because the ratio of two continuous functions is continuous. By l'Hospital's rule,

$$\lim_{t\to 0} x(t) = \lim_{t\to 0} \frac{\sin(2t)}{t} = \lim_{t\to 0} \frac{2\cos(2t)}{1} = 2 \quad \Rightarrow \quad \lim_{t\to 0} x(t) \neq x(0) = 1;$$

that is, $x(t)$ is not continuous at $t = 0$. Thus, $\mathbf{r}(t)$ is continuous everywhere, but $t = 0$. □

79.3. Space Curve as a Continuous Vector Function. A curve in space can be understood as a continuous transformation (or a deformation without breaking) of a straight line segment in space. Conversely, every space curve can be continuously deformed to a straight line segment. So *a space curve is a continuous deformation of a straight line segment, and this deformation has a continuous inverse.* This motivates the following (simpler) definition of a spatial curve that is sufficient for all applications discussed in this course.

Definition 12.4. (Curve in Space).
A curve in space is the range of a continuous vector function.

If a curve in space is defined as a point set by geometrical means (e.g., as an intersection of two surfaces), then this definition implies that there is a continuous vector function whose range coincides with the point set. It should be understood that there are different continuous vector functions with the same range, and a continuous vector function may traverse the same point set several times. For example, the vector function $\mathbf{r}(t) = (t^2, t^2, t^2)$ is continuous on the interval $[-1, 1]$ and traces out the straight line segment, $x = y = z$, between the points $(0, 0, 0)$ and $(1, 1, 1)$ twice.

A curve is said to be *simple* if it does not intersect itself at any point; that is, a simple curve is a continuous vector function $\mathbf{r}(t)$ for which $\mathbf{r}(t_1) \neq \mathbf{r}(t_2)$ for any $t_1 \neq t_2$ in the open interval (a, b). A simple curve is always *oriented* because the function $\mathbf{r}(t)$ traces out its range only once from the initial point $\mathbf{r}(a)$ to the terminal point $\mathbf{r}(b)$. A curve is *closed* if $\mathbf{r}(a) = \mathbf{r}(b)$. The definition of a space curve as a continuous vector function is rather fruitful because it allows us to give a precise algebraic description of the geometrical properties a space curve may have.

79.4. Study Problems.

Problem 12.1. *Find a vector function that traces out a helix of radius R that climbs up along its axis by h.*

Solution: Let the helix axis be the z axis. The motion in the xy plane must be circular with radius R. Suitable parametric equations of the circle are $x(t) = R\cos t$, $y(t) = R\sin t$. With this parameterization of the circle, the motion has a period of 2π. On the other hand, $z(t)$ must rise linearly by h as t changes over the period. Therefore, $z(t) = ht/(2\pi)$. The vector function may be chosen in the form $\mathbf{r}(t) = \langle R\cos t, R\sin t, ht/(2\pi)\rangle$. □

Problem 12.2. *Sketch and/or describe the curve traced out by the vector function $\mathbf{r}(t) = \langle \cos t, \sin t, \sin(4t)\rangle$ if t ranges in the interval $[0, 2\pi]$.*

Solution: In the xy plane, the motion goes along the circle of unit radius, counterclockwise, starting from the point $(1, 0, 0)$. As t ranges over the specified interval, the circle is traced out only once. The height $z(t) = \sin(4t)$ has a period of $2\pi/4 = \pi/2$. Therefore, the graph of $\sin(4t)$ makes four ups and four downs if $0 \leq t \leq 2\pi$. The curve looks like the graph of $\sin(4t)$ wrapped around the cylinder of unit radius. It makes one up and one down in each quarter of the cylinder. The procedure is shown in Figure 12.3. □

Problem 12.3. *Sketch and/or describe the curve traced out by the vector function $\mathbf{r}(t) = \langle t\cos t, t\sin t, t\rangle$.*

Solution: The components of $\mathbf{r}(t)$ satisfy the equation $x^2(t)+y^2(t) = z^2(t)$ for all values of t. Therefore, the curve lies on the double cone $x^2 + y^2 = z^2$. Since $x^2(t) + y^2(t) = t^2$, the motion in the xy plane is a spiral (think of a rotational motion about the origin such that the radius increases linearly with the angle of rotation). If t increases from $t = 0$, the curve in question is traced by a point that rises linearly with

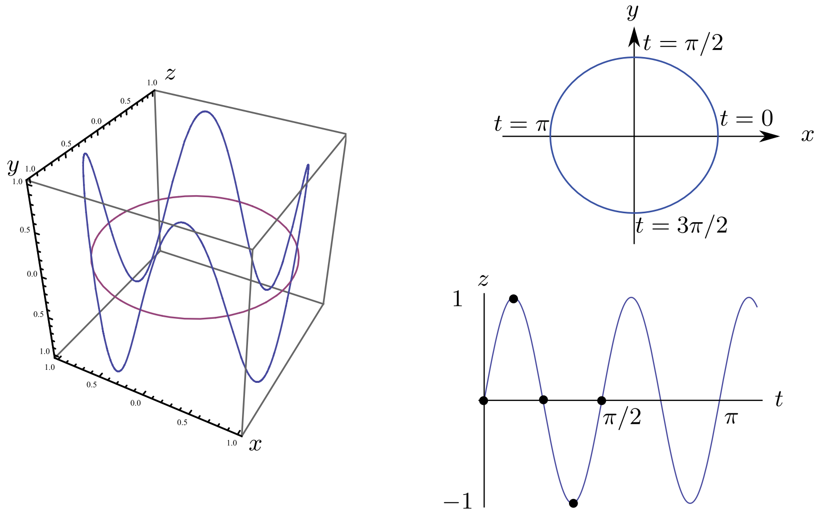

Figure 12.3. Illustration to Study Problem 12.2. **Left**: The curve lies on the cylinder of unit radius. It may be viewed as the graph of $z = \sin(4t)$ on the interval $0 \le t \le 2\pi$ wrapped around the cylinder.
Right: The circle traced out by $\mathbf{R}(t) = \langle \cos t, \sin t, 0\rangle$ (top). It defines the cylindrical surface on which the curve lies. The graph $z = z(t) = \sin(4t)$, which defines the height of points of the curve relative to the circle in the xy plane (bottom).

the distance from the origin as it travels along the spiral. If t decreases from $t = 0$, instead of rising, the point would descend ($z(t) = t < 0$). So the curve winds about the axis of the double cone while remaining on its surface. The procedure is shown in Figure 12.4. □

Problem 12.4. *Find the portion of the elliptic helix* $\mathbf{r}(t) = \langle 2\cos(\pi t), t, \sin(\pi t)\rangle$ *that lies inside the ellipsoid* $x^2 + y^2 + 4z^2 = 13$.

Solution: The helix here is called *elliptic* because it lies on the surface of an elliptic cylinder. Indeed, in the xz plane, the motion goes along the ellipse $x^2/4 + z^2 = 1$. So the curve remains on the surface of the elliptic cylinder parallel to the y axis. One turn around the ellipse occurs as t changes from 0 to 2. The helix rises by 2 along the y axis per turn. Now, to solve the problem, one has to find the values of t at which the helix intersects the ellipsoid. The intersection happens when the components of $\mathbf{r}(t)$ satisfy the equation of the ellipsoid, that is, when $x^2(t)+y^2(t)+4z^2(t) = 1$ or $4+t^2 = 13$ and hence $t = \pm 3$. The

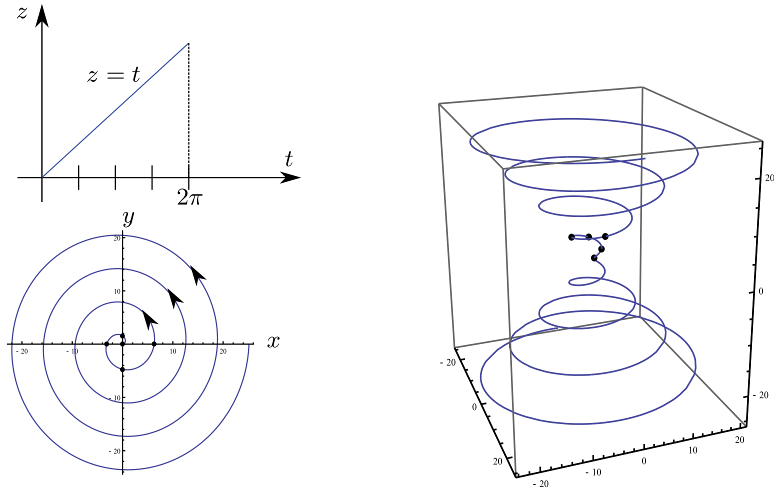

FIGURE 12.4. Illustration to Study Problem 12.3. **Left**: The height of the graph relative to the xy plane (top). The curve $\mathbf{R} = \langle t\cos t, t\sin t, 0\rangle$. For $t \geq 0$, it looks like an unwinding spiral (bottom).
Right: For $t > 0$, the curve is traversed by the point moving along the spiral while rising linearly with the distance traveled along the spiral. It can be viewed as a straight line wrapped around the cone $x^2 + y^2 = z^2$.

position vectors of the points of intersection are $\mathbf{r}(\pm 3) = \langle -2, \pm 3, 0\rangle$. The portion of the helix that lies inside the ellipsoid corresponds to the range $-3 \leq t \leq 3$. □

Problem 12.5. *Consider two curves C_1 and C_2 traced out by the vector functions $\mathbf{r}_1(t) = \langle t^2, t, t^2 + 2t - 8\rangle$ and $\mathbf{r}_2(s) = \langle 8 - 4s, 2s, s^2 + s - 2\rangle$, respectively. Do the curves intersect? If so, find the points of intersection. Suppose two particles have the trajectories $\mathbf{r}_1(t)$ and $\mathbf{r}_2(t)$, where t is time. Do the particles collide?*

SOLUTION: The curves intersect if there are values of the pair (t, s) such that $\mathbf{r}_1(t) = \mathbf{r}_2(s)$. This vector equation is equivalent the system of three equations $x_1(t) = x_2(s)$, $y_1(t) = y_2(s)$, $z_1(t) = z_2(s)$. It follows from the second equation that $t = 2s$. Substituting this into the first equation, one finds that $(2s)^2 = 8 - 4s$ whose solutions are $s = -2$ and $s = 1$. One has yet to verify that the third equation $t^2 + 2t - 8 = s^2 + s - 2$ holds for the pairs $(t, s) = (-4, -2)$ and $(t, s) = (2, 1)$ (otherwise, the z components do not match). A simple calculation shows that indeed both pairs satisfy the equation. So the position vectors of the points of intersection are $\mathbf{r}_1(-4) = \mathbf{r}_2(-2) = \langle 16, -4, 0\rangle$ and $\mathbf{r}_1(2) = \mathbf{r}_2(1) = \langle 4, 2, 0\rangle$. Although the curves along which the particles travel intersect, this does not mean that the particles would necessarily collide because they may not arrive at a point of intersection at the same moment of time, just like two cars traveling along intersecting streets may or may not collide at the street intersection. The collision condition is more restrictive, $\mathbf{r}_1(t) = \mathbf{r}_2(t)$ (i.e., the time t must satisfy three conditions). For the problem at hand, these conditions cannot be fulfilled for any t because, among all the solutions of $\mathbf{r}_1(t) = \mathbf{r}_2(s)$, there is no solution for which $t = s$. Thus, the particles do not collide. □

Problem 12.6. *Find a vector function that traces out the curve of intersection of the paraboloid $z = x^2 + y^2$ and the plane $2x + 2y + z = 2$ counterclockwise as viewed from the top of the z axis.*

Solution: One has to find the components $x(t)$, $y(t)$, and $z(t)$ such that they satisfy the equations of the paraboloid and plane simultaneously for all values of t. This ensures that the endpoint of the vector $\mathbf{r}(t)$ remains on both surfaces, that is, traces out their curve of intersection (see Figure 12.5). Consider first the motion in the xy plane. Solving the plane equation for z, $z = 2 - 2x - 2y$, and substituting the solution into the paraboloid equation, one finds $2 - 2x - 2y = x^2 + y^2$. After

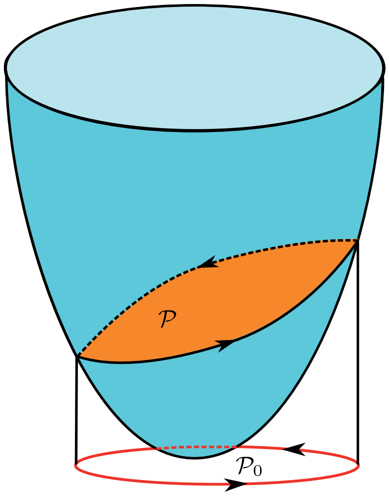

Figure 12.5. Illustration to Study Problem 12.6. The curve is an intersection of the paraboloid and the plane $\mathcal{P}$. It is traversed by the point moving counterclockwise about the circle in the xy plane (indicated by $\mathcal{P}_0$) and rising so that it remains on the paraboloid.

completing the squares, this equation becomes $4 = (x+1)^2 + (y+1)^2$, which describes a circle of radius 2 centered at $(-1,-1)$. The circle lies in the plane $\mathcal{P}_0$ in Figure 12.5. Its parametric equations read $x = x(t) = -1 + 2\cos t$, $y = y(t) = -1 + 2\sin t$. As t increases from 0 to 2π, the circle is traced out counterclockwise as required. Thus, $\mathbf{r}(t) = \langle -1 + 2\cos t, -1 + 2\sin t, 6 - 2\cos t - 2\sin t\rangle$, where $t \in [0, 2\pi]$. □

79.5. Exercises. **(1)** Find the domain of each of the following vector functions:
(i) $\mathbf{r}(t) = \langle \sqrt{9-t^2}, \ln t, \cos t\rangle$
(ii) $\mathbf{r}(t) = \langle \ln(9-t^2), \ln|t|, (1+t)/(2+t)\rangle$

(2) Find each of the following limits or show that it does not exist:
(i) $\lim_{t\to 0^+}\langle (e^{2t}-1)/t,\ (\sqrt{1+t}-1)/t,\ t\ln t\rangle$
(ii) $\lim_{t\to 0}\langle \sin^2(2t)/t^2,\ t^2+2,\ (\cos t - 1)/t^2\rangle$
(iii) $\lim_{t\to 0}\langle (e^{2t}-t)/t,\ t\cot t,\ \sqrt{1+t}\rangle$

(3) Sketch each of the following curves and identify the direction in which the curve is traced out as the parameter t increases:
(i) $\mathbf{r}(t) = \langle t, \cos(3t), \sin(3t)\rangle$
(ii) $\mathbf{r}(t) = \langle 2\sin(5t), 4, 3\cos(5t)\rangle$
(iii) $\mathbf{r}(t) = \langle 2t\sin t, 3t\cos t, t\rangle$
(iv) $\mathbf{r}(t) = \langle \sin t, \cos t, \ln t\rangle$

(4) Two objects are said to collide if they are at the same position *at the same time*. Two trajectories are said to intersect if they have common points. Let t be the physical time. Let two objects travel along the space curves $\mathbf{r}_1(t) = \langle t, t^2, t^3\rangle$ and $\mathbf{r}_2(t) = \langle 1+2t, 1+6t, 1+14t\rangle$. Do the objects collide? Do their trajectories intersect? If so, find the collision and intersection points.

(5) Find a vector function that traces out the curve of intersection of two surfaces:
(i) $x^2+y^2 = 9$ and $z = xy$
(ii) $x^2+y^2 = z^2$ and $x+y+z = 1$
(iii) $z = x^2+y^2$ and $y = x^2$

(6) Suppose that the limits $\lim_{t\to a}\mathbf{v}(t)$ and $\lim_{t\to a}\mathbf{u}(t)$ exist. Prove the basic laws of limits for the following vector functions:

$$\begin{aligned}
\lim_{t\to a}(\mathbf{v}(t)+\mathbf{u}(t)) &= \lim_{t\to a}\mathbf{v}(t) + \lim_{t\to a}\mathbf{u}(t),\\
\lim_{t\to a}(s\mathbf{v}(t)) &= s\lim_{t\to a}\mathbf{v}(t),\\
\lim_{t\to a}(\mathbf{v}(t)\cdot\mathbf{u}(t)) &= \lim_{t\to a}\mathbf{v}(t)\cdot\lim_{t\to a}\mathbf{u}(t),\\
\lim_{t\to a}(\mathbf{v}(t)\times\mathbf{u}(t)) &= \lim_{t\to a}\mathbf{v}(t)\times\lim_{t\to a}\mathbf{u}(t).
\end{aligned}$$

(7) To appreciate the basic laws of limits established in exercise 6, put $\mathbf{v}(t) = \langle (e^{2t}-1)/t,\ (\sqrt{1+t}-1)/t,\ t\ln t\rangle$ and $\mathbf{u}(t) = \langle \sin^2(2t)/t^2,$ $t^2+2,\ (\cos t - 1)/t^2\rangle$ (see exercise 2) and find:
(i) $\lim_{t\to 0^+}(\mathbf{v}(t)+\mathbf{u}(t))$
(ii) $\lim_{t\to 0^+}(\mathbf{v}(t)\cdot\mathbf{u}(t))$
(iii) $\lim_{t\to 0^+}(\mathbf{v}(t)\times\mathbf{u}(t))$
Think of the amount of technicalities needed to obtain the answers without the laws of limits (e.g., calculating the cross product first and then finding the limit value).

(8) Find the values of the parameters a and b at which the curve $\mathbf{r}(t) = \langle 1+at^2, b-t, t^3\rangle$ passes through the point $(1,2,8)$.

(9) Let $\mathbf{r}(0) = \langle a,b,c\rangle$ and let $\mathbf{r}(t) = \langle t\cot(2t), t^{1/3}\ln|t|, t^2+2\rangle$ for $t\neq 0$. Find the values of a, b, and c at which the vector function is continuous.

(10) Suppose that the vector function $\mathbf{v}(t)\times\mathbf{u}(t)$ is continuous. Does this imply that both vector functions $\mathbf{v}(t)$ and $\mathbf{u}(t)$ are continuous? Support your arguments by examples.

80. Differentiation of Vector Functions

Definition 12.5. (Derivative of a Vector Function).
Suppose a vector function $\mathbf{r}(t)$ *is defined on an interval* $[a,b]$ *and* $t_0\in[a,b]$. *If the limit*

$$\lim_{h\to 0}\frac{\mathbf{r}(t_0+h)-\mathbf{r}(t_0)}{h} = \mathbf{r}'(t_0) = \frac{d\mathbf{r}}{dt}(t_0)$$

exists, then it is called the derivative of a vector function $\mathbf{r}(t)$ *at* $t=t_0$, *and* $\mathbf{r}(t)$ *is said to be differentiable at* t_0. *For* $t_0=a$ *or* $t_0=b$, *the limit is understood as the right* ($h>0$) *or left* ($h<0$) *limit, respectively. If the derivative exists for all points in* $[a,b]$, *then the vector function* $\mathbf{r}(t)$ *is said to be differentiable on* $[a,b]$.

It follows from the definition of the limit that a vector function is differentiable if and only if all its components are differentiable:

$$\mathbf{r}'(t) = \lim_{h\to 0}\Big\langle \frac{x(t+h)-x(t)}{h},\ \frac{y(t+h)-y(t)}{h},\ \frac{z(t+h)-z(t)}{h}\Big\rangle$$
$$= \langle x'(t)\,,\ y'(t)\,,\ z'(t)\rangle\,. \tag{12.1}$$

For example,

$$\mathbf{r}(t) = \langle \sin(2t),\ t^2-t,\ e^{-3t}\rangle \quad\Rightarrow\quad \mathbf{r}'(t) = \langle 2\cos(2t),\ 2t-1,\ -3e^{-3t}\rangle.$$

Definition 12.6. (Continuously Differentiable Vector Function).
If the derivative $\mathbf{r}'(t)$ *is a continuous vector function on an interval*

$[a, b]$, then the vector function $\mathbf{r}(t)$ is said to be continuously differentiable on $[a, b]$.

Higher-order derivatives are defined similarly: the second derivative is the derivative of $\mathbf{r}'(t)$, $\mathbf{r}''(t) = (\mathbf{r}'(t))'$, the third derivative is the derivative of $\mathbf{r}''(t)$, $\mathbf{r}'''(t) = (\mathbf{r}''(t))'$, and $\mathbf{r}^{(n)}(t) = (\mathbf{r}^{(n-1)}(t))'$, provided they exist.

80.1. Differentiation Rules.

The following rules of differentiation of vector functions can deduced from (12.1).

Theorem 12.2. (Differentiation Rules).
Suppose $\mathbf{u}(t)$ and $\mathbf{v}(t)$ are differentiable vector functions and $f(t)$ is a real-valued differentiable function. Then

$$\begin{aligned}\frac{d}{dt}\Big[\mathbf{v}(t) + \mathbf{u}(t)\Big] &= \mathbf{v}'(t) + \mathbf{u}'(t),\\ \frac{d}{dt}\Big[f(t)\mathbf{v}(t)\Big] &= f'(t)\mathbf{v}(t) + f(t)\mathbf{v}'(t),\\ \frac{d}{dt}\Big[\mathbf{v}(t)\cdot\mathbf{u}(t)\Big] &= \mathbf{v}'(t)\cdot\mathbf{u}(t) + \mathbf{v}(t)\cdot\mathbf{u}'(t),\\ \frac{d}{dt}\Big[\mathbf{v}(t)\times\mathbf{u}(t)\Big] &= \mathbf{v}'(t)\times\mathbf{u}(t) + \mathbf{v}(t)\times\mathbf{u}'(t).\\ \frac{d}{dt}\Big[\mathbf{v}(f(t))\Big] &= f'(t)\mathbf{v}'(f(t)).\end{aligned}$$

The proof is based on a straightforward use of the rule (12.1) and basic rules of differentiation for ordinary functions and left as an exercise to the reader.

80.2. Differential of a Vector Function.

If $\mathbf{r}(t)$ is differentiable, then

$$\Delta\mathbf{r}(t) = \mathbf{r}(t + \Delta t) - \mathbf{r}(t) = \mathbf{r}'(t)\,\Delta t + \mathbf{u}(\Delta t)\,\Delta t, \tag{12.2}$$

where $\mathbf{u}(\Delta t) \to \mathbf{0}$ as $\Delta t \to 0$. Indeed, by the definition of the derivative, $\mathbf{u}(\Delta t) = \Delta\mathbf{r}/\Delta t - \mathbf{r}'(t) \to \mathbf{0}$ as $\Delta t \to 0$. Therefore, the components of the difference $\Delta\mathbf{r} - \mathbf{r}'\,\Delta t$ converge to 0 faster than Δt. As for ordinary functions, situations in which all such terms can be neglected is described by the concept of the differential.

Definition 12.7. (Differential of a Vector Function).
Let $\mathbf{r}(t)$ be a differentiable vector function. Then the vector

$$d\mathbf{r}(t) = \mathbf{r}'(t)\,dt$$

is called the differential *of $\mathbf{r}(t)$.*

In particular, the derivative is the ratio of the differentials, $\mathbf{r}'(t) = d\mathbf{r}/dt$. In a small enough neighborhood of any particular $t = t_0$, a differentiable vector function can be well approximated by a linear vector function because $\Delta\mathbf{r} \approx d\mathbf{r}(t_0)$ or with $dt = \Delta t$:

$$\mathbf{r}(t_0 + \Delta t) \approx \mathbf{L}(t) = \mathbf{r}(t_0) + \mathbf{r}'(t_0)\Delta t, \quad \Delta t = t - t_0.$$

The linear vector function $\mathbf{L}(t)$ defines a line passing through the point $\mathbf{r}(t_0)$. This line is called the *tangent line* to the curve traced out by $\mathbf{r}(t)$. Thus, *the differential* $d\mathbf{r}(t)$ *at a point of the curve* $\mathbf{r}(t)$ *is the increment of the position vector along the line tangent to the curve at that point.*

80.3. Geometrical Significance of the Derivative. Consider a vector function that traces out a line parallel to a vector $\mathbf{v}$, $\mathbf{r}(t) = \mathbf{r}_0 + t\mathbf{v}$. Then $\mathbf{r}'(t) = \mathbf{v}$; that is, the derivative is a vector parallel or tangent to the line. This observation is of a general nature; that is, *the vector* $\mathbf{r}'(t_0)$ *is tangent to the curve traced out by* $\mathbf{r}(t)$ *at the point whose position vector is* $\mathbf{r}(t_0)$. Let P_0 and P_h have position vectors $\mathbf{r}(t_0)$ and $\mathbf{r}(t_0+h)$. Then $\overrightarrow{P_0P_h} = \mathbf{r}(t_0+h) - \mathbf{r}(t_0)$ is a secant vector. As $h \to 0$, $\overrightarrow{P_0P_h}$ approaches a vector that lies on the tangent line as depicted in Figure 12.6. On the other hand, it follows from (12.2) that, for small enough $h = dt$, $\overrightarrow{P_0P_h} = d\mathbf{r}(t_0) = \mathbf{r}'(t_0)h$, and therefore the tangent line is parallel to $\mathbf{r}'(t_0)$. The direction of the tangent vector also defines the orientation

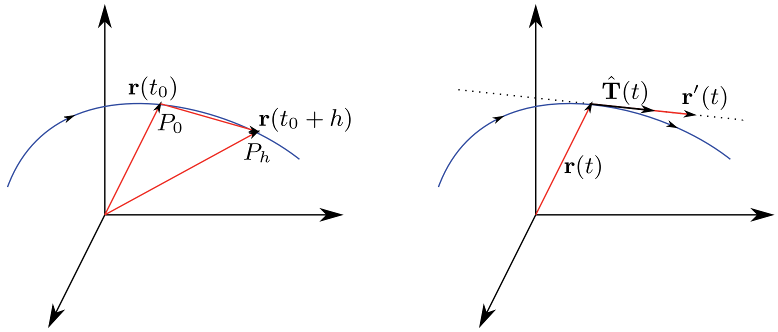

FIGURE 12.6. **Left**: A secant line through two points of the curve, P_0 and P_h. As h gets smaller, the direction of the vector $\overrightarrow{P_0P_h} = \mathbf{r}(t_0 + h) - \mathbf{r}(t_0)$ becomes closer to the tangent to the curve at P_0.
Right: The derivative $\mathbf{r}'(t)$ defines a tangent vector to the curve at the point with the position vector $\mathbf{r}(t)$. It also specifies the direction in which $\mathbf{r}(t)$ traverses the curve with increasing t. $\hat{\mathbf{T}}(t)$ is the unit tangent vector.

of the curve, that is, the direction in which the curve is traced out by $\mathbf{r}(t)$.

EXAMPLE 12.4. *Find the line tangent to the curve* $\mathbf{r}(t) = \langle 2t, t^2 - 1, t^3 + 2t\rangle$ *at the point* $P_0(2, 0, 3)$.

SOLUTION: By the geometrical property of the derivative, a vector parallel to the line is $\mathbf{v} = \mathbf{r}'(t_0)$, where t_0 is the value of the parameter t at which $\mathbf{r}(t_0) = \langle 2, 0, 3\rangle$ is the position vector of P_0. Therefore, $t_0 = 1$. Then $\mathbf{v} = \mathbf{r}'(1) = \langle 2, 2t, 3t + 2\rangle|_{t=1} = \langle 2, 2, 5\rangle$. Parametric equations of the line through P_0 and parallel to $\mathbf{v}$ are $x = 2 + 2t$, $y = 2t$, $z = 3 + 5t$. □

If the derivative $\mathbf{r}'(t)$ exists and does not vanish, then, at any point of the curve traced out by $\mathbf{r}(t)$, a *unit tangent vector* can be defined by

$$\hat{\mathbf{T}}(t) = \frac{\mathbf{r}'(t)}{\|\mathbf{r}'(t)\|}.$$

In Section 79.3, spatial curves were identified with continuous vector functions. Intuitively, a smooth curve as a point set in space should have a unit tangent vector that is continuous along the curve. Recall also that, for any curve as a point set in space, there are many vector functions whose range coincides with the curve.

DEFINITION 12.8. (Smooth Curve).
A point set in space is called a smooth curve *if there is a continuously differentiable vector function whose range coincides with the point set and whose derivative does not vanish.*

A smooth curve $\mathbf{r}(t)$ is *oriented* by the direction of the unit tangent vector $\hat{\mathbf{T}}(t)$.

Consider the planar curve $\mathbf{r} = \langle t^3, t^2, 0\rangle$. The vector function is differentiable everywhere, $\mathbf{r}'(t) = \langle 2t, 3t^2, 0\rangle$, and the derivative vanishes at the origin, $\mathbf{r}'(0) = \mathbf{0}$. The unit tangent vector $\hat{\mathbf{T}}(t)$ is not defined at $t = 0$. In the xy plane, the curve traces out the graph $y = x^{2/3}$, which has a *cusp* at $x = 0$. The graph is not smooth at the origin. The tangent line is the vertical line $x = 0$ because $y'(x) = (2/3)x^{-1/3} \to \pm\infty$ as $x \to 0^\pm$. The graph approaches it from the positive half-plane $y > 0$, forming a hornlike shape at the origin. A cusp does not necessarily occur at a point where the derivative $\mathbf{r}'(t)$ vanishes. For example, consider $\mathbf{r}(t) = \langle t^3, t^5, 0\rangle$ such that $\mathbf{r}'(0) = \mathbf{0}$. The curve traces out the graph $y = x^{5/3}$, which has no cusp at $x = 0$ (it has an inflection point at $x = 0$). There is another vector function $\mathbf{R}(s) = \langle s, s^{5/3}, 0\rangle$ that traces out the same graph, but $\mathbf{R}'(0) = \langle 1, 0, 0\rangle \neq \mathbf{0}$, and the curve is smooth. So the vanishing of the derivative is merely associated with a poor choice of the vector function. Note that $\mathbf{r}(t) = \mathbf{R}(s)$ identically if $s = t^3$. By the chain rule, $\frac{d}{dt}\mathbf{r}(t) = \frac{d}{dt}\mathbf{R}(s) = \mathbf{R}'(s)(ds/dt)$. This shows that, even if $\mathbf{R}'(s)$ never vanishes, the derivative $\mathbf{r}'(t)$ can vanish, provided ds/dt vanishes at some point, which is indeed the case in the considered example as $ds/dt = 3t^2$ vanishes at $t = 0$.

80.4. Study Problems.

Problem 12.7. *Prove that, for any smooth curve on a sphere, a tangent vector at any point P is perpendicular to the vector from the sphere center to P.*

Solution: Let $\mathbf{r}_0$ be the position vector of the center of a sphere of radius R. The position vector $\mathbf{r}$ of any point of the sphere satisfies the equation $\|\mathbf{r} - \mathbf{r}_0\| = R$ or $(\mathbf{r} - \mathbf{r}_0) \cdot (\mathbf{r} - \mathbf{r}_0) = R^2$ (because $\|\mathbf{a}\|^2 = \mathbf{a} \cdot \mathbf{a}$ for any vector $\mathbf{a}$). Let $\mathbf{r}(t)$ be a vector function that traces out a curve on the sphere. Then, for all values of t, $(\mathbf{r}(t) - \mathbf{r}_0) \cdot (\mathbf{r}(t) - \mathbf{r}_0) = R^2$. Differentiating both sides of the latter relation, one infers $\mathbf{r}'(t) \cdot (\mathbf{r}(t) - \mathbf{r}_0) = 0$. This algebraic condition is equivalent to the geometrical one: $\mathbf{r}'(t) \perp \mathbf{r}(t) - \mathbf{r}_0$. If $\mathbf{r}(t)$ is the position vector of P and O is the center of the sphere, then $\vec{OP} = \mathbf{r}(t) - \mathbf{r}_0$, and hence the tangent vector $\mathbf{r}'(t)$ at P is perpendicular to $\vec{OP}$ for any t or at any point P of the curve. □

80.5. Exercises. **(1)** Find the derivatives and differentials of each of the following curves:
(i) $\mathbf{r}(t) = \langle \cos t, \sin^2(t), t^2\rangle$
(ii) $\mathbf{r}(t) = \langle \ln(t), e^{2t}, te^{-t}\rangle$
(iii) $\mathbf{r}(t) = \langle \sqrt[3]{t-2}, \sqrt{t^2-4}, t\rangle$
(iv) $\mathbf{r}(t) = \mathbf{a} + \mathbf{b}t^2 - \mathbf{c}e^t$
(v) $\mathbf{r}(t) = t\mathbf{a} \times (\mathbf{b} - \mathbf{c}e^t)$

(2) Determine if the curve traced out by each of the following vector functions is smooth for a specified interval of the parameter. If the curve is not smooth at a particular point, graph it near that point.
(i) $\mathbf{r}(t) = \langle t, t^2, t^3\rangle$, $0 \le t \le 1$
(ii) $\mathbf{r}(t) = \langle t^2, t^3, 2\rangle$, $-1 \le t \le 1$
(iii) $\mathbf{r}(t) = \langle t^{1/3}, t, t^3\rangle$, $-1 \le t \le 1$
(vi) $\mathbf{r}(t) = \langle t^5, t^3, t^4\rangle$, $-1 \le t \le 1$

(3) Find the parametric equations of the tangent line to each of the following curves at a specified point:
(i) $\mathbf{r}(t) = \langle t^2 - t, t^3/3, 2t\rangle$, $P_0 = (6, 9, 6)$
(ii) $\mathbf{r}(t) = \langle \ln t, 2\sqrt{t}, t^2\rangle$, $P_0 = (0, 2, 1)$

(4) Is there a point on the curve $\mathbf{r}(t) = \langle t^2 - t, t^3/3, 2t\rangle$ at which the tangent line is parallel to the vector $\mathbf{v} = \langle -5/2, 2, 1\rangle$? If so, find the point.

(5) Let $\mathbf{r}(t) = \langle e^t, 2\cos t, \sin(2t)\rangle$. Use the tangent line approximation to find $\mathbf{r}(0.2)$. Use a calculator to assess the accuracy of the approximation.

(6) Suppose a smooth curve $\mathbf{r}(t)$ does not intersect a plane through a point P_0 and perpendicular to a vector $\mathbf{n}$. What is the angle between $\mathbf{n}$ and $\mathbf{r}'(t)$ at the point of the curve that is the closest to the plane?

(7) Does the curve $\mathbf{r}(t) = \langle 2t^2, 2t, 2 - t^2\rangle$ intersect the plane $x + y + z = -3$? If not, find a point on the curve that is closest to the plane. What is the distance between the curve and the plane.

(8) Find the point intersection of two curves $\mathbf{r}_1(t) = \langle 1, 1 - t, 3 + t^2\rangle$ and $\mathbf{r}_1(s) = \langle 3 - s, s - 2, s^2\rangle$. If the angle at which two curves intersect is defined as the angle between their tangent lines at the point of intersection, find the angle at which the above two curves intersect.

(9) Suppose $\mathbf{r}(t)$ is twice differentiable. Show that $(\mathbf{r}(t) \times \mathbf{r}'(t))' = \mathbf{r}(t) \times \mathbf{r}''(t)$.

(10) Suppose that $\mathbf{r}(t)$ is differentiable three times. Put $\mathbf{v} = \mathbf{r} \cdot (\mathbf{r}' \times \mathbf{r}'')$. Show that $\mathbf{v}' = \mathbf{r} \cdot (\mathbf{r}' \times \mathbf{r}''')$.

81. Integration of Vector Functions

DEFINITION 12.9. (Definite Integral of a Vector Function). *Let $\mathbf{r}(t)$ be defined on the interval $[a, b]$. The vector whose components are the definite integrals of the corresponding components of $\mathbf{r}(t) = \langle x(t), y(t), z(t)\rangle$ is called the* definite integral *of $\mathbf{r}(t)$ over the interval $[a, b]$ and denoted as*

$$\int_a^b \mathbf{r}(t)\,dt = \Big\langle \int_a^b x(t)\,dt\,,\ \int_a^b y(t)\,dt\,,\ \int_a^b x(t)\,dt \Big\rangle. \tag{12.3}$$

If the integral (12.3) exists, then $\mathbf{r}(t)$ is said to be integrable on $[a, b]$.

By this definition, a vector function is integrable if and only if all its components are integrable functions. Recall that a continuous real-valued function is integrable. Therefore, the following theorem holds.

THEOREM 12.3. *If a vector function is continuous on the interval $[a, b]$, then it is integrable on $[a, b]$.*

DEFINITION 12.10. (Indefinite Integral of a Vector Function). *A vector function $\mathbf{R}(t)$ is called an* indefinite integral *of $\mathbf{r}(t)$ if $\mathbf{R}'(t) = \mathbf{r}(t)$.*

If $\mathbf{R}(t) = \langle X(t), Y(t), Z(t)\rangle$ and $\mathbf{r}(t) = \langle x(t), y(t), z(t)\rangle$. Then, according to (12.1), the functions $X(t)$, $Y(t)$, and $Z(t)$ are antiderivatives of $x(t)$, $y(t)$, and $z(t)$, respectively,

$$X(t) = \int x(t)\, dt + c_1\,, \quad Y(t) = \int y(t)\, dt + c_2\,, \quad Z(t) = \int z(t)\, dt + c_3\,,$$

where c_1, c_2, and c_3 are constants. The latter relations can be combined into a single vector relation:

$$\mathbf{R}(t) = \int \mathbf{r}(t)\, dt + \mathbf{c}\,,$$

where $\mathbf{c}$ is an arbitrary constant vector. Applying the fundamental theorem of calculus to every component on the right side of (12.3), the fundamental theorem of calculus can be extended to vector functions.

Theorem 12.4. (Fundamental Theorem of Calculus for Vector Functions).
If $\mathbf{R}(t)$ is an indefinite integral of $\mathbf{r}(t)$, then

$$\int_a^b \mathbf{r}(t)\, dt = \mathbf{R}(b) - \mathbf{R}(a).$$

Example 12.5. *Find $\mathbf{r}(t)$ if $\mathbf{r}'(t) = \langle 2t, 1, 6t^2\rangle$ and $\mathbf{r}(1) = \langle 2, 1, 0\rangle$.*

Solution: Taking the antiderivative of $\mathbf{r}'(t)$, one finds $\mathbf{r}(t) = \int\langle 2t, 1, 6t^2\rangle\, dt + \mathbf{c} = \langle t^2, t, 3t^3\rangle + \mathbf{c}$. The constant vector $\mathbf{c}$ is determined by the condition $\mathbf{r}(1) = \langle 2, 1, 0\rangle$, which gives $\langle 1, 1, 3\rangle + \mathbf{c} = \langle 2, 1, 0\rangle$. Hence, $\mathbf{c} = \langle 1, 0, -3\rangle$ and $\mathbf{r}(t) = \langle t^2 + 1, t, 3t^3 - 3\rangle$. □

In general, the solution of the equation $\mathbf{r}'(t) = \mathbf{v}(t)$ satisfying the condition $\mathbf{r}(t_0) = \mathbf{r}_0$ can be written in the form

$$\mathbf{r}'(t) = \mathbf{v}(t) \;\text{ and }\; \mathbf{r}(t_0) = \mathbf{r}_0 \quad\Rightarrow\quad \mathbf{r}(t) = \mathbf{r}_0 + \int_{t_0}^{t} \mathbf{v}(s)\, ds$$

if $\mathbf{v}(t)$ is a continuous vector function. Recall that if the integrand is a continuous function, then the derivative of the integral with respect to its upper limit is the value of the integrand at that limit. Therefore, $(d/dt)\int_{t_0}^{t} \mathbf{v}(s)\, ds = \mathbf{v}(t)$, and hence $\mathbf{r}(t)$ is an antiderivative of $\mathbf{v}(t)$. When $t = t_0$, the integral vanishes and $\mathbf{r}(t_0) = \mathbf{r}_0$ as required.

81.1. Applications to Mechanics. Let $\mathbf{r}(t)$ be the position vector of a particle as a function of time t. The first derivative $\mathbf{r}'(t) = \mathbf{v}(t)$ is called the *velocity* of the particle. The magnitude of the velocity vector $v(t) = \|\mathbf{v}(t)\|$ is called the *speed*. The speed of a car is a number shown on the speedometer. The velocity defines the direction in which the particle travels and the instantaneous rate at which it moves in that direction. The second derivative $\mathbf{r}''(t) = \mathbf{v}'(t) = \mathbf{a}(t)$ is called the *acceleration.* If m is the mass of a particle and $\mathbf{F}$ is the force acting on the particle, according to Newton's second law, the acceleration and force are related as

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

If the force is known as a vector function of time, then this relation is an equation of motion that determines a particle's trajectory. The problem of finding the trajectory amounts to reconstructing the vector function $\mathbf{r}(t)$ if its second derivative $\mathbf{r}''(t) = (1/m)\mathbf{F}(t)$ is known; that is, $\mathbf{r}(t)$ is given by the second antiderivative of $(1/m)\mathbf{F}(t)$. Indeed, the velocity $\mathbf{v}(t)$ is an antiderivative of $(1/m)\mathbf{F}(t)$, and the position vector $\mathbf{r}(t)$ is an antiderivative of the velocity $\mathbf{v}(t)$. As shown in the previous section, an antiderivative is not unique, unless its value at a particular point is specified. So *the trajectory of motion is uniquely determined by Newton's equation, provided the position and velocity vectors are specified at a particular moment of time*, for example, $\mathbf{r}(t_0) = \mathbf{r}_0$ and $\mathbf{v}(t_0) = \mathbf{v}_0$. The latter conditions are called *initial conditions.* Given the initial conditions, the trajectory of motion is uniquely defined by the relations:

$$\mathbf{v}(t) = \mathbf{v}_0 + \frac{1}{m}\int_{t_0}^{t} \mathbf{F}(s)\,ds\,, \quad \mathbf{r}(t) = \mathbf{r}_0 + \int_{t_0}^{t} \mathbf{v}(s)\,ds$$

if the force is a continuous vector function of time.

Remark. If the force is a function of a particle's position, then the Newton's equation becomes a system of *ordinary differential equations* that is a set of some relations between components of the vector functions, its derivatives, and time.

Example 12.6. (Motion Under a Constant Force).
Prove that the trajectory of motion under a constant force is a parabola.

Solution: Let $\mathbf{F}$ be a constant force. Without loss of generality, the initial conditions can be set at $t = 0$, $\mathbf{r}(0) = \mathbf{r}_0$, and $\mathbf{v}(0) = \mathbf{v}_0$. Then

$$\mathbf{v}(t) = \mathbf{v}_0 + \frac{t}{m}\,\mathbf{F}\,, \quad \mathbf{r}(t) = \mathbf{r}_0 + t\mathbf{v}_0 + \frac{t^2}{2m}\,\mathbf{F}\,.$$

The vector $\mathbf{r}(t)-\mathbf{r}_0$ is a linear combination of $\mathbf{v}_0$ and $\mathbf{F}$ and hence must be perpendicular to $\mathbf{n} = \mathbf{v}_0\times\mathbf{F}$ by the geometrical property of the cross product. Therefore, the particle remains in the plane through $\mathbf{r}_0$ that is parallel to $\mathbf{F}$ and $\mathbf{v}_0$ or perpendicular to $\mathbf{n}$, that is, $(\mathbf{r}(t)-\mathbf{r}_0)\cdot\mathbf{n} = 0$ (see Figure 12.7, left panel). The shape of a space curve does not depend on the choice of the coordinate system. Let us choose the coordinate system such that the origin is at the initial position $\mathbf{r}_0$ and the plane

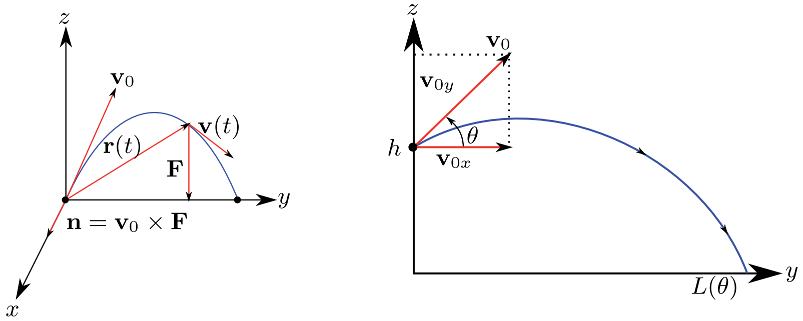

FIGURE 12.7. **Left**: Motion under a constant force $\mathbf{F}_0$. The trajectory is a parabola that lies in the plane through the initial point of the motion $\mathbf{r}_0$ and perpendicular to the vector $\mathbf{n} = \mathbf{v}_0 \times \mathbf{F}$, where $\mathbf{v}_0$ is the initial velocity. **Right**: Motion of a projectile thrown at an angle θ and an initial height h. The trajectory is a parabola. The point of impact defines the range $L(\theta)$.

in which the trajectory lies coincides with the zy plane so that $\mathbf{F}$ is parallel to the z axis. In this coordinate system, $\mathbf{r}_0 = \mathbf{0}$, $\mathbf{F} = \langle 0, 0, -F\rangle$, and $\mathbf{v}_0 = \langle 0, v_{0y}, v_{0z}\rangle$. The parametric equations of the trajectory of motion assume the form $x = 0$, $y = v_{0y}t$, and $z = v_{0z}t - t^2F/(2m)$. The substitution of $t = y/v_{0y}$ into the latter equation yields $z = ay^2 + by$, where $a = -Fv_{0y}^2/(2m)$ and $b = v_{0z}/v_{0y}$, which defines a parabola in the zy plane. Thus, the trajectory of motion under a constant force is a parabola through the point $\mathbf{r}_0$ that lies in the plane containing the force and initial velocity vectors $\mathbf{F}$ and $\mathbf{v}_0$. □

81.2. Motion Under a Constant Gravitational Force. The magnitude of the gravitational force that acts on an object of mass m near the surface of the Earth is mg, where $g \approx 9.8 \text{ m/s}^2$ is a universal constant called the *acceleration of a free fall*. According to the previous section, any projectile fired from some point follows a parabolic trajectory. This fact allows one to predict the exact positions of the projectile and, in particular, the point at which it impacts the ground. In practice, the initial speed v_0 of the projectile and angle of elevation θ at which the projectile is fired are known (see Figure 12.7, right panel). Some practical questions are: At what elevation angle is the maximal range reached? At what elevation angle does the range attain a specified value (e.g., to hit a target)?

To answer these and related questions, choose the coordinate system such that the z axis is directed upward from the ground and the parabolic trajectory lies in the zy plane. The projectile is fired from the point $(0, 0, h)$, where h is the initial elevation of the projectile above the ground (firing from a hill). In the notation of the previous section, $F = -mg$ (F is negative because the gravitational force is directed toward the ground, while the z axis points upward), $v_{0y} = v_0 \cos\theta$, and $v_{0z} = v_0 \sin\theta$. The trajectory is

$$y = tv_0 \cos\theta\,, \quad z = h + tv_0 \sin\theta - \frac{1}{2}gt^2\,, \quad t \geq 0\,.$$

It is interesting to note that the trajectory is independent of the mass of the projectile. Light and heavy projectiles would follow the same parabolic trajectory, provided they are fired from the same position, at the same speed, and at the same angle of elevation. The height of the projectile relative to the ground is given by $z(t)$. The horizontal displacement is $y(t)$. Let $t_L > 0$ be the moment of time when the projectile lands; that is, when $t = t_L$, the height vanishes, $z(t_L) = 0$. A positive solution of this equation is

$$t_L = \frac{v_0 \sin\theta + \sqrt{v_0^2 \sin^2\theta + 2gh}}{g}\,.$$

The distance L traveled by the projectile in the horizontal direction until it lands is the *range*:

$$L = y(t_L) = t_L v_0 \cos\theta\,.$$

For example, if the projectile is fired from the ground, $h = 0$, then $t_L = 2v_0 \sin\theta/g$ and the range is $L = v_0^2 \sin(2\theta)/g$. The range attains its maximal value v_0^2/g when the projectile is fired at an angle of elevation $\theta = \pi/4$. The angle of elevation at which the projectile hits a target at a given range $L = L_0$ is $\theta = (1/2)\sin^{-1}(L_0 g/v_0^2)$. For $h \neq 0$, the angle at which $L = L(\theta)$ attains its maximal values can be found by solving the equation $L'(\theta) = 0$, which defines critical points of the function $L(\theta)$. The angle of elevation at which the projectile hits a target at a given range is found by solving the equation $L(\theta) = L_0$. The technicalities are left to the reader.

Remark. In reality, the trajectory of a projectile deviates from a parabola because there is an additional force acting on a projectile moving in the atmosphere, the friction force. The friction force depends on the velocity of the projectile. So a more accurate analysis of the projectile motion in the atmosphere requires methods of ordinary differential equations.

81.3. Study Problems.

Problem 12.8. *The acceleration of a particle is* $\mathbf{a} = \langle 2, 6t, 0\rangle$. *Find the position vector of the particle and its velocity in two units of time* t *if the particle was initially at the point* $(-1, -4, 1)$ *and had the velocity* $\langle 0, 2, 1\rangle$.

Solution: The velocity vector is $\mathbf{v}(t) = \int \mathbf{a}(t)\,dt + \mathbf{c} = \langle 2t, 3t^2, 0\rangle + \mathbf{c}$. The constant vector $\mathbf{c}$ is fixed by the initial condition $\mathbf{v}(0) = \langle 0, 2, 1\rangle$, which yields $\mathbf{c} = \langle 0, 2, 1\rangle$. Thus, $\mathbf{v}(t) = \langle 2t, 3t^2 + 2, 1\rangle$ and $\mathbf{v}(2) = \langle 2, 8, 1\rangle$. The position vector is $\mathbf{r}(t) = \int \mathbf{v}(t)\,dt + \mathbf{c} = \langle t^2, t^3 + 2t, t\rangle + \mathbf{c}$. Here the constant vector $\mathbf{c}$ is determined by the initial condition $\mathbf{r}(0) = \langle 0, 2, 1\rangle$, which yields $\mathbf{c} = \langle -1, -4, 1\rangle$. Thus, $\mathbf{r}(t) = \langle t^2 - 1, t^3 + 2t - 4, t + 1\rangle$ and $\mathbf{r}(2) = \langle 3, 4, 3\rangle$. □

Problem 12.9. *Show that if the velocity and position vectors of a particle remains orthogonal during the motion, then the trajectory lies on a sphere.*

Solution: If $\mathbf{v}(t) = \mathbf{r}'(t)$ and $\mathbf{r}(t)$ are orthogonal, then $\mathbf{r}'(t)\cdot\mathbf{r}(t) = 0$ for all t. Since $(\mathbf{r}\cdot\mathbf{r})' = \mathbf{r}'\cdot\mathbf{r} + \mathbf{r}\cdot\mathbf{r}' = 2\mathbf{r}'\cdot\mathbf{r} = 0$, one concludes that $\mathbf{r}(t)\cdot\mathbf{r}(t) = R^2 = \text{const}$ or $\|\mathbf{r}(t)\| = R$ for all t; that is, the particle remains at a fixed distance R from the origin all the time. □

Problem 12.10. *A. charged particle moving in a magnetic field* $\mathbf{B}$ *is subject to the Lorentz force* $\mathbf{F} = (e/c)\mathbf{v}\times\mathbf{B}$, *where* e *is the electric charge of the particle and* c *is the speed of light in vacuum. Assume that the magnetic field is a constant vector parallel to the* z *axis and the initial velocity is* $\mathbf{v}(0) = \langle v_\perp, 0, v_\parallel\rangle$. *Show that the trajectory is a helix:*

$$\mathbf{r}(t) = \langle R\sin(\omega t),\ R\cos(\omega t),\ v_\parallel t\rangle, \quad \omega = \frac{eB}{mc}, \quad R = \frac{v_\perp}{\omega},$$

where $B = \|\mathbf{B}\|$ *is the magnetic field magnitude and* m *is the particle mass.*

Solution: Newton's second law reads

$$m\mathbf{v}' = \frac{e}{c}\mathbf{v}\times\mathbf{B}.$$

Put $\mathbf{B} = \langle 0, 0, B\rangle$. Then

$$\begin{aligned}\mathbf{v} &= \mathbf{r}' = \langle \omega R\cos(\omega t), -\omega R\sin(\omega t), v_\parallel\rangle,\\ \mathbf{v}\times\mathbf{B} &= \langle -\omega B\cos(\omega t), -\omega R\sin(\omega t), 0\rangle,\\ \mathbf{v}' &= \langle -\omega^2 R\cos(\omega t), -\omega^2 R\sin(\omega t), 0\rangle.\end{aligned}$$

The substitution of these relations into Newton's second law yields $m\omega^2 R = eBR\omega/c$ and hence $\omega = (eB)/(mc)$. Since $\mathbf{v}(0) = \langle \omega R, 0, v_\| \rangle = \langle v_\perp, 0, v_\| \rangle$, it follows that $R = v_\perp/\omega$. □

Remark. Note that the equations of motion involve only the velocity $\mathbf{v}$. For this reason, the velocity vector is uniquely determined by the initial condition $\mathbf{v}(0) = \mathbf{v}_0$, while the initial condition for the position vector is not needed (the vector function $\mathbf{r}(t) + \mathbf{r}_0$ is also a solution for an arbitrary constant vector $\mathbf{r}_0$). The rate at which the helix rises along the magnetic field is determined by the magnitude (speed) of the initial velocity component $v_\|$ parallel to the magnetic field, whereas the radius of the helix is determined by the magnitude of the initial velocity component $v_\perp$ perpendicular to the magnetic field. A particle makes one full turn about the magnetic field in time $T = 2\pi/\omega = 2\pi mc/(eB)$, that is, the larger the magnetic field, the faster the particle rotates about it.

81.4. Exercises. **(1)** Find the indefinite and definite integrals over a specified interval for each of the following functions:
(i) $\mathbf{r}(t) = \langle \sin t, t^3, \cos t \rangle$, $-\pi \le t \le \pi$
(ii) $\mathbf{r}(t) = \langle t^2, t\sqrt{t-1}, \sqrt{t} \rangle$, $0 \le t \le 1$
(iii) $\mathbf{r}(t) = \langle t \ln t, t^2, e^{2t} \rangle$, $0 \le t \le 0$

(2) Find $\mathbf{r}(t)$ if the derivatives $\mathbf{r}'(t)$ and $\mathbf{r}(t_0)$ are given:
(i) $\mathbf{r}'(t) = \langle t-1, t^2, \sqrt{t} \rangle$, $\mathbf{r}(1) = \langle 1, 0, 1 \rangle$
(ii) $\mathbf{r}'(t) = \langle \sin(2t), 2\cos t, \sin^2 t \rangle$, $\mathbf{r}(\pi) = \langle 1, 2, 3 \rangle$

(3) If a particle was initially at point $(1, 2, 1)$ and had velocity $\mathbf{v} = \langle 0, 1, -1 \rangle$. Find the position vector of the particle after it has been moving with acceleration $\mathbf{a}(t) = \langle 1, 0, t \rangle$ for 2 units of time.

(4) A particle of unit mass moves under a constant force $\mathbf{F}$. If a particle was initially at the point $\mathbf{r}_0$ and passed through the point $\mathbf{r}_1$ after 2 units of time, find the initial velocity of the particle. What was the velocity of the particle when it passed through $\mathbf{r}_1$?

(5) The position vector of a particle is $\mathbf{r}(t) = \langle t^2, 5t, t^2 - 16t \rangle$. Find $\mathbf{r}(t)$ when the speed of the particle is maximal.

(6) A projectile is fired at an initial speed of 400 m/s and at an angle of elevation of 30°. Find the range of the projectile, the maximum height reached, and the speed at impact.

(7) A ball of mass m is thrown southward into the air at an initial speed of v_0 at an angle of θ to the ground. An east wind applies a steady force of magnitude F to the ball in a westerly direction. Find the trajectory of the ball. Where does the ball land and at what speed? Find the deviation of the impact point from the impact point A when no wind is present. Is there any way to correct the direction in which the ball is thrown so that the ball still hits A?

(8) A rocket burns its onboard fuel while moving through space. Let $\mathbf{v}(t)$ and $m(t)$ be the velocity and mass of the rocket at time t. It can be shown that the force exerted by the rocket jet engines is $m'(t)\mathbf{v}_g$, where $\mathbf{v}_g$ is the velocity of the exhaust gases relative to the rocket. Show that $\mathbf{v}(t) = \mathbf{v}(0) - \ln(m(0)/m(t))\mathbf{v}_g$. The rocket is to accelerate in a straight line from rest to twice the speed of its own exhaust gases. What fraction of its initial mass would the rocket have to burn as fuel?

(9) The acceleration of a projectile is $\mathbf{a}(t) = \langle 0, 2, 6t\rangle$. The projectile is shot from $(0, 0, 0)$ with an initial velocity $\mathbf{v}(0) = \langle 1, -2, -10\rangle$. It is supposed to destroy a target located at $(2, 0, -12)$. The target can be destroyed if the projectile's speed is at least 3.1 at impact. Will the target be destroyed?

82. Arc Length of a Curve

Consider a partition of a curve C, that is, a collection of points of C, P_k, $k = 0, 1, ..., N$, where P_0 and P_N are the endpoints of the curve. A partition is said to be *refined* if the number of partition points increases so that $\max_k |P_{k-1}P_k| \to 0$ as $N \to \infty$; here $|P_{k-1}P_k|$ is the distance between the points P_{k-1} and P_k.

Definition 12.11. (Arc Length of a Curve).
The arc length of a curve C is the limit

$$L = \lim_{N\to\infty} \sum_{k=1}^{N} |P_{k-1}P_k|,$$

provided it exists and is independent of the choice of partition. If $L < \infty$, the curve is called measurable *or* rectifiable.

The geometrical meaning of this definition is rather simple. Here the sum of $|P_{k-1}P_k|$ is the length of a polygonal path with vertices at P_0, P_1,..., P_N in this order. As the partition becomes finer and finer, this polygonal path approaches the curve more and more closely (see Figure 12.8, left panel). In certain cases, the arc length is given by the Riemann integral.

Theorem 12.5. (Arc Length of a Curve).
Let C be a curve traced out by a continuously differentiable vector function $\mathbf{r}(t)$, which defines a one-to-one correspondence between points of C and the interval $t \in [a, b]$. Then

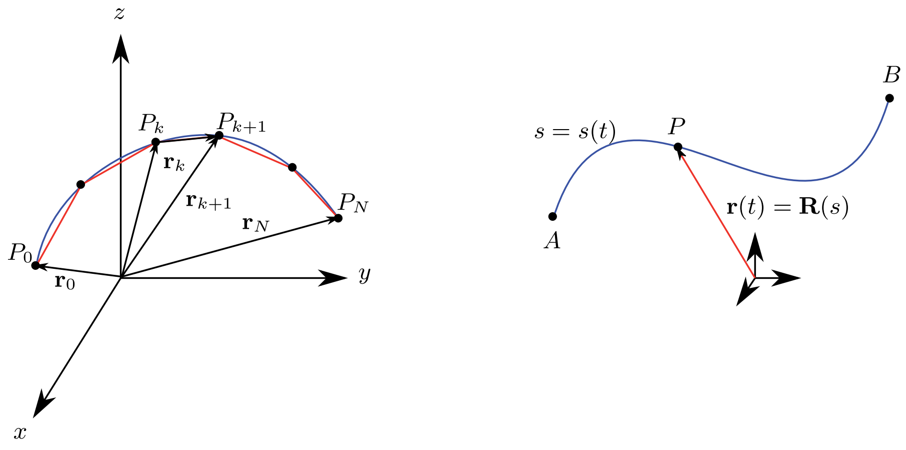

FIGURE 12.8. **Left**: The arc length of a curve is defined as the limit of the sequence of lengths of polygonal paths through partition points of the curve.
Right: Natural parameterization of a curve. Given a point A of the curve, the arc length s is counted from it to any point P of the curve. The position vector of P is a vector $\mathbf{R}(s)$. If the curve is traced out by another vector function $\mathbf{r}(t)$, then there is a relation $s = s(t)$ such that $\mathbf{r}(t) = \mathbf{R}(s(t))$.

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

PROOF. For any partition P_k of C, there is a partition t_k of $[a, b]$ such that $t_0 = a < t_1 < \cdots < t_{N-1} < t_N = b$ and $\mathbf{r}_k = \mathbf{r}(t_k)$ are position vectors of P_k, $k = 0, 1, ..., N$. Put $\Delta t_k = t_k - t_{k-1} > 0$, $k = 1, 2, ..., N$. In the limit $N \to \infty$, $\Delta t_k \to 0$ because $\mathbf{r}_k - \mathbf{r}_{k-1} \to \mathbf{0}$ for all k. Let $\mathbf{r}'_{k-1} = \mathbf{r}'(t_{k-1})$. The differentiability of $\mathbf{r}(t)$ implies that $\mathbf{r}_k - \mathbf{r}_{k-1} = \mathbf{r}'_{k-1}\,\Delta t_k + \mathbf{u}_k\,\Delta t_k$, where $\mathbf{u}_k \to \mathbf{0}$ as $\Delta t_k \to 0$ for every k (cf. (12.2)). Then, by the triangle inequality (11.7),

$$\|\mathbf{r}'_{k-1}\|\Delta t_k - \|\mathbf{u}_k\|\Delta t_k \le \|\mathbf{r}_k - \mathbf{r}_{k-1}\| \le \|\mathbf{r}'_{k-1}\|\Delta t_k + \|\mathbf{u}_k\|\Delta t_k.$$

By the continuity of the derivative, the function $\|\mathbf{r}'(t)\|$ is continuous and hence integrable. Therefore, its Riemann sum converges:

$$\sum_{k=1}^{N} \|\mathbf{r}'_{k-1}\|\Delta t_k \to \int_a^b \|\mathbf{r}'(t)\|\, dt \quad \text{as} \quad N \to \infty\,.$$

Put $\max_k \|\mathbf{u}_k\| = M_N$ (the largest $\|\mathbf{u}_k\|$ for a given partition size N). Then

$$\sum_{k=1}^{N} \|\mathbf{u}_k\|\Delta t_k \le M_N \sum_{k=1}^{N} \Delta t_k = M_N(b-a) \to 0 \quad \text{as} \quad N \to \infty$$

because $\|\mathbf{u}_k\| \to 0$ as $\Delta t_k \to 0$ for all k, and hence $M_N \to 0$ as $N \to \infty$. It follows from the squeeze principle that the limit of $\sum_{k=1}^{N} \|\mathbf{r}_k - \mathbf{r}_{k-1}\|$ as $N \to \infty$ exists and equals $\int_a^b \|\mathbf{r}'(t)\|\,dt$. □

Remark. Let $\mathbf{r}(t)$ be continuously differentiable on $[a,b]$ but does not necessarily define a one-to-one correspondence with its range C. Then the integral $\int_a^b \|\mathbf{r}'(t)\|\,dt$ is not the length of the curve C as a point set in space because $\mathbf{r}(t)$ may traverse a part of C several times. However, it is also useful in practical applications. Suppose $\mathbf{r}(t)$ is a trajectory of a particle. Then its velocity is $\mathbf{v}(t) = \mathbf{r}'(t)$ and its speed is $v(t) = \|\mathbf{v}(t)\|$. The distance traveled by the particle in the time interval $[a,b]$ is given by

$$D = \int_a^b v(t)\,dt = \int_a^b \|\mathbf{r}'(t)\|\,dt.$$

If a particle travels along the same space curve (or some of its parts) several times, then the distance traveled does not coincide with the arc length of the curve.

Example 12.7. *Find the arc length of one turn of a helix of radius R that rises by h per each turn.*

Solution: Let the helix axis be the z axis. The helix is traced out by the vector function $\mathbf{r}(t) = \langle R\cos t, R\sin t, th/(2\pi)\rangle$. One turn corresponds to the interval $t \in [0, 2\pi]$. Therefore,

$$\|\mathbf{r}'(t)\| = \|\langle -R\sin t,\ R\cos t, h/(2\pi)\rangle\| = \sqrt{R^2 + (h/(2\pi))^2}\,.$$

So the norm of the derivative turns out be constant. The arc length is

$$L = \int_0^{2\pi} \|\mathbf{r}'(t)\|\,dt = \sqrt{R^2 + (h/(2\pi))^2}\int_0^{2\pi} dt = \sqrt{(2\pi R)^2 + h^2}\,.$$

This result is rather easy to obtain without calculus. The helix lies on a cylinder of radius R. If the cylinder is cut parallel to its axis and unfolded into a strip, then one turn of the helix becomes the hypotenuse of the right-angled triangle with catheti $2\pi R$ and h. The result follows from the Pythagorean theorem. □

82.1. Reparameterization of a Curve. In Section 79, it was shown that a space curve defined as a point set in space can be traced by different vector functions. For example, a semicircle of radius R is traced out by the vector functions

$$\mathbf{r}(t) = \langle R\cos t,\ R\sin t,\ 0\rangle\,, \quad t \in [0,\pi]\,,$$
$$\mathbf{R}(u) = \langle u,\ \sqrt{R^2-u^2},\ 0\rangle\,, \quad u \in [-R,R]\,.$$

These vector functions are related to one another by the composition rule:

$$\mathbf{R}(u) = \mathbf{r}(t(u))\,, \quad t(u) = \cos^{-1}(u/R) \quad \text{or}$$
$$\mathbf{r}(t) = \mathbf{R}(u(t))\,, \quad u(t) = R\cos t\,.$$

This example illustrates the concept of a reparameterization of a curve. A reparameterization of a curve is a change of the parameter that labels points of the curve.

Definition 12.12. (Reparameterization of a Curve).
Let $\mathbf{r}(t)$ trace out a curve C if $t \in [a,b]$. Consider a one-to-one mapping $[a,b] \to [a',b']$, that is, a function $u = u(t)$ with the domain $[a,b]$ and the range $[a',b']$ that has the inverse $t = t(u)$. The vector function $\mathbf{R}(u) = \mathbf{r}(t(u))$ is called a reparameterization *of C.*

It should be emphasized that the geometrical properties of the curve (e.g., its shape or length) do not depend on a parameterization of the curve because the vector functions $\mathbf{r}(t)$ and $\mathbf{R}(u)$ have the *same* range. A reparameterization of a curve is a technical tool to find an algebraic description of the curve convenient for particular applications.

82.2. A Natural Parameterization of a Smooth Curve. Suppose one is traveling along a highway from town A to town B and comes upon an accident. How can the location of the accident be reported to the police? If one has a GPS navigator, one can report coordinates on the surface of the Earth. This implies that the police should use a specific (GPS) coordinate system to locate the accident. Is it possible to avoid any reference to a coordinate system? A simpler way to define the position of the accident is to report the distance traveled from A along the highway to the point where the accident happened (by using, e.g., mile markers). No coordinate system is needed to uniquely label all points of the highway by specifying the distance from a particular point A to the point of interest along the highway. This observation can be extended to all smooth curves (see Figure 12.8, right panel).

Definition 12.13. (Natural or Arc Length Parameterization). *Let C be a smooth curve of length L between points A and B. Let $\mathbf{r}(t)$, $t \in [a,b]$, be a one-to-one vector function that traces out C so that $\mathbf{r}(a)$ and $\mathbf{r}(b)$ are position vectors of A and B, respectively. Then the arc length $s = s(t)$ of the portion of the curve between $\mathbf{r}(a)$ and $\mathbf{r}(t)$ is a function of the parameter t:*

$$s = s(t) = \int_a^t \|\mathbf{r}'(u)\| du\,, \quad s \in [0, L]\,.$$

The vector function $\mathbf{R}(s) = \mathbf{r}(t(s))$ is called a natural or arc length parameterization *of C, where $t(s)$ is the inverse function of $s(t)$.*

Example 12.8. *Find the coordinates of a point P that is $5\pi/3$ units of length away from the point $(4, 0, 0)$ along the helix $\mathbf{r}(t) = \langle 4\cos(\pi t), 4\sin(\pi t), 3\pi t\rangle$.*

Solution: The initial point of the helix corresponds to $t = 0$. So the arc length counted from $(4, 0, 0)$ as a function of t is

$$s(t) = \int_0^t \|\mathbf{r}'(u)\|\, du = \int_0^t 5\pi\, du = 5\pi t$$

because $\mathbf{r}'(u) = \langle -4\pi\sin(\pi u), 4\pi\cos(\pi u), 3\pi\rangle$ and therefore $\|\mathbf{r}'(u)\| = 5\pi$. Hence, the inverse is $t = s/(5\pi)$, and the natural parameterization reads $\mathbf{R}(s) = \mathbf{r}(t(s)) = \langle 4\cos(s/5), 4\sin(s/5), 3s/5\rangle$. The position vector of P is $\mathbf{R}(5\pi/3) = \langle 2, 2\sqrt{3}, \pi\rangle$. Note that there are two points of the helix at the specified distance from $(4, 0, 0)$. One such point is upward along the helix, and the other is downward along it. The downward point corresponds to $t < 0$. Hence, its position vector is $\mathbf{R}(-5\pi/3) = \langle 2, -2\sqrt{3}, -\pi\rangle$. □

By definition, the arc length is independent of a parameterization of a space curve. For smooth curves, this can also be established through the change of variables in the integral that determines the arc length. Indeed, let $\mathbf{r}(t)$, $t \in [a, b]$, be a one-to-one continuously differentiable vector function that traces out a curve C of length L. Consider the change of the integration variable $t = t(s)$, $s \in [0, L]$. Then $ds = s'(t)\, dt = \|\mathbf{r}'(t)\|\, dt$ (by differentiating the integral for $s(t)$ with respect to the upper limit) and

$$L = \int_a^b \|\mathbf{r}'(t)\|\, dt = \int_0^L ds$$

for any parameterization of the curve C.

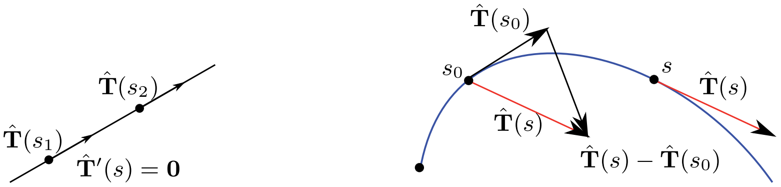

FIGURE 12.9. **Left**: A straight line does not bend. The unit tangent vector has zero rate of change relative to the arc length parameter s.
Right: Curvature of a smooth curve. The more a smooth curve bends, the larger the rate of change of the unit tangent vector relative to the arc length parameter becomes. So the magnitude of the derivative (curvature) $\|\hat{\mathbf{T}}'(s)\| = \kappa(s)$ can be taken as a geometrical measure of bending.

82.3. Exercises. **(1)** Find the arc length of each of the following curves:
(i) $\mathbf{r}(t) = \langle 3\cos t, 2t, 3\sin t\rangle$, $-2 \le t \le 2$
(ii) $\mathbf{r}(t) = \langle 2t, t^3/3, t^2\rangle$, $0 \le t \le 1$

(2) Find the arc length of the portion of the helix $\mathbf{r}(t) = \langle \cos t, \sin t, t\rangle$ that lies inside the sphere $x^2 + y^2 + z^2 = 2$.

(3) Find the arc length of the portion of the curve $\mathbf{r}(t) = \langle 2t, 3t^2, 3t^3\rangle$ that lies between the planes $z = 3$ and $z = 24$.

(4) Let C be the curve of intersection of the surfaces $z^2 = 2y$ and $3x = yz$. Find the length of C from the origin to the point $(36, 18, 6)$.

(5) Reparameterize each of the following curves with respect to the arc length measure from the point where $t = 0$ in the direction of increasing t:
(i) $\mathbf{r} = \langle t, 1 - 2t, 5 + 3t\rangle$
(ii) $\mathbf{r} = \frac{2t}{t^2+1}\hat{\mathbf{e}}_1 + (\frac{2}{t^2+1} - 1)\hat{\mathbf{e}}_3$

(6) A particle travels along a helix of radius R that rises h units of length per turn. Let the z axis be the symmetry axis of the helix. If a particle travels the distance $4\pi R$ from the point $(R, 0, 0)$, find the position vector of the particle.

83. Curvature of a Space Curve

Consider two curves passing through a point P. Both curves bend at P. Which one bends more than the other and how much more? The answer to this question requires a numerical characterization of bending, that is, a number computed at P for each curve with the property that it becomes larger as the curve bends more. Naturally, this number should not depend on a parameterization of a curve. Suppose that a curve is smooth so that a unit tangent vector can be attached to every point of the curve. A straight line does not bend (does not "curve") so it has the same unit tangent vector at all its points. If a curve bends, then its unit tangent vector becomes a function of its position on the curve. The position on the curve can be specified in a coordinate- and parameterization-independent way by the arc length s counted from a particular point of the curve. If $\hat{\mathbf{T}}(s)$ is the unit tangent vector as a function of s, then its derivative $\hat{\mathbf{T}}'(s)$ vanishes for a straight line (see Figure 12.9), while this would not be the case for a general curve. From the definition of the derivative

$$\hat{\mathbf{T}}'(s_0) = \lim_{s\to s_0} \frac{\hat{\mathbf{T}}(s) - \hat{\mathbf{T}}(s_0)}{s - s_0},$$

it follows that the magnitude $\|\hat{\mathbf{T}}'(s_0)\|$ becomes larger when the curve "bends more." For a fixed distance $s - s_0$ between two neighboring points of the curve, the magnitude $\|\hat{\mathbf{T}}(s) - \hat{\mathbf{T}}(s_0)\|$ becomes larger when the curve bends more at the point corresponding to s_0. So the number $\|\mathbf{T}'(s_0)\|$ can be used as a numerical measure of the bending or *curvature* of a curve.

DEFINITION 12.14. (Curvature of a Smooth Curve).
Let C be a smooth curve and let $\hat{\mathbf{T}}(s)$ be the unit tangent vector as a differentiable function of the arc length counted from a particular point of C. The number

$$\kappa(s) = \left\|\frac{d}{ds}\mathbf{T}(s)\right\|$$

is called the curvature *of C at the point corresponding to the value s of the arc length.*

In practice, a curve may not be given in the natural parameterization. Therefore, a question of interest is to find a method to calculate the curvature in any parameterization.

Let $\mathbf{r}(t)$ be a vector function in $[a, b]$ that traces out a curve C such that the arc length parameter can be defined as a function of t, $s = s(t)$, and it has the inverse function $t = t(s)$. The unit tangent vector as a function of the parameter t has the form $\hat{\mathbf{T}}(t) = \mathbf{r}'(t)/\|\mathbf{r}'(t)\|$. So, to calculate the curvature as a function of t, the relation between the derivatives d/ds and d/dt has to be found. The graphs of $s(t)$ and its inverse are obtained from one another by the reflection about the line $s = t$. Let (t_1, s_1) and (t_2, s_2) be points on the graph of $s(t)$, that is,

$s_{1,2} = s(t_{1,2})$. Then the points (s_1, t_1) and (s_2, t_2) lie on the graph of the inverse of $s(t)$, where $t_{1,2} = t(s_{1,2})$. Consider the identity

$$\frac{t(s_1) - t(s_2)}{s_1 - s_2} = \frac{t_1 - t_2}{s_1 - s_2} = \frac{1}{\frac{s_1 - s_2}{t_1 - t_2}} = \frac{1}{\frac{s(t_1) - s(t_2)}{t_1 - t_2}}.$$

When $t_2 \to t_1$, the right side tends to $1/s'(t_1)$ because $s(t)$ is differentiable and, moreover, $s'(t) = \|\mathbf{r}'(t)\| > 0$ for $t > a$. Hence, the limit of the left side as $s_2 \to s_1$ exists too and, by the definition of the derivative, must be equal to $t'(s_1)$. This is known as the inverse function theorem for a real-valued function of one real argument.

Theorem 12.6. (Inverse Function Theorem).
Let $s(t)$ have a continuous derivative such that $s'(t) > 0$. Then there exists an inverse differentiable function $t = t(s)$ and $t'(s) = 1/s'(t)$, where $t = t(s)$ on the right side.

The condition $s'(t) > 0$ guarantees the existence of a one-to-one correspondence between the variables s and t and hence the existence of the inverse function (see Calculus I). Recall that the derivative can be written as the ratio of differentials $s'(t) = ds/dt$. The advantage of this representation is that the differentials can be manipulated as numbers. So the theorem can be stated in the compact form

$$\frac{ds(t)}{dt} = \frac{1}{\frac{dt(s)}{ds}}\,, \qquad s = s(t)\,.$$

Making use of this relation, one finds

$$\frac{d}{ds} = \frac{1}{s'(t)}\frac{d}{dt} = \frac{1}{\|\mathbf{r}'(t)\|}\frac{d}{dt}$$

and therefore

$$\kappa(t) = \frac{\|\hat{\mathbf{T}}'(t)\|}{\|\mathbf{r}'(t)\|}\,. \tag{12.4}$$

Note that the existence of the curvature requires that $\mathbf{r}(t)$ be twice differentiable because $\hat{\mathbf{T}}(t)$ is proportional to $\mathbf{r}'(t)$. Differentiation of the unit vector $\hat{\mathbf{T}}$ can sometimes be a tedious technical task. The following theorem provides a more convenient way to calculate the curvature.

Theorem 12.7. (Curvature of a Curve).
Let a curve be traced out by a twice-differentiable vector function $\mathbf{r}(t)$. Then the curvature is

$$\kappa(t) = \frac{\|\mathbf{r}'(t) \times \mathbf{r}''(t)\|}{\|\mathbf{r}'(t)\|^3}. \tag{12.5}$$

PROOF. Put $v(t) = \|\mathbf{r}'(t)\|$. With this notation,

$$\mathbf{r}'(t) = v(t)\hat{\mathbf{T}}(t)\,.$$

Differentiating both sides of this relation, one infers

$$\mathbf{r}''(t) = v'(t)\hat{\mathbf{T}}(t) + v(t)\hat{\mathbf{T}}'(t) = \frac{v'(t)}{v(t)}\,\mathbf{r}'(t) + v(t)\hat{\mathbf{T}}'(t)\,. \tag{12.6}$$

Since the cross product of two parallel vectors vanishes, it follows from (12.6) that

$$\mathbf{r}'(t) \times \mathbf{r}''(t) = v(t)\Big(\mathbf{r}'(t) \times \hat{\mathbf{T}}'(t)\Big)\,. \tag{12.7}$$

Therefore,

$$\|\mathbf{r}'(t) \times \mathbf{r}''(t)\| = v(t)\|\mathbf{r}'(t) \times \hat{\mathbf{T}}'(t)\| = \|\mathbf{r}'(t)\|^2\|\hat{\mathbf{T}}'(t)\|\sin\theta, \tag{12.8}$$

where θ is the angle between $\hat{\mathbf{T}}'(t)$ and the tangent vector $\mathbf{r}'(t)$. By construction, $\hat{\mathbf{T}}(t)$ is a unit vector, $\|\hat{\mathbf{T}}(t)\|^2 = \hat{\mathbf{T}}(t) \cdot \hat{\mathbf{T}}(t) = 1$. By taking the derivative of both sides of the latter relation, one obtains $\hat{\mathbf{T}}'(t)\cdot\hat{\mathbf{T}}(t) = 0$, which means that the derivative of a unit tangent vector is always perpendicular to the unit tangent vector. Since $\mathbf{v}$ is parallel to $\hat{\mathbf{T}}$, the vector $\hat{\mathbf{T}}'$ is perpendicular to $\mathbf{v}$ as well. Hence, $\theta = \pi/2$ and $\sin\theta = 1$. Substituting the latter relation and $\|\hat{\mathbf{T}}'(t)\| = \kappa(t)\|\mathbf{r}'(t)\|$ (see (12.4)) into (12.8), the expression (12.5) is derived. □

EXAMPLE 12.9. *Find the curvature of the curve* $\mathbf{r}(t) = \langle \ln t, t^2, 2t\rangle$ *at the point* $P_0(0, 1, 2)$.

SOLUTION: The point P_0 corresponds to $t = 1$ because $\mathbf{r}(1) = \langle 0, 1, 2\rangle$ coincides with the position vector of P_0. Hence, one has to calculate $\kappa(1)$:

$$\begin{aligned}
\mathbf{r}'(1) &= \langle t^{-1},\ 2t,\ 2\rangle\Big|_{t=1} = \langle 1, 2, 2\rangle \quad\Rightarrow\quad \|\mathbf{r}'(1)\| = 3\,,\\
\mathbf{r}''(1) &= \langle -t^{-2},\ 2,\ 0\rangle\Big|_{t=1} = \langle -1, 2, 0\rangle \quad\Rightarrow\quad \mathbf{r}'(1) \times \mathbf{r}''(1)\\
&= -2\langle 2, 1, -2\rangle\,,\\
\kappa(1) &= \frac{\|\mathbf{r}'(1) \times \mathbf{r}''(1)\|}{\|\mathbf{r}'(1)\|^3} = \frac{2\,\|\langle 2, 1, -2\rangle\|}{3^3} = \frac{6}{27} = \frac{2}{9}.
\end{aligned}$$

□

Equation (12.5) can be simplified in two particularly interesting cases. If a curve is planar (i.e., it lies in a plane), then, by choosing the coordinate system so that the xy plane coincides with the plane in which the curve lies, one has $\mathbf{r}(t) = \langle x(t), y(t), 0\rangle$. Since $\mathbf{r}'$ and

$\mathbf{r}''$ are in the xy plane, their cross product is parallel to the z axis: $\mathbf{r}' \times \mathbf{r}'' = \langle 0, 0, x'y'' - x''y' \rangle$. Then we have the following result.

Corollary 12.1. (Curvature of a Planar Curve).
For a planar curve $\mathbf{r}(t) = \langle x(t), y(t), 0 \rangle$, the curvature is given by

$$\kappa = \frac{|x'y'' - x''y'|}{[(x')^2 + (y')^2]^{3/2}}.$$

A further simplification occurs when the planar curve is a graph $y = f(x)$. The graph is traced out by the vector function $\mathbf{r}(t) = \langle t, f(t), 0 \rangle$. Then, in the above corollary, $x'(t) = 1$, $x''(t) = 0$, and $y''(t) = f''(t) = f''(x)$, which leads to the following result.

Corollary 12.2. (Curvature of a Graph).
The curvature of the graph $y = f(x)$ is given by

$$\kappa(x) = \frac{|f''(x)|}{[1 + (f'(x))^2]^{3/2}}.$$

83.1. Geometrical Significance of the Curvature. Let us calculate the curvature of a circle of radius R. Put $\mathbf{r}(t) = \langle R\cos t,\ R\sin t,\ 0 \rangle$. Then $\|\mathbf{r}'(t)\| = \|\langle -R\sin t, R\cos t, 0 \rangle\| = R$ and $|x'y''-x''y'| = R^2$. Therefore, the curvature is constant along the circle and equals a reciprocal of its radius, $\kappa = 1/R$. The fact that the curvature is independent of its position on the circle can be anticipated from the rotational symmetry of the circle (it bends uniformly). Naturally, if two circles of different radii pass through the same point, then the circle of smaller radius bends more. Note also that the curvature has the dimension of the inverse length. This motivates the following definition.

Definition 12.15. (Curvature Radius).
The reciprocal of the curvature of a curve is called the curvature radius *$\rho(t) = 1/\kappa(t)$.*

Let a planar curve have a curvature κ at a point P. Consider a circle of radius $\rho = 1/\kappa$ through the same point P. The curve and the circle have the same curvature at P; that is, in a sufficiently small neighborhood of P, the circle approximates well the curve as they are equally bent at P. So, if one says that the curvature of a curve at a point P is κ inverse meters, then the curve looks like a circle of radius $1/\kappa$ meters near P.

For a general spatial curve, not every circle of radius $\rho = 1/\kappa$ that passes through P would approximate well the curve near P. The best approximation is attained when the circle and the curve "bend" in the

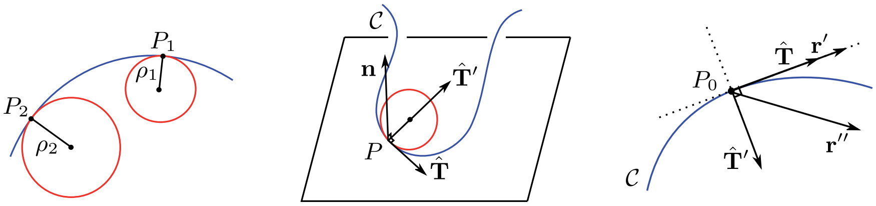

FIGURE 12.10. **Left**: Curvature radius. A smooth curve near a point P can be approximated by a portion of a circle of radius $\rho = 1/\kappa$. The curve bends in the same way as a circle of radius that is the reciprocal of the curvature. A large curvature at a point corresponds to a small curvature radius.
Middle: Osculating plane and osculating circle. The osculating plane at a point P contains the tangent vector $\hat{\mathbf{T}}$ and its derivative $\hat{\mathbf{T}}'$ at P and hence is perpendicular to $\mathbf{n} = \hat{\mathbf{T}} \times \hat{\mathbf{T}}'$. The osculating circle lies in the osculating plane, it has radius $\rho = 1/\kappa$, and its center is a distance ρ from P in the direction of $\hat{\mathbf{T}}'$. One says that the curve "bends" in the osculating plane.
Right: For a curve traced out by a vector function $\mathbf{r}(t)$, the derivatives $\mathbf{r}'$ and $\mathbf{r}''$ at any point P_0 lie in the osculating plane through P_0. So the normal to the osculating plane can also be computed as $\mathbf{n} = \mathbf{r}(t_0)' \times \mathbf{r}''(t_0)$, where $\mathbf{r}(t_0)$ is the position vector of P_0.

same plane. From the discussion at the beginning of this section, it might be concluded that a curve bends in the plane that contains the unit tangent vector $\hat{\mathbf{T}}$ and its derivative $\hat{\mathbf{T}}'$.

DEFINITION 12.16. (Osculating Plane and Circle).
The plane through a point P of a curve that is parallel to the unit tangent vector $\hat{\mathbf{T}}$ and its derivative $\hat{\mathbf{T}}' \neq \mathbf{0}$ at P is called the osculating plane *at P. The circle of radius $\rho = 1/\kappa$, where κ is the curvature at P, through P that lies in the osculating plane and whose center is in the direction of $\hat{\mathbf{T}}'$ from P is called the* osculating circle *at P.*

Recall that $\hat{\mathbf{T}}' \perp \hat{\mathbf{T}}$ (see the proof of Theorem 12.7). By the geometrical interpretation of the derivative, $\hat{\mathbf{T}}'$ should point in the direction in which the curve bends. Hence, the osculating circle must have the same $\hat{\mathbf{T}}'$ at a common point P in order to make the best approximation to the curve near P. Therefore, its center must be in the direction of $\hat{\mathbf{T}}'$ from P, not in the opposite one.

THEOREM 12.8. (Equation of the Osculating Plane). *Let a curve C be traced out by a twice-differentiable vector function $\mathbf{r}(t)$. Let P_0 be a point of C such that its position vector is $\mathbf{r}(t_0) = \langle x_0, y_0, z_0\rangle$ at which the vector $\mathbf{n} = \mathbf{r}'(t_0) \times \mathbf{r}''(t_0)$ does not vanish. An equation of the osculating plane through P_0 is*

$$n_1(x - x_0) + n_2(y - y_0) + n_3(z - z_0) = 0\,, \quad \mathbf{n} = \langle n_1, n_2, n_3\rangle.$$

PROOF. It follows from (12.6) that the second derivative $\mathbf{r}''(t_0)$ lies in the osculating plane because it is a linear combination of $\hat{\mathbf{T}}(t_0)$ and $\hat{\mathbf{T}}'(t_0)$. Hence, the osculating plane contains the first and second derivatives $\mathbf{r}'(t_0)$ and $\mathbf{r}''(t_0)$. Therefore, their cross product $\mathbf{n} = \mathbf{r}'(t_0) \times \mathbf{r}''(t_0)$ is perpendicular to the osculating plane, and the conclusion of the theorem follows. □

EXAMPLE 12.10. *For the curve $\mathbf{r}(t) = \langle t, t^2, t^3\rangle$, find the osculating plane through the point $(1, 1, 1)$.*

SOLUTION: The point in question corresponds to $t = 1$. Therefore, the normal of the osculating plane is $\mathbf{n} = \mathbf{r}'(1)\times\mathbf{r}''(1) = \langle 1, 2, 3\rangle\times\langle 0, 2, 6\rangle = \langle 6, -6, 2\rangle$. The osculating plane is $6(x - 1) - 6(y - 1) + 2(z - 1) = 0$ or $3x - 3y + z = 1$. □

83.2. Study Problems.

Problem 12.11. *Find the maximal curvature of the graph of the exponential, $y = e^x$, and the point(s) at which it occurs.*

SOLUTION: The curvature of the graph is given by $\kappa(x) = e^x/(1 + e^{2x})^{3/2}$. Critical points are determined by $\kappa'(x) = 0$ or

$$\kappa'(x) = \frac{e^x(1 + e^{2x})^{1/2}[2e^{2x} - 1]}{(1 + e^{2x})^3} = 0 \quad \Rightarrow \quad 2e^{2x} - 1 = 0$$
$$\Rightarrow \quad x = -\frac{\ln 2}{2}.$$

From the shape of the graph of the exponential, it is clear that $\kappa(x)$ attains its absolute maximum (maximal bending) and $\kappa_{\max} = \kappa(-\ln(2)/2) = 2/3^{3/2}$. □

Problem 12.12. (Equation of the Osculating Circle). *Find a vector function that traces out the osculating circle of a curve $\mathbf{r}(t)$ at a point $\mathbf{r}(t_0)$.*

SOLUTION: Put $\mathbf{r}_0 = \mathbf{r}(t_0)$ and $\hat{\mathbf{T}}_0 = \hat{\mathbf{T}}(t_0)$ (the unit tangent vector to the curve at the point with the position vector $\mathbf{r}_0$). Put $\hat{\mathbf{N}}_0 = \hat{\mathbf{T}}'(t_0)/\|\hat{\mathbf{T}}'(t_0)\|$; it is a unit vector in the direction of $\hat{\mathbf{T}}'(t_0)$. Let $\rho_0 = 1/\kappa(t_0)$ be the curvature radius at the point $\mathbf{r}_0$. The center of the osculating circle must lie ρ_0 units of length from the point $\mathbf{r}_0$ in the direction of $\hat{\mathbf{N}}_0$. Thus, its position vector is $\mathbf{R}_0 = \mathbf{r}_0 + \rho_0\hat{\mathbf{N}}_0$. Let $\mathbf{R}(t)$ be the position vector of a generic point of the osculating circle. Then the vector $\mathbf{R}(t) - \mathbf{R}_0$ lies in the osculating plane and hence must be a linear combination of $\hat{\mathbf{T}}_0$ and $\hat{\mathbf{N}}_0$, that is, $\mathbf{R}(t) - \mathbf{R}_0 = a(t)\hat{\mathbf{N}}_0 + b(t)\hat{\mathbf{T}}_0$. To find the functions $a(t)$ and $b(t)$, note that the vector $\mathbf{R}(t) - \mathbf{R}_0$ traces out a circle of radius ρ_0. In a coordinate system in which $\hat{\mathbf{N}}_0$ coincides with $\hat{\mathbf{e}}_x$ and $\hat{\mathbf{T}}_0$ with $\hat{\mathbf{e}}_y$ (such a coordinate system always exists because $\hat{\mathbf{N}}_0$ and $\hat{\mathbf{T}}_0$ are unit orthogonal vectors), the vector $-\rho_0\cos(t)\hat{\mathbf{e}}_x + \rho_0\sin(t)\hat{\mathbf{e}}_y$ traces out a circle of radius ρ_0 in the xy plane. Thus, one can always put $a(t) = -\rho_0\cos t$ and $b(t) = \rho_0\sin t$, and the vector function that traces out the osculating circle is

$$\mathbf{R}(t) = \mathbf{r}_0 + \rho_0\Big(1 - \cos t\Big)\,\hat{\mathbf{N}}_0 + \rho_0 \sin t\,\hat{\mathbf{T}}_0,$$

where $t \in [0, 2\pi]$. □

Problem 12.13. *Consider a helix* $\mathbf{r}(t) = \langle R\cos(\omega t), R\sin(\omega t), ht\rangle$, *where* ω *and* h *are numerical parameters. The arc length of one turn of the helix is a function of the parameter* ω, $L = L(\omega)$, *and the curvature at any fixed point of the helix is also a function of* ω, $\kappa = \kappa(\omega)$. *Use only geometrical arguments (no calculus) to find the limits of* $L(\omega)$ *and* $\kappa(\omega)$ *as* $\omega \to \infty$.

SOLUTION: The vector function $\mathbf{r}(t)$ traces out one turn of the helix when t ranges over the period of $\cos(\omega t)$ or $\sin(\omega t)$ (i.e., over the interval of length $2\pi/\omega$). Thus, the helix rises by $2\pi h/\omega = H(\omega)$ along the z axis per each turn. When $\omega \to \infty$, the height $H(\omega)$ tends to 0 so that each turn of the helix becomes closer and closer to a circle of radius R. Therefore, $L(\omega) \to 2\pi R$ (the circumference) and $\kappa(\omega) \to 1/R$ (the curvature of the circle) as $\omega \to \infty$.

A calculus approach requires a lot more work to establish this result:

$$\begin{aligned} L(\omega) &= \int_0^{2\pi/\omega} \|\mathbf{r}'(t)\|\, dt = \frac{2\pi}{\omega}\sqrt{(R\omega)^2 + h^2} \\ &= 2\pi\sqrt{R^2 + (h/\omega)^2} \to 2\pi R, \end{aligned}$$

$$\begin{aligned}\kappa(\omega) &= \frac{\|\mathbf{r}'(t)\times\mathbf{r}''(t)\|}{\|\mathbf{r}'(t)\|^3} = \frac{R\omega^2[(R\omega)^2+h^2]^{1/2}}{[(R\omega)^2+h^2]^{3/2}}\\ &= \frac{R}{R^2+(h/\omega)^2}\to\frac{1}{R}\end{aligned}$$

as $\omega\to\infty$. □

83.3. Exercises. **(1)** Find the curvature of $\mathbf{r}(t)=\langle t, t^2/2, t^3/3\rangle$ at the point of its intersection with the plane $z=2xy+1/3$.

(2) Find the maximal and minimal curvatures of the graph $y=\cos(ax)$ and the points at which they occur.

(3) Use a geometrical interpretation of the curvature to guess the point on the graphs $y=ax^2$ and $y=ax^4$ where the maximal curvature occurs. Then verify your guess by calculations.

(4) Let $\mathbf{r}(t)=\langle t^3, t^2, 0\rangle$. This curve has a cusp at $t=0$. Find the curvature for $t\neq 0$ and investigate its limit as $t\to 0$.

(5) Find an equation for the osculating and normal planes for the curve $\mathbf{r}(t)=\langle\,\ln(t)\,,\,2t\,,\,t^2\rangle$ at the point P_0 of its intersection with the plane $y-z=1$. A plane is normal to a curve at a point if the tangent to the curve at that point is normal to the plane.

(6) Prove that the trajectory of a particle is planar if its velocity $\mathbf{v}(t)$ remains perpendicular to a constant vector $\mathbf{n}$ and its acceleration is $\mathbf{a}(t)=\mathbf{n}\times\mathbf{v}(t)+\mu(t)\mathbf{v}(t)$, where $\mu(t)$ is a function of time. Find an equation of the plane in which the trajectory lies if the particle is known to pass a point $\mathbf{r}_0$.

84. Practical Applications

84.1. Tangential and Normal Accelerations. Let $\mathbf{r}(t)$ be the trajectory of a particle (t is time). Then $\mathbf{v}(t)=\mathbf{r}'(t)$ and $\mathbf{a}(t)=\mathbf{v}'(t)$ are the velocity and acceleration of the particle. The magnitude of the velocity vector is the speed, $v(t)=\|\mathbf{v}(t)\|$. If $\hat{\mathbf{T}}(t)$ is the unit tangent vector to the trajectory, then $\hat{\mathbf{T}}'(t)$ is perpendicular to it. The unit vector $\hat{\mathbf{N}}(t)=\hat{\mathbf{T}}'(t)/\|\hat{\mathbf{T}}'(t)\|$ is called a unit *normal* to the trajectory. In particular, the osculating plane at any point of the trajectory contains $\hat{\mathbf{T}}(t)$ and $\hat{\mathbf{N}}(t)$. It follows from (12.6) that the acceleration always lies in the osculating plane:

$$\mathbf{a}=v'\hat{\mathbf{T}}+v\hat{\mathbf{T}}'=v'\mathbf{T}+v\|\mathbf{T}'\|\hat{\mathbf{N}}\,.$$

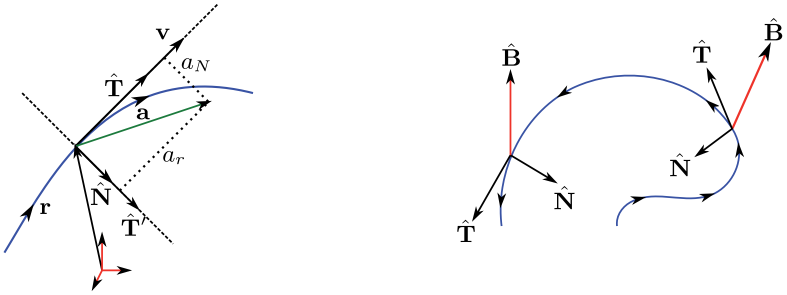

FIGURE 12.11. **Left**: Decomposition of the acceleration $\mathbf{a}$ of a particle into normal and tangential components. The tangential component a_T is the scalar projection of $\mathbf{a}$ onto the unit tangent vector $\hat{\mathbf{T}}$. The normal component is the scalar projection of $\mathbf{a}$ onto the unit normal vector $\hat{\mathbf{N}}$. The vectors $\mathbf{r}$ and $\mathbf{v}$ are the position and velocity vectors of the particle.
Right: The tangent, normal, and binormal vectors associated with a smooth curve. These vectors are mutually orthogonal and have unit length. The binormal is defined by $\hat{\mathbf{B}} = \hat{\mathbf{T}} \times \hat{\mathbf{N}}$. The shape of the curve is uniquely determined by the orientation of the triple of vectors $\hat{\mathbf{T}}$, $\hat{\mathbf{N}}$, and $\hat{\mathbf{B}}$ as functions of the arc length parameter up to general rigid rotations and translations of the curve as the whole.

Furthermore, substituting the relations $\kappa = \|\hat{\mathbf{T}}'\|/v$ and $\rho = 1/\kappa$ into the latter equation, one finds (see Figure 12.11, left panel) that

$$\begin{aligned} \mathbf{a} &= a_T\hat{\mathbf{T}} + a_N\hat{\mathbf{N}}, \\ a_T &= v' = \hat{\mathbf{T}} \cdot \mathbf{a} = \frac{\mathbf{v} \cdot \mathbf{a}}{v}, \\ a_N &= \kappa v^2 = \frac{v^2}{\rho} = \frac{\|\mathbf{v} \times \mathbf{a}\|}{v}\,. \end{aligned}$$

DEFINITION 12.17. (Tangential and Normal Accelerations). *Scalar projections a_T and a_N of the acceleration vector onto the unit tangent and normal vectors at any point of the trajectory of motion are called* tangential and normal accelerations*, respectively.*

The tangential acceleration a_T determines the rate of change of a particle's speed, while the normal acceleration appears only when the particle makes a "turn." In particular, a circular motion with a constant speed, $v = v_0$, has no tangential acceleration, $a_T = 0$, and the normal acceleration is constant, $a_N = v_0^2/R$, where R is the circle radius.

To gain an intuitive understanding of the tangential and normal accelerations, consider a car moving along a road. The speed of the car can be changed by pressing the gas or brake pedals. When one of these pedals is suddenly pressed, one can feel a force along the direction of motion of the car (the tangential direction). The car speedometer also shows that the speed changes, indicating that this force is due to the acceleration along the road (i.e., the tangential acceleration $a_T = v' \neq 0$). When the car moves along a straight road with a constant speed, its acceleration is 0. When the road takes a turn, the steering wheel must be turned in order to keep the car on the road, while the car maintains a constant speed. In this case, one can feel a force normal to the road. It is larger for sharper turns (larger curvature or smaller curvature radius) and also grows when the same turn is passed with a greater speed. This force is due to the normal acceleration, $a_N = v^2/\rho$, and is called a *centrifugal force*. When making a turn, the car does not slide off the road as long as the friction force between the tires and the road compensates for the centrifugal force. The maximal friction force depends on the road and tire conditions (e.g., a wet road and worn tires reduce substantially the maximal friction force). The centrifugal force is determined by the speed (the curvature of the road is fixed by the road shape). So, for a high enough speed, the centrifugal force can no longer be compensated for by the friction force and the car would skid off the road. For this reason, suggested speed limit signs are often placed at highway exits. If one drives a car on a highway exit with a speed twice as high as the suggested speed, *the risk of skidding off the road is quadrupled, not doubled*, because the normal acceleration $a_N = v^2/\rho$ quadruples when the speed v is doubled.

EXAMPLE 12.11. *A road has a parabolic shape, $y = x^2/(2R)$, where (x, y) are coordinates of points of the road and R is a constant (all measured in units of length, e.g., meters). A safety assessment requires that the normal acceleration on the road should not exceed a threshold value a_m (e.g., meters per second squared) to avoid skidding off the road. If a car moves with a constant speed v_0 along the road, find the portion of the road where the car might skid off the road.*

SOLUTION: The normal acceleration of the car as a function of *position* (not time!) is $a_N(x) = \kappa(x)v_0^2$. The curvature of the graph $y = x^2/(2R)$ is $\kappa(x) = (1/R)[1+(x/R)^2]^{-3/2}$. The maximal curvature and hence the maximal normal acceleration are attained at $x = 0$. So, if the speed is such that $a_N(0) = v_0^2/R < a_m$, no accident can happen. Otherwise, the inequality $a_N(x) \le a_m$ yields

$$\frac{v_0^2}{R}\frac{1}{[1+(x/R)^2]^{3/2}} \le a_m \quad \Rightarrow \quad |x| \le R\sqrt{\nu - 1}\,, \quad \nu = \Big(\frac{v_0^2}{Ra_m}\Big)^{2/3}.$$

Here the constant ν always exceeds 1 if $a_N(0) = v_0^2/R > a_m$. The car can skid off the road when moving on its portion corresponding to the interval $-R(\nu-1)^{1/2} \le x \le R(\nu-1)^{1/2}$. □

84.2. Frenet-Serret Formulas. The shape of a space curve as a point set is independent of a parameterization of the curve. A natural question arises: What parameters of the curve determine its shape? Suppose the curve is smooth enough so that the unit tangent vector $\hat{\mathbf{T}}(s)$ and its derivative $\hat{\mathbf{T}}'(s)$ can be defined as functions of the arc length s counted from an endpoint of the curve. Let $\hat{\mathbf{N}}(s)$ be the unit normal vector of the curve.

Definition 12.18. (Binormal Vector).
Let $\hat{\mathbf{T}}$ and $\hat{\mathbf{N}}$ be the unit tangent and normal vectors at a point of a curve. The unit vector $\hat{\mathbf{B}} = \hat{\mathbf{T}} \times \hat{\mathbf{N}}$ is called the binormal (unit) vector.

So, with every point of a smooth curve, one can associate a triple unit of mutually orthogonal vectors so that one of them is tangent to the curve while the other two span the plane normal to the tangent vector (normal to the curve). By a suitable rotation, the triple of vectors $\hat{\mathbf{T}}$, $\hat{\mathbf{N}}$, and $\hat{\mathbf{B}}$ can be oriented parallel to the axes of any given coordinate system, that is, parallel to $\hat{\mathbf{e}}_x$, $\hat{\mathbf{e}}_y$, and $\hat{\mathbf{e}}_z$, respectively (note that $\hat{\mathbf{e}}_x \times \hat{\mathbf{e}}_y = \hat{\mathbf{e}}_z$; this is why the binormal is defined as $\hat{\mathbf{T}} \times \hat{\mathbf{N}}$, not as $\hat{\mathbf{N}} \times \hat{\mathbf{T}} = -\hat{\mathbf{T}} \times \hat{\mathbf{N}}$). The orientation of the unit tangent, normal, and binormal vectors relative to some coordinate system depends on the point of the curve. The triple of these vectors can only rotate as the point slides along the curve (the vectors are mutually orthogonal and unit at any point). Therefore, the rates with respect to the arc length at which these vectors change must be characteristic for the shape of the curve (see Figure 12.11, right panel).

By the definition of the curvature, $\hat{\mathbf{T}}'(s) = \kappa(s)\hat{\mathbf{N}}(s)$. Next, consider the rate:

$$\hat{\mathbf{B}}' = (\hat{\mathbf{T}} \times \hat{\mathbf{N}})' = \hat{\mathbf{T}}' \times \hat{\mathbf{N}} + \hat{\mathbf{T}} \times \hat{\mathbf{N}}' = \hat{\mathbf{T}} \times \hat{\mathbf{N}}'$$

because $\hat{\mathbf{T}}'(s)$ is parallel to $\hat{\mathbf{N}}(s)$. It follows from this equation that $\hat{\mathbf{B}}'$ is perpendicular to $\hat{\mathbf{T}}$, and, since $\hat{\mathbf{B}}$ is a unit vector, its derivative must also be perpendicular to $\hat{\mathbf{B}}$. Thus, $\hat{\mathbf{B}}'$ must be parallel to $\hat{\mathbf{N}}$.

This conclusion establishes the existence of another scalar quantity that characterizes the curve shape.

Definition 12.19. (Torsion of a Curve).
Let $\hat{\mathbf{N}}(s)$ *and* $\hat{\mathbf{B}}(s)$ *be unit normal and binormal vectors of the curve as functions of the arc length* s*. Then*

$$\frac{d\hat{\mathbf{B}}(s)}{ds} = -\tau(s)\hat{\mathbf{N}}(s)$$

and the number $\tau(s)$ *is called the* torsion of the curve.

By definition, the torsion is measured in units of a reciprocal length, just like the curvature, because the unit vectors $\hat{\mathbf{T}}$, $\hat{\mathbf{N}}$, and $\hat{\mathbf{B}}$ are dimensionless.

At any point of a curve, the binormal $\hat{\mathbf{B}}$ is perpendicular to the osculating plane. So, if the curve is planar, then $\hat{\mathbf{B}}$ does not change along the curve, $\hat{\mathbf{B}}'(s) = \mathbf{0}$, because the osculating plane at any point coincides with the plane in which the curve lies. Thus, the torsion is a local numerical characteristic that determines how fast the curve deviates from the osculating plane while bending in it with some curvature radius.

It follows from the relation $\hat{\mathbf{N}} = \hat{\mathbf{B}} \times \hat{\mathbf{T}}$ (compare $\hat{\mathbf{e}}_y = \hat{\mathbf{e}}_z \times \hat{\mathbf{e}}_x$) that

$$\hat{\mathbf{N}}' = (\hat{\mathbf{B}} \times \hat{\mathbf{T}})' = \hat{\mathbf{B}}' \times \hat{\mathbf{T}} + \hat{\mathbf{B}} \times \hat{\mathbf{T}}' = -\tau\hat{\mathbf{N}} \times \hat{\mathbf{T}} + \kappa\hat{\mathbf{B}} \times \hat{\mathbf{N}} = \tau\hat{\mathbf{B}} - \kappa\hat{\mathbf{T}}\,,$$

where the definitions of the torsion and curvature have been used. The obtained rates of the unit vectors are known as the *Frenet-Serret formulas*:

$$\hat{\mathbf{T}}'(s) = \kappa(s)\hat{\mathbf{N}}(s)\,, \tag{12.9}$$

$$\hat{\mathbf{N}}'(s) = -\kappa(s)\hat{\mathbf{T}}(s) + \tau(s)\hat{\mathbf{B}}(s)\,, \tag{12.10}$$

$$\hat{\mathbf{B}}'(s) = -\tau(s)\hat{\mathbf{N}}(s)\,. \tag{12.11}$$

Suppose that the vectors $\hat{\mathbf{T}}(0)$, $\hat{\mathbf{N}}(0)$, and $\hat{\mathbf{B}}(0)$ are given at an initial point of the curve. Then $\hat{\mathbf{T}}(s)$, $\hat{\mathbf{N}}(s)$, and $\hat{\mathbf{B}}(s)$ are uniquely determined by solving the Frenet-Serret equations, provided the curvature and torsion are given as functions of the arc length. This establishes a fundamental theorem about the shape of a space curve.

Theorem 12.9. (Shape of a Smooth Curve in Space).
A curve in space is determined by its curvature and torsion as functions of the arc length up to rigid rotations and translations of the curve as a whole.

A curve with zero curvature and torsion is a straight line. Indeed, in this case, the tangent, normal, and binormal vectors remain constant along the curve, $\hat{\mathbf{T}}(s) = \hat{\mathbf{T}}(0)$, $\hat{\mathbf{N}}(s) = \hat{\mathbf{N}}(0)$, and $\hat{\mathbf{B}}(s) = \hat{\mathbf{B}}(0)$; that is, it has a constant unit tangent vector, which is the characteristic property of a straight line.

EXAMPLE 12.12. *Prove that a curve with a constant curvature $\kappa(s) = \kappa_0 \neq 0$ and zero torsion $\tau(s) = 0$ is a circle (or its portion) of radius $R = 1/\kappa_0$.*

SOLUTION: Let the coordinate system be set so that $\mathbf{T}(0) = \hat{\mathbf{e}}_x$, $\mathbf{N}(0) = \hat{\mathbf{e}}_y$, and $\hat{\mathbf{B}}(0) = \hat{\mathbf{e}}_z$. Since the torsion is 0, the binormal does not change along the curve, $\hat{\mathbf{B}}(s) = \hat{\mathbf{e}}_z$. Any unit vector $\hat{\mathbf{T}}$ in the xy plane can be written as $\hat{\mathbf{T}} = \langle \cos\varphi, \sin\varphi, 0\rangle$, where $\varphi = \varphi(s)$ such that $\varphi(0) = 0$. Then a unit vector $\hat{\mathbf{N}}$ perpendicular to $\hat{\mathbf{T}}$ such that $\hat{\mathbf{T}} \times \hat{\mathbf{N}} = \hat{\mathbf{B}} = \hat{\mathbf{e}}_z$ must have the form $\hat{\mathbf{N}} = \langle -\sin\varphi, \cos\varphi, 0\rangle$. Equation (12.9) gives $\hat{\mathbf{T}}' = \varphi'\hat{\mathbf{N}} = \kappa_0\hat{\mathbf{N}}$ and therefore $\varphi'(s) = \kappa_0$ or $\varphi(s) = \kappa_0 s$. Let $\mathbf{r}(s) = \langle x(s), y(s), 0\rangle$ be a natural parameterization of the curve. It follows from the definition of the arc length parameter s that $\mathbf{r}'(s) = \hat{\mathbf{T}}(s)$ (see the relation above (12.4)). Hence,

$$\begin{aligned} \mathbf{r}'(s) &= \langle \cos(\kappa_0 s), \sin(\kappa_0 s), 0\rangle \quad \Rightarrow \\ \mathbf{r}(s) &= \mathbf{r}_0 + \langle \kappa_0^{-1}\sin(\kappa_0 s), -\kappa_0^{-1}\cos(\kappa_0 s), 0\rangle, \end{aligned}$$

where $\mathbf{r}_0 = \langle x_0, y_0, z_0\rangle$. Thus, the curve lies in the plane $z = z_0$ and $(x(s) - x_0)^2 + (y(s) - y_0)^2 = R^2$, where $R = 1/\kappa_0$, for all values of s; that is, the curve is a circle (or its portion) of radius R. □

THEOREM 12.10. (Torsion of a Curve).
The torsion of a curve traced out by $\mathbf{r}(t)$ is given by

$$\tau(t) = \frac{(\mathbf{r}'(t) \times \mathbf{r}''(t)) \cdot \mathbf{r}'''(t)}{\|\mathbf{r}'(t) \times \mathbf{r}''(t)\|^2}.$$

PROOF. Put $\|\mathbf{r}'(t)\| = v(t)$ (if $s = s(t)$ is the arc length as a function of t, then $s' = v$). By (12.6) and the definition of the curvature,

$$\mathbf{r}'' = v'\hat{\mathbf{T}} + \kappa v^2\hat{\mathbf{N}}\,, \tag{12.12}$$

and by (12.7) and the definition of the binormal,

$$\mathbf{r}' \times \mathbf{r}'' = v\hat{\mathbf{T}} \times \mathbf{r}'' = \kappa v^3\hat{\mathbf{B}}. \tag{12.13}$$

Differentiation of both sides of (12.12) gives

$$\mathbf{r}''' = v''\hat{\mathbf{T}} + v'\hat{\mathbf{T}}' + (\kappa' v^2 + 2\kappa v v')\hat{\mathbf{N}} + \kappa v^2\hat{\mathbf{N}}'.$$

The derivatives $\hat{\mathbf{T}}'(t)$ and $\hat{\mathbf{N}}'(t)$ are found by making use of the differentiation rule $d/ds = (1/s'(t))(d/dt) = (1/v)(d/dt)$ in the Frenet-Serret equations (12.9) and (12.10):

$$\hat{\mathbf{T}}' = \kappa v \hat{\mathbf{N}}\,, \quad \hat{\mathbf{N}}' = -\kappa v \hat{\mathbf{T}} + \tau v \hat{\mathbf{B}}\,.$$

Therefore,

$$\mathbf{r}''' = (v'' - \kappa^2 v^3)\hat{\mathbf{T}} + (3\kappa v v' + \kappa' v^2)\hat{\mathbf{N}} + \kappa\tau v^3 \hat{\mathbf{B}}. \tag{12.14}$$

Since the tangent, normal, and binormal vectors are unit and orthogonal to each other, $(\mathbf{r}' \times \mathbf{r}'') \cdot \mathbf{r}''' = \kappa v^3 (\mathbf{r}' \times \mathbf{r}'') \cdot \hat{\mathbf{B}} = \kappa^2 v^6 \tau$. Therefore,

$$\tau = \frac{(\mathbf{r}' \times \mathbf{r}'') \cdot \mathbf{r}'''}{\kappa^2 v^6}$$

and the conclusion of the theorem follows from Theorem 12.7, $\kappa = \|\mathbf{r}' \times \mathbf{r}''\|/v^3$. □

Remark. Relation (12.13) is often more convenient for calculating the unit binormal vector rather than its definition. The unit tangent, normal, and binormal vectors at a particular point $\mathbf{r}(t_0)$ of the curve $\mathbf{r}(t)$ are

$$\hat{\mathbf{T}}(t_0) = \frac{\mathbf{r}'(t_0)}{\|\mathbf{r}'(t_0)\|}\,, \quad \hat{\mathbf{B}}(t_0) = \frac{\mathbf{r}'(t_0) \times \mathbf{r}''(t_0)}{\|\mathbf{r}'(t_0) \times \mathbf{r}''(t_0)\|}\,, \quad \hat{\mathbf{N}}(t_0) = \hat{\mathbf{B}}(t_0) \times \hat{\mathbf{T}}(t_0).$$

84.3. Study Problems.

Problem 12.14. *Find the position vector $\mathbf{r}(t)$ of a particle as a function of time t if the particle moves clockwise along a circular path of radius R in the xy plane through $\mathbf{r}(0) = \langle R, 0, 0\rangle$ with a constant speed v_0.*

Solution: For a circle of radius R in the xy plane through the point $(R, 0, 0)$, $\mathbf{r}(t) = \langle R\cos\varphi, R\sin\varphi, 0\rangle$, where $\varphi = \varphi(t)$ such that $\varphi(0) = 0$. Then the velocity is $\mathbf{v}(t) = \mathbf{r}'(t) = \varphi'\langle -R\sin\varphi, R\cos\varphi, 0\rangle$. Hence, the condition $\|\mathbf{v}(t)\| = v_0$ yields $R|\varphi'(t)| = v_0$ or $\varphi(t) = \pm(v_0/R)t$ and

$$\mathbf{r}(t) = \langle R\cos(\omega t), \pm R\sin(\omega t), 0\rangle\,,$$

where $\omega = v_0/R$ is the angular velocity. The second component must be taken with the minus sign because the particle revolves clockwise (the second component should become negative immediately after $t = 0$). □

Problem 12.15. *Let the particle position vector as a function of time t be $\mathbf{r}(t) = \langle \ln(t), t^2, 2t\rangle$, $t > 0$. Find the speed, tangential and normal accelerations, the unit tangent, normal, and binormal vectors, and the torsion of the trajectory at the point $P_0(0, 1, 2)$.*

Solution: By Example 12.9, the velocity and acceleration vectors at P_0 are $\mathbf{v} = \langle 1, 2, 2\rangle$ and $\mathbf{a} = \langle -1, 2, 0\rangle$. So the speed is $v = \|\mathbf{v}\| = 3$. The tangential acceleration is $a_T = \mathbf{v}\cdot\mathbf{a}/v = 1$. As $\mathbf{v}\times\mathbf{a} = 2\langle -2, -1, 2\rangle$, the normal acceleration is $a_N = \|\mathbf{v}\times\mathbf{a}\|/v = 6/3 = 2$. The unit tangent vector is $\hat{\mathbf{T}} = \mathbf{v}/v = (1/3)\langle 1, 2, 2\rangle$ and the unit binormal vector is $\hat{\mathbf{B}} = \mathbf{v} \times \mathbf{a}/\|\mathbf{v} \times \mathbf{a}\| = (1/3)\langle -2, -1, 2\rangle$ as the unit vector along $\mathbf{v} \times \mathbf{a}$. Therefore, the unit normal vector is $\hat{\mathbf{N}} = \hat{\mathbf{T}} \times \hat{\mathbf{B}} = (1/9)\mathbf{v} \times (\mathbf{v} \times \mathbf{a}) = (1/3)\langle -2, 2, -1\rangle$. To find the torsion at P_0, the third derivative at $t = 0$ has to be calculated, $\mathbf{r}'''(1) = \langle 2/t^2, 0, 0\rangle|_{t=1} = \langle 2, 0, 0\rangle = \mathbf{b}$. Therefore, $\tau(1) = (\mathbf{v} \times \mathbf{a}) \cdot \mathbf{b}/\|\mathbf{v} \times \mathbf{a}\|^2 = -8/36 = -2/9$. □

Problem 12.16. (Curves with Constant Curvature and Torsion). *Find the shape of a curve that has constant, nonzero curvature and torsion.*

Solution: Put $\kappa(s) = \kappa_0 \neq 0$ and $\tau(s) = \tau_0 \neq 0$. It follows from (12.9) and (12.11) that the vector $\mathbf{w} = \tau\hat{\mathbf{T}} + \kappa\hat{\mathbf{B}}$ does not change along the curve, $\mathbf{w}'(s) = 0$. Indeed, because $\kappa'(s) = \tau'(s) = 0$, one has $\mathbf{w}' = \tau\hat{\mathbf{T}}' + \kappa\hat{\mathbf{B}}' = (\tau\kappa - \tau\kappa)\hat{\mathbf{N}} = \mathbf{0}$. It is therefore convenient to introduce new unit vectors orthogonal to $\hat{\mathbf{N}}$:

$$\hat{\mathbf{u}} = \cos(\alpha)\hat{\mathbf{T}} - \sin(\alpha)\hat{\mathbf{B}}\,, \quad \hat{\mathbf{w}} = \sin(\alpha)\hat{\mathbf{T}} + \cos(\alpha)\hat{\mathbf{B}},$$

where $\cos\alpha = \kappa_0/\omega$, $\sin\alpha = \tau_0/\omega$, and $\omega = (\kappa_0^2 + \tau_0^2)^{1/2}$. By construction, the unit vectors $\hat{\mathbf{u}}$, $\hat{\mathbf{w}}$, and $\hat{\mathbf{N}}$ are mutually orthogonal unit vectors, which is easy to verify by calculating the corresponding dot products, $\hat{\mathbf{u}} \cdot \hat{\mathbf{u}} = \hat{\mathbf{w}} \cdot \hat{\mathbf{w}} = 1$ and $\hat{\mathbf{u}} \cdot \hat{\mathbf{w}} = 0$. Also,

$$\hat{\mathbf{u}} \times \hat{\mathbf{w}} = \hat{\mathbf{N}}\,.$$

It has been established that the vector $\mathbf{w}(s)$ is constant along the curve and so is its unit vector $\hat{\mathbf{w}}(s) = \mathbf{w}(s)/\omega$. Therefore, one can always choose the coordinate system so that

$$\hat{\mathbf{w}}(s) = \hat{\mathbf{w}}(0) = \hat{\mathbf{e}}_z.$$

By differentiating the vector $\hat{\mathbf{u}}$ and using the Frenet-Serret equations,

$$\hat{\mathbf{u}}' = \omega\mathbf{N}\,.$$

Put $\hat{\mathbf{u}} = \langle \cos\varphi, \sin\varphi, 0\rangle$, where $\varphi = \varphi(s)$. Then the unit normal vector is $\hat{\mathbf{N}} = \hat{\mathbf{u}} \times \hat{\mathbf{w}} = \langle -\sin\varphi, \cos\varphi, 0\rangle$ and $\hat{\mathbf{u}}' = \varphi'\hat{\mathbf{N}}$. Hence, $\varphi'(s) = \omega$

or $\varphi(s) = \omega s$ (the integration constant is set to 0; see Example 12.12). Expressing the vector $\hat{\mathbf{T}}$ via $\hat{\mathbf{u}}$ and $\hat{\mathbf{w}}$,

$$\hat{\mathbf{T}} = \cos(\alpha)\hat{\mathbf{u}} + \sin(\alpha)\hat{\mathbf{w}}\,,$$

one infers (compare Example 12.12)

$$\mathbf{r}'(s) = \hat{\mathbf{T}}(s) = \Big\langle \frac{\kappa_0}{\omega}\cos(\omega s),\ \frac{\kappa_0}{\omega}\sin(\omega s),\ \frac{\tau_0}{\omega}\Big\rangle\,,$$

where $\mathbf{r}(s)$ is a natural parameterization of the curve. The integration of this equation gives

$$\mathbf{r}(s) = \mathbf{r}_0 + \langle R\sin(\omega s),\ -R\cos(\omega s),\ hs\rangle\,,\quad R = \frac{\kappa_0}{\omega^2}\,,\quad h = \frac{\tau_0}{\omega},$$

where $\omega = (\kappa_0^2 + \tau_0^2)^{1/2}$. This is a helix of radius R whose axis goes through the point $\mathbf{r}_0$ parallel to the z axis; the helix climbs along its axis by $2\pi h/\omega$ per each turn. □

Remark. A curve in a neighborhood of its particular point P_0 can be well approximated by a helix through that point whose curvature and torsion coincide with the curvature and torsion of the curve at P_0. Such an approximation is better than the approximation by an osculating circle because the latter does not take into account the rate at which the curve deviates from the osculating plane (which is determined by the torsion).

Problem 12.17. (Motion in a Constant Magnetic Field, Revisited). *The force acting on a charged particle moving in the magnetic field* $\mathbf{B}$ *is given by* $\mathbf{F} = (e/c)\mathbf{v}\times\mathbf{B}$*, where* e *is the electric charge of the particle,* c *is the speed of light, and* $\mathbf{v}$ *is its velocity. Show that the trajectory of the particle in a constant magnetic field is a helix whose axis is parallel to the magnetic field.*

Solution: In contrast to Study Problem 12.10, here the shape of the trajectory is to be obtained directly from Newton's second law with arbitrary initial conditions. Choose the coordinate system so that the magnetic field is parallel to the z axis, $\mathbf{B} = B\hat{\mathbf{e}}_z$, where B is the magnitude of the magnetic field. Newton's law of motion, $m\mathbf{a} = \mathbf{F}$, where m is the mass of the particle, determines the acceleration, $\mathbf{a} = \mu\mathbf{v}\times\mathbf{B} = \mu B\mathbf{v}\times\hat{\mathbf{e}}_z$, where $\mu = e/(mc)$. First, note that $a_3 = \hat{\mathbf{e}}_z\cdot\mathbf{a} = 0$. Hence, $v_3 = v_\parallel = \text{const}$. Second, the acceleration, and velocity remain orthogonal during the motion, and therefore the tangential acceleration vanishes, $a_T = \mathbf{v}\cdot\mathbf{a} = 0$. Hence, the speed of the particle is a constant of motion, $v = v_0$ (because $v' = a_T = 0$). Put $\mathbf{v} = \mathbf{v}_\perp + v_\parallel\hat{\mathbf{e}}_z$, where $\mathbf{v}_\perp$ is the projection of $\mathbf{v}$ onto the xy plane. Since $\|\mathbf{v}\| = v_0$, the magnitude of

$\mathbf{v}_\perp$ is also constant, $\|\mathbf{v}_\perp\| = v_\perp = (v_0^2 - v_\|^2)^{1/2}$. The velocity vector may therefore be written in the form $\mathbf{v} = \langle v_\perp \cos\varphi, v_\perp \sin\varphi, v_{z0}\rangle$, where $\varphi = \varphi(t)$. Taking the derivative $\mathbf{a} = \mathbf{v}'$ and substituting it into Newton's equation, one finds $\varphi'(t) = \mu B$ or $\varphi(t) = \mu B t + \varphi_0$. Integration of the equation $\mathbf{r}' = \mathbf{v}$ yields the trajectory of motion:

$$\mathbf{r}(t) = \mathbf{r}_0 + \langle R\sin(\omega t + \varphi_0), -R\cos(\omega t + \varphi_0), v_\| t\rangle,$$

where $\omega = eB/(mc)$ is the so-called cyclotron frequency and $R = v_\perp/\omega$. This equation describes a helix of radius R whose axis goes through $\mathbf{r}_0$ parallel to the z axis. So a charged particle moves along a helix that winds about force lines of the magnetic field. The particle revolves in the plane perpendicular to the magnetic field with frequency $\omega = eB/(mc)$. In each turn, the particle moves along the magnetic field a distance $h = 2\pi v_\|/\omega$. In particular, if the initial velocity is perpendicular to the magnetic field (i.e., $v_\| = 0$), then the trajectory is a circle of radius R. □

Problem 12.18. *Suppose that the force acting on a particle of mass m is proportional to the position vector of the particle (such forces are called* central*). Prove that the angular momentum of the particle, $\mathbf{L} = m\mathbf{r}\times\mathbf{v}$, is a constant of motion (i.e., $d\mathbf{L}/dt = 0$).*

Solution: Since a central force $\mathbf{F}$ is parallel to the position vector $\mathbf{r}$, their cross product vanishes, $\mathbf{r}\times\mathbf{F} = \mathbf{0}$. By Newton's second law, $m\mathbf{a} = \mathbf{F}$ and hence $m\mathbf{r}\times\mathbf{a} = \mathbf{0}$. Therefore,

$$\frac{d\mathbf{L}}{dt} = m(\mathbf{r}\times\mathbf{v})' = m(\mathbf{r}'\times\mathbf{v} + \mathbf{r}\times\mathbf{v}') = m\mathbf{r}\times\mathbf{a} = \mathbf{0},$$

where $\mathbf{r}' = \mathbf{v}$, $\mathbf{v} = \mathbf{a}$, and $\mathbf{v}\times\mathbf{v} = \mathbf{0}$ have been used. □

Problem 12.19. (Kepler's Laws of Planetary Motion). *Newton's law of gravity states that two masses m and M at a distance r are attracted by a force of magnitude GmM/r^2, where G is the universal constant (called* Newton's constant*). Prove Kepler's laws of planetary motion:*
1. A planet revolves around the Sun in an elliptical orbit with the Sun at one focus.
2. The line joining the Sun to a planet sweeps out equal areas in equal times.
3. The square of the period of revolution of a planet is proportional to the cube of the length of the major axis of its orbit.

Solution: Let the Sun be at the origin of a coordinate system and let $\mathbf{r}$ be the position vector of a planet. Let $\hat{\mathbf{r}} = \mathbf{r}/r$ be the unit vector parallel to $\mathbf{r}$. Then the gravitational force is

$$\mathbf{F} = -\frac{GMm}{r^2}\,\hat{\mathbf{r}} = -\frac{GMm}{r^3}\,\mathbf{r},$$

where M is the mass of the Sun and m is the mass of a planet. The minus sign is necessary because an attractive force must be opposite to the position vector. By Newton's second law, the trajectory of a planet satisfies the equation $m\mathbf{a} = \mathbf{F}$ and hence

$$\mathbf{a} = -\frac{GM}{r^3}\mathbf{r}.$$

The gravities force is a central force, and, by Study Problem 12.18, the vector $\mathbf{r} \times \mathbf{v} = \mathbf{l}$ is a constant of motion. One has $\mathbf{v} = \mathbf{r}' = (r\hat{\mathbf{r}})' = r'\hat{\mathbf{r}} + r\hat{\mathbf{r}}'$. Using this identity, the constant of motion can also be written as

$$\mathbf{l} = \mathbf{r} \times \mathbf{v} = r\hat{\mathbf{r}} \times \mathbf{v} = r(r'\hat{\mathbf{r}} \times \hat{\mathbf{r}} + r\hat{\mathbf{r}} \times \hat{\mathbf{r}}') = r^2(\hat{\mathbf{r}} \times \hat{\mathbf{r}}').$$

Using the rule for the double cross product (see Study Problem 11.17), one infers that

$$\mathbf{a} \times \mathbf{l} = -\frac{GM}{r^2}\hat{\mathbf{r}} \times \mathbf{l} = -GM\hat{\mathbf{r}} \times (\hat{\mathbf{r}} \times \hat{\mathbf{r}}') = GM\hat{\mathbf{r}}',$$

where $\hat{\mathbf{r}} \cdot \hat{\mathbf{r}} = 1$ has been used. On the other hand,

$$(\mathbf{v} \times \mathbf{l})' = \mathbf{v}' \times \mathbf{l} + \mathbf{v} \times \mathbf{l}' = \mathbf{a} \times \mathbf{l}$$

because $\mathbf{l}' = \mathbf{0}$. It follows from these two equations that

$$(\mathbf{v} \times \mathbf{l})' = GM\hat{\mathbf{r}}' \quad \Longrightarrow \quad \mathbf{v} \times \mathbf{l} = GM\hat{\mathbf{r}} + \mathbf{c},$$

where $\mathbf{c}$ is a constant vector. The motion is characterized by two constant vectors $\mathbf{l}$ and $\mathbf{c}$. It occurs in the plane through the origin that is perpendicular to the constant vector $\mathbf{l}$ because $\mathbf{l} = \mathbf{r} \times \mathbf{v}$ must be orthogonal to $\mathbf{r}$. It is therefore convenient to choose the coordinate system so that $\mathbf{l}$ is parallel to the z axis and $\mathbf{c}$ to the x axis as shown in Figure 12.12 (left panel).

The vector $\mathbf{r}$ lies in the xy plane. Let θ be the polar angle of $\mathbf{r}$ (i.e., $\mathbf{r} \cdot \mathbf{c} = rc\cos\theta$, where $c = \|\mathbf{c}\|$ is the length of $\mathbf{c}$). Then

$$\mathbf{r} \cdot (\mathbf{v} \times \mathbf{l}) = \mathbf{r} \cdot (GM\hat{\mathbf{r}} + \mathbf{c}) = GMr + rc\cos\theta.$$

On the other hand, using a cyclic permutation in the triple product,

$$\mathbf{r} \cdot (\mathbf{v} \times \mathbf{l}) = \mathbf{l} \cdot (\mathbf{r} \times \mathbf{v}) = \mathbf{l} \cdot \mathbf{l} = l^2,$$

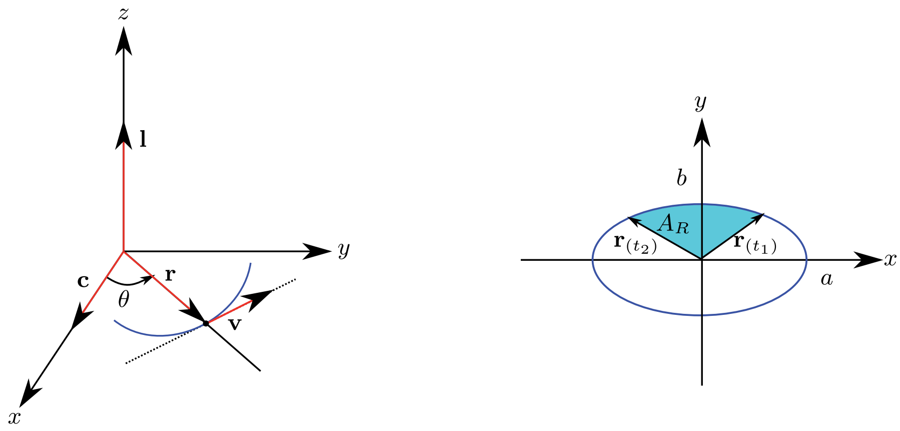

FIGURE 12.12. **Left**: The setup of the coordinate system for the derivation of Kepler's first law.
Right: An illustration to the derivation of Kepler's second law.

where $l = \|\mathbf{l}\|$ is the length of $\mathbf{l}$. The comparison of the last two equations yields the equation for the trajectory:

$$l^2 = r(GM + b\cos\theta) \quad\Longrightarrow\quad r = \frac{ed}{1 + e\cos\theta},$$

where $d = l^2/c$ and $e = c/(GM)$. This is the polar equation of a conic section with focus at the origin and eccentricity e (see Calculus II). Thus, *all possible trajectories of any massive body in a solar system are conic sections*! This is a quite remarkable result. Parabolas and hyperbolas do not correspond to a periodic motion. So a planet must follow an elliptic trajectory with the Sun at one focus. All objects coming to the solar system from outer space (i.e., those that are not confined by the gravitational pull of the Sun) should follow either parabolic or hyperbolic trajectories.

To prove Kepler's second law, put $\hat{\mathbf{r}} = \langle \cos\theta, \sin\theta, 0\rangle$ and hence $\hat{\mathbf{r}}' = \langle -\theta'\sin\theta, \theta'\cos\theta, 0\rangle$. Therefore,

$$\mathbf{l} = r^2(\hat{\mathbf{r}} \times \hat{\mathbf{r}}') = \langle 0, 0, r^2\theta'\rangle \quad\Longrightarrow\quad l = r^2\theta'\,.$$

The area of a sector with angle $d\theta$ swept by $\mathbf{r}$ is $dA = \frac{1}{2}r^2\,d\theta$ (see Calculus II; the area bounded by a polar graph $r = r(\theta)$). Hence,

$$\frac{dA}{dt} = \frac{1}{2}r^2\frac{d\theta}{dt} = \frac{l}{2}.$$

For any moments of time t_1 and t_2, the area of the sector between $\mathbf{r}(t_1)$ and $\mathbf{r}(t_2)$ is

$$A_{12} = \int_{t_1}^{t_2} \frac{dA}{dt}\,dt = \int_{t_1}^{t_2} \frac{l}{2}\,dt = \frac{l}{2}(t_2 - t_1).$$

Thus, the position vector $\mathbf{r}$ sweeps out equal areas in equal times (see Figure 12.12, right panel).

Kepler's third law follows from the last equation. Indeed, the entire area of the ellipse A is swept when $t_2 - t_1 = T$ is the period of the motion. If the major and minor axes of the ellipse are $2a$ and $2b$, respectively, $a > b$, then $A = \pi ab = lT/2$ and $T = 2\pi ab/l$. Now recall that $ed = b^2/a$ for an elliptic conic section (see Calculus II) or $b^2 = eda = l^2a/(GM)$. Hence,

$$T^2 = \frac{4\pi^2a^2b^2}{l^2} = \frac{4\pi^2}{GM}\,a^3.$$

Note that the proportionality constant $4\pi^2/(GM)$ is independent of the mass of a planet; therefore, Kepler's laws are *universal* for all massive objects trapped by the Sun (planets, asteroids, and comets). □

84.4. Exercises. **(1)** Find the normal and tangential accelerations of a particle with the position vector $\mathbf{r}(t) = \langle t^2+1, t^3, t^2-1\rangle$ when the particle is at the least distance from the origin.

(2) Find the tangential and normal accelerations of a particle with the position vector $\mathbf{r}(t) = \langle R\sin(\omega t+\varphi_0), -R\cos(\omega t+\varphi_0), v_0t\rangle$, where R, ω, φ_0, and v_0 are constants (see Study Problem 12.17).

(3) The shape of a winding road can be approximated by the graph $y = L\cos(x/L)$, where the coordinates are in miles and $L = 1$ mile. The condition of the road is such that if the normal acceleration of a car on it exceeds $10\,\mathrm{m/s}^2$, the car may skid off the road. Recommend a speed limit for this portion of the road.

(4) Suppose a particle moves so that its tangential acceleration is constant, while the normal acceleration remains 0. What is the trajectory of the particle?

(5) Suppose a particle moves so that its tangent acceleration remains 0, while the normal acceleration is constant. What is the trajectory of the particle?

Hint: Prove first that the acceleration of the particle has the form $\mathbf{a} = \mathbf{b}\times\mathbf{v}$, where $\mathbf{v}$ is the velocity and $\mathbf{b}$ is a vector that can depend on time. Use this fact to prove that the torsion of the trajectory is constant. Then see Study Problem 12.16.

CHAPTER 13

Differentiation of Multivariable Functions

85. Functions of Several Variables

The concept of a function of several variables can be qualitatively understood from simple examples in everyday life. The temperature in a room may vary from point to point. A point in space can be defined by an ordered triple of numbers that are coordinates of the point in some coordinate system, say, (x, y, z). Measurements of the temperature at every point from a set D in space assign a real number T (the temperature) to every point of D. The dependence of T on coordinates of the point is indicated by writing $T = T(x, y, z)$. Similarly, the concentration of a chemical can depend on a point in space. In addition, if the chemical reacts with other chemicals, its concentration at a point may also change with time. In this case, the concentration C depends on four variables—three spatial coordinates and the time t—$C = C(x, y, z, t)$. In general, if the value of a quantity f depends on values of several other quantities, say, x_1, x_2,..., x_n, this dependence is indicated by writing $f = f(x_1, x_2, ..., x_n)$. In other words, $f = f(x_1, x_2, ..., x_n)$ indicates a rule that assigns a number f to each ordered n-tuple of real numbers $(x_1, x_2, ..., x_n)$. Each number in the n-tuple may be of a different nature. In the above example, the concentration depends on ordered quadruples (x, y, z, t), where x, y, and z are the coordinates of a point in space and t is time.

Definition 13.1. (Real-Valued Function of Several Variables). *Let D be a set of ordered n-tuples of real numbers $(x_1, x_2, ..., x_n)$. A function f of n variables is a rule that assigns to each n-tuple in the set D a unique real number denoted by $f(x_1, x_2, ..., x_n)$. The set D is the domain of f, and its range is the set of values that f takes on it, that is, $\{f(x_1, x_2, ..., x_n) \,|\, (x_1, x_2, ..., x_n) \in D\}$.*

The rule may be defined by different means. If D is a finite set, a function f can be defined by a table $(P_i, f(P_i))$, where $P_i \in D$, $i = 1, 2, ..., N$, are elements (ordered n-tuples) of D, and $f(P_i)$ is the value of f at P_i. A function f can be defined geometrically. For example, the height of a mountain relative to sea level is a function of its position on the globe. So the height is a function of two variables, the longitude and latitude. A function can be defined by an algebraic rule that prescribes algebraic operations to be carried out with real numbers in any n-tuple to obtain the value of the function. For example, $f(x, y, z) = x^2-y+z^3$. The value of this function at $(1, 2, 3)$ is $f(1, 2, 3) = 1^2 - 2 + 3^3 = 28$. Unless specified otherwise, the domain D of a function defined by an algebraic rule is the set of n-tuples for which the rule makes sense.

Example 13.1. *Find the domain and the range of the function of two variables* $f(x, y) = \ln(1 - x^2 - y^2)$.

Solution: The logarithm is defined for any strictly positive number. Therefore, the doublets (x, y) must be such that $1 - x^2 - y^2 > 0$ or $x^2+y^2 < 1$. Hence, $D = \{(x, y) \,|\, x^2+y^2 < 1\}$. Since any doublet (x, y) can be uniquely associated with a point on a plane, the set D can be given a geometrical description as a disk of radius 1 whose boundary, the circle $x^2 + y^2 = 1$, is not included in D. For any point in the interior of the disk, the argument of the logarithm lies in the interval $0 \leq 1 - x^2 - y^2 < 1$. So the range of f is the set of values of the logarithm in the interval $(0, 1]$, which is $-\infty < f \leq 0$. □

Example 13.2. *Find the domain and the range of the function of three variables* $f(x, y, z) = x^2\sqrt{z - x^2 - y^2}$.

Solution: The square root is defined only for nonnegative numbers. Therefore, ordered triples (x, y, z) must be such that $z - x^2 - y^2 \geq 0$, that is, $D = \{(x, y, z) \,|\, z \geq x^2+y^2\}$. This set can be given a geometrical description as a point set in space because any triple can be associated with a unique point in space. The equation $z = x^2 + y^2$ describes a circular paraboloid. So the domain is the spatial (solid) region containing points that lie on or above the paraboloid. The function is nonnegative. By fixing x and y and increasing z, one can see that the value of f can be any positive number. So the range is $0 \leq f(x, y, z) < \infty$. □

85.1. The Graph of a Function of Two Variables. The graph of a function of one variable $f(x)$ is the set of points of a plane $\{(x, y) \,|\, y = f(x)\}$. The domain D is a set of points on the x axis. The graph is obtained by moving a point of the domain parallel to the y axis by an amount determined by the value of the function $y = f(x)$. The graph provides a useful picture of the behavior of the function. The idea can be extended to functions of two variables.

Definition 13.2. (Graph of a Function of Two Variables).
The graph of a function $f(x, y)$ with domain D is the point set in space

$$\{(x, y, z) \,|\, z = f(x, y),\ (x, y) \in D\}.$$

The domain D is a set of points in the xy plane. The graph is then obtained by moving each point of D parallel to the z axis by an amount equal to the corresponding value of the function $z = f(x, y)$. If D is a portion of the plane, then the graph of f is generally a surface. One can think of the graph as "mountains" of height $f(x, y)$ on the xy plane.

Example 13.3. *Sketch the graph of the function* $f(x, y) = \sqrt{1 - (x/2)^2 - (y/3)^2}$.

Solution: The domain is the portion of the xy plane $(x/2)^2+(y/3)^2 \leq 1$; that is, it is bounded by the ellipse with semiaxes 2 and 3. The graph is the surface defined by the equation $z = \sqrt{1 - (x/2)^2 - (y/3)^2}$. By squaring both sides of this equation, one finds $(x/2)^2 + (y/3)^2 + z^2 = 1$, which defines an ellipsoid. The graph is its upper portion with $z \geq 0$. □

The concept of the graph is obviously hard to extend to functions of more than two variables. The graph of a function of three variables would be a three-dimensional surface in four-dimensional space. So the qualitative behavior of a function of three variables should be studied by different graphical means.

85.2. Level Curves. When visualizing the shape of quadric surfaces, the method of cross sections by coordinate planes has been helpful. It can also be applied to visualize the shape of the graph $z = f(x, y)$. In particular, consider the cross sections of the graph with horizontal planes $z = k$. The curve of intersection is defined by the equation $f(x, y) = k$. Continuing the analogy that $f(x, y)$ defines the height of a mountain, a hiker traveling along the path $f(x, y) = k$ does not have to climb or descend as the height along the path remains constant.

Definition 13.3. (Level Curves).
The level curves of a function f of two variables are the curves along which the function remains constant; that is, they are determined by the equation $f(x, y) = k$, where k is a number from the range of f.

Definition 13.4. (Contour Map).
A collection of level curves is called a contour map *of the function f.*

The contour map of the function in Example 13.3 consists of ellipses. Indeed, the range is the interval $[0, 1]$. For any $k \in [0, 1]$, a level curve is an ellipse, $1 - (x/2)^2 + (y/3)^2 = k^2$ or $(x/a)^2 + (y/b)^2 = 1$, where $a = 2\sqrt{1-k^2}$ and $b = 3\sqrt{1-k^2}$.

A contour map is a useful tool for studying the qualitative behavior of a function. Consider the contour map that consists of level curves C_i, $i = 1, 2, ...$, $f(x, y) = k_i$, where $k_{i+1} - k_i = \Delta k$ is fixed. The values of the function along the neighboring curves C_i and C_{i+1} differ by Δk. So, in the region where the level curves are dense (close to one another), the function $f(x, y)$ changes rapidly. Indeed, let P be a point of C_i and let Δs be the distance from P to C_{i+1} along the normal to C_i. Then the slope of the graph of f or the rate of change of f at P in that direction is $\Delta k/\Delta s$. Thus, the closer the curves C_i are to one another, the faster the function changes. Such contour maps are used in topography to indicate the steepness of mountains on maps.

85.3. Level Surfaces. In contrast to the graph of a function, the method of level curves does not require a higher-dimensional space to study the behavior of a function of two variables. So the concept can be extended to functions of three variables.

Definition 13.5. (Level Surface).
The level surfaces of a function f of three variables are the surfaces along which the function remains constant; that is, they are determined by the equation $f(x, y, z) = k$, where k is a number from the range of f.

The shape of the level surfaces may be studied, for example, by the method of cross sections with coordinate planes. A collection of level surfaces S_i, $f(x, y, z) = k_i$, $k_{i+1} - k_i = \Delta k$, $i = 1, 2, ...$, can be depicted in the domain of f. The closer the level surfaces S_i are to one another, the faster the function changes.

Example 13.4. *Sketch and/or describe the level surfaces of the function*
$f(x, y, z) = z/(1 + x^2 + y^2)$.

Solution: The domain is the entire space, and the range contains all real numbers. The equation $f(x, y, z) = k$ can be written in the form $z - k = k(x^2 + y^2)$, which defines a circular paraboloid whose symmetry axis is the z axis and whose vertex is at $(0, 0, k)$. For larger k, the paraboloid rises faster. For $k = 0$, the level surface is the xy plane. For $k > 0$, the level surfaces are paraboloids above the xy plane, (i.e., they are concave downward). For $k < 0$, the paraboloids are below the xy plane (i.e., they are concave upward). □

85.4. Euclidean Spaces. When the number of variables is greater than 3, the geometrical visualization of the domain is not so simple. The goal is achieved with the help of the concept of a higher-dimensional Euclidean space. The plane and space are particular cases of two- and three-dimensional Euclidean spaces.

With every ordered pair of numbers (x, y), one can associate a point in a plane and its position vector relative to a fixed point $(0, 0)$ (the origin), $\mathbf{r} = \langle x, y\rangle$. With every ordered triple of numbers (x, y, z), one can associate a point in space and its position vector (again relative to the origin $(0, 0, 0)$), $\mathbf{r} = \langle x, y, z\rangle$. So the plane can be viewed as the set of all two-component vectors; similarly, space is the set of all three-component vectors. From this point of view, the plane and space have characteristic common features. First, their elements are vectors. Second, they are closed relative to addition of vectors and multiplication of vectors by a real number; that is, if $\mathbf{a}$ and $\mathbf{b}$ are elements of space or a plane and c is a real number, then $\mathbf{a} + \mathbf{b}$ and $c\mathbf{a}$ are also elements of space (ordered triples of numbers) or a plane (ordered pairs of numbers). Third, the norm or length of a vector $\|\mathbf{r}\|$ vanishes if and only if the vector has zero components. Consequently, two elements of space or a plane coincide if and only if the norm of their difference vanishes, that is, $\mathbf{a} = \mathbf{b} \Leftrightarrow \|\mathbf{a} - \mathbf{b}\| = 0$. From this point of view, there is no difference between a vector $\langle a_1, a_2, a_3\rangle$ and an ordered triple (a_1, a_2, a_3) as they represent the very same point in space; that is, there is no confusion in writing $\mathbf{a} = (a_1, a_2, a_3)$. Finally, the dot product $\mathbf{a}\cdot\mathbf{b}$ of two elements is defined in the same way for two- or three-component vectors (plane or space) so that $\|\mathbf{a}\|^2 = \mathbf{a}\cdot\mathbf{a}$. These observations can be extended to ordered n-tuples for any n and lead to the notion of a *Euclidean space*.

Definition 13.6. (Euclidean Space).

For each positive integer n, consider the set of all ordered n-tuples of real numbers. For any two elements $\mathbf{a} = (a_1, a_2, ..., a_n)$ and $\mathbf{b} = (b_1, b_2, ..., b_n)$ and a number c, put

$$\begin{aligned}\mathbf{a} + \mathbf{b} &= (a_1 + b_1, a_2 + b_2, ..., a_n + b_n),\\ c\,\mathbf{a} &= (ca_1, ca_2, ..., ca_n),\\ \mathbf{a}\cdot\mathbf{b} &= a_1b_1 + a_2b_2 + \cdots + a_nb_n,\\ \|\mathbf{a}\| &= \sqrt{\mathbf{a}\cdot\mathbf{a}} = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}\ .\end{aligned}$$

*The set of all ordered n-tuples in which the addition, the multiplication by a number, the dot product, and the norm are defined by these rules is called an n-*dimensional Euclidean space*.*

Two points of a Euclidean space are said to coincide, $\mathbf{a} = \mathbf{b}$, if the corresponding components are equal, that is, $a_i = b_i$ for $i = 1, 2, ..., n$. It follows that $\mathbf{a} = \mathbf{b}$ if and only if $\|\mathbf{a} - \mathbf{b}\| = 0$. Indeed, by the definition of the norm, $\|\mathbf{c}\| = 0$ if and only if $\mathbf{c} = (0, 0, ..., 0)$. Put $\mathbf{c} = \mathbf{a} - \mathbf{b}$. Then $\|\mathbf{a} - \mathbf{b}\| = 0$ if and only if $\mathbf{a} = \mathbf{b}$. The number $\|\mathbf{a} - \mathbf{b}\|$ is called the *distance* between points $\mathbf{a}$ and $\mathbf{b}$ of a Euclidean space.

The dot product in a Euclidean space has the same geometrical properties as in two and three dimensions. The Cauchy-Schwarz inequality can be extended to any Euclidean space (cf. Theorem 11.3).

Theorem 13.1. (Cauchy-Schwarz Inequality).

$$|\mathbf{a} \cdot \mathbf{b}| \leq \|\mathbf{a}\| \|\mathbf{b}\|$$

for any vectors $\mathbf{a}$ *and* $\mathbf{b}$ *in a Euclidean space, and the equality is reached if and only if* $\mathbf{a} = t\mathbf{b}$ *for some number* t.

Proof. Put $a = \|\mathbf{a}\|$ and $b = \|\mathbf{b}\|$, that is, $a^2 = \mathbf{a} \cdot \mathbf{a}$ and similarly for b. If $b = 0$, then $\mathbf{b} = \mathbf{0}$, and the conclusion of the theorem holds. For $b \neq 0$ and any real variable t, $\|\mathbf{a} - t\mathbf{b}\|^2 = (\mathbf{a} - t\mathbf{b}) \cdot (\mathbf{a} - t\mathbf{b}) \geq 0$. Therefore, $a^2 - 2tc + t^2 b^2 \geq 0$, where $c = \mathbf{a} \cdot \mathbf{b}$. The right side of this inequality is a downward concave parabola with respect to t, which attains its absolute minimum at $t = c/b^2$. Since the inequality is valid for any t, it is satisfied for $t = b^2/c$, that is, $a^2 - c^2/b^2 \geq 0$ or $c^2 \leq a^2 b^2$ or $|c| \leq ab$, which is the conclusion of the theorem. The inequality becomes an equality if and only if $\|\mathbf{a} - t\mathbf{b}\|^2 = 0$ and hence if and only if $\mathbf{a} = t\mathbf{b}$. □

It follows from the Schwarz inequality that $\mathbf{a} \cdot \mathbf{b} = s\|\mathbf{a}\| \|\mathbf{b}\|$, where s is a number such that $|s| \leq 1$. So one can always put $s = \cos\theta$, where $\theta \in [0, \pi]$. If $\theta = 0$, then $\mathbf{a} = t\mathbf{b}$ for some positive $t > 0$ (i.e., the vectors are parallel), and $\mathbf{a} = t\mathbf{b}$, $t < 0$, when $\theta = \pi$ (i.e., the vectors are antiparallel). The dot product vanishes when $\theta = \pi/2$. This allows one to define θ as the angle between vectors in any Euclidean space: $\cos\theta = \mathbf{a} \cdot \mathbf{b}/(\|\mathbf{a}\| \|\mathbf{b}\|)$ much like in two and three dimensions. Consequently, the triangle inequality (11.7) holds in a Euclidean space of any dimension.

In what follows, the domain of a function of n variables is viewed as a subset in an n-dimensional Euclidean space. It is also convenient to adopt the vector notation of the argument:

$$f(x_1, x_2, ..., x_n) = f(\mathbf{r}), \quad \mathbf{r} = (x_1, x_2, ..., x_n).$$

For example, the domain of the function $f(\mathbf{r}) = (1 - x_1^2 - x_2^2 - \cdots - x_n^2)^{1/2} = (1 - \|\mathbf{r}\|^2)^{1/2}$ is the set of points in the n-dimensional Euclidean space whose distance from the origin (the zero vector) does not exceed 1, $D = \{\mathbf{r} \,|\, \|\mathbf{r}\| \le 1\}$; that is, it is an n-dimensional ball of radius 1. So the domain of a multivariable function defined by an algebraic rule can be described by conditions on the components (coordinates) of the ordered n-tuple $\mathbf{r}$ under which the rule makes sense.

85.5. Exercises. **(1)** Find and sketch the domain of each of the following functions:
(i) $f(x, y) = \ln(9 - x^2 - (y/2)^2)$
(ii) $f(x, y) = \sqrt{1 - (x/2)^2 - (y/3)^2}$
(iii) $f(x, y, z) = \ln(1 - z + x^2 + y^2)$
(iv) $f(x, y) = \ln(9 - x^2 - (y/2)^2)$
(v) $f(x, y, z) = \sqrt{x^2 - y^2 - z^2}$
(vi) $f(t, \mathbf{x}) = (t^2 - \|\mathbf{x}\|^2)^{-1}$, $\mathbf{x} = (x_1, x_2, ..., x_n)$

(2) For each of the following functions, sketch the graph and a contour map:
(i) $f(x, y) = x^2 + 4y^2$
(ii) $f(x, y) = xy$
(iii) $f(x, y) = x^2 - y^2$
(iv) $f(x, y) = \sqrt{x^2 + 9y^2}$
(v) $f(x, y) = \sin x$

(3) Describe and sketch the level surfaces of each of the following functions:
(i) $f(x, y, z) = x + 2y + 3z$
(ii) $f(x, y, z) = x^2 + 4y^2 + 9z^2$
(iii) $f(x, y, z) = z + x^2 + y^2$
(iv) $f(x, y, z) = x^2 + y^2 - z^2$

(4) Explain how the graph $z = g(x, y)$ can be obtained from the graph of $f(x, y)$ if
(i) $g(x, y) = k + f(x, y)$, where k is a constant
(ii) $g(x, y) = mf(x, y)$, where m is a nonzero constant
(iii) $g(x, y) = f(x - a, y - b)$, where a and b are constants
(iv) $g(x, y) = f(px, qy)$, where p and q are nonzero constants
Let $f(x, y) = x^2 + y^2$. Sketch the graphs of $g(x, y)$ defined above. Analyze carefully various cases for values of the constants, for example, $m > 0$, $m < 0$, $p > 1$, $0 < p < 1$, and $p = -1$.

86. Limits and Continuity

The function $f(x) = \sin(x)/x$ is defined for all reals except $x = 0$. So the domain D of the function contains points arbitrarily close to the point $x = 0$, and therefore the limit of $f(x)$ can be studied as

$x \to 0$. It is known (see Calculus I) that $\sin(x)/x \to 1$ as $x \to 0$. A similar question can be asked for functions of several variables. For example, the domain of the function $f(x, y) = \sin(x^2 + y^2)/(x^2 + y^2)$ is the entire plane except the point $(x, y) = (0, 0)$. If $(x, y) \neq (0, 0)$, then, in contrast to the one-dimensional case, the point (x, y) may approach $(0, 0)$ along various paths. So the very notion that (x, y) approaches $(0, 0)$ needs to be accurately defined.

As noted before, the domain of a function f of several variables is a set in n-dimensional Euclidean space. Two points $\mathbf{x} = (x_1, x_2, ..., x_n)$ and $\mathbf{y} = (y_1, y_2, ..., y_n)$ coincide if and only if the distance

$$\|\mathbf{x} - \mathbf{y}\| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

vanishes.

Definition 13.7. *A point* $\mathbf{r}$ *is said to approach a fixed point* $\mathbf{r}_0$ *if the distance* $\|\mathbf{r} - \mathbf{r}_0\|$ *tends to 0. The limit* $\|\mathbf{r} - \mathbf{r}_0\| \to 0$ *is also denoted by* $\mathbf{r} \to \mathbf{r}_0$.

In the above example, the limit $(x, y) \to (0, 0)$ means that $\sqrt{x^2 + y^2} \to 0$ or $x^2 + y^2 \to 0$. Therefore,

$$\frac{\sin(x^2 + y^2)}{x^2 + y^2} = \frac{\sin u}{u} \to 1 \quad \text{as} \quad x^2 + y^2 = u \to 0.$$

Note that here the limit point $(0, 0)$ can be approached from any direction in the plane. This is not always so. For example, the domain of the function $f(x, y) = \sin(xy)/(\sqrt{x} + \sqrt{y})$ is the first quadrant, including its boundaries except the point $(0, 0)$. The points $(0, 0)$ and $(-1, -1)$ are not in the domain of the function. However, the limit of f as $(x, y) \to (0, 0)$ can be defined, whereas the limit of f as $(x, y) \to (-1, -1)$ does not make any sense. The difference between these two points is that any neighborhood of $(0, 0)$ contains points of the domain, while this is not so for $(-1, -1)$. So the limit can be defined only for some special class of points called *limit points* of a set D.

Definition 13.8. (Limit Point of a Set).
A point $\mathbf{r}_0$ *is said to be a limit point of a set* D *if any open ball* $B_\delta = \{\mathbf{r} \,|\, \|\mathbf{r} - \mathbf{r}_0\| < \delta\}$ *centered at* $\mathbf{r}_0$ *contains a point of* D *that does not coincide with* $\mathbf{r}_0$ *if* $\mathbf{r}_0 \in D$.

In other words, a limit point $\mathbf{r}_0$ of D may or may not be in D, but it can always be approached from within the set D in the sense that $\mathbf{r} \to \mathbf{r}_0$ and $\mathbf{r} \in D$ because, no matter how small δ is, one can always find a point $\mathbf{r} \in D$ whose distance from $\mathbf{r}_0$ is less than δ. In the above example of D being the first quadrant, the limit $(x, y) \to (0, 0)$ is understood as $x^2 + y^2 \to 0$ while $(x, y) \neq (0, 0)$ and $x \geq 0$, $y \geq 0$.

86.1. Limits of Functions of Several Variables.

Definition 13.9. (Limit of a Function of Several Variables). *Let f be a function of several variables whose domain is a set D in a Euclidean space. Let $\mathbf{r}_0$ be a limit point of D. Then the limit of $f(\mathbf{r})$ as $\mathbf{r} \to \mathbf{r}_0$ is said to be a number f_0 if, for every number $\varepsilon > 0$, there exists a corresponding number $\delta > 0$ such that if $\mathbf{r} \in D$ and $0 < \|\mathbf{r} - \mathbf{r}_0\| < \delta$, then $|f(\mathbf{r}) - f_0| < \varepsilon$. In this case, one writes*

$$\lim_{\mathbf{r} \to \mathbf{r}_0} f(\mathbf{r}) = f_0.$$

The number $|f(\mathbf{r}) - f_0|$ determines a deviation of the value of f from the number f_0. The existence of the limit means that no matter how small ε is, there is a neighborhood $N_\delta(\mathbf{r}_0)$ of $\mathbf{r}_0$ in D, which contains all points of D whose distance from $\mathbf{r}_0$ does not exceed a number δ, such that the values of the function f in $N_\delta(\mathbf{r}_0)$ deviate from the limit value f_0 no more than ε, that is, $f_0 - \varepsilon < f(\mathbf{r}) < f_0 + \varepsilon$ for all $\mathbf{r} \in N_\delta(\mathbf{r}_0) \subset D$.

The use of this definition is illustrated by the following example.

Example 13.5. *Show that*

$$\lim_{(x,y,z) \to (0,0,0)} (x^2 y + y z^2 - 6z^3) = 0.$$

Solution: The distance between $\mathbf{r} = (x, y, z)$ and the limit point $\mathbf{r}_0 = (0, 0, 0)$ is $R = \|\mathbf{r} - \mathbf{r}_0\| = \sqrt{x^2 + y^2 + z^2}$. Then $|x| \leq R$, $|y| \leq R$, and $|z| \leq R$. Consider the deviation of values of the function from the limiting value $f_0 = 0$:

$$|f(\mathbf{r}) - f_0| = |x^2 y + y z^2 - 6z^3| \leq |x^2 y| + |y z^2| + 6|z^3| \leq 8R^3,$$

where the inequality $|a \pm b| \leq |a| + |b|$ and $|ab| = |a||b|$ have been used. Now fix $\varepsilon > 0$. To establish the existence of $\delta > 0$, note that the inequality $8R^3 < \varepsilon$ or $R < \sqrt[3]{\varepsilon}/2$ guarantees that $|f(\mathbf{r}) - f_0| < \varepsilon$. Therefore, $\delta = \sqrt[3]{\varepsilon}/2$. For example, put $\varepsilon = 10^{-6}$. Then, in the interior of a ball of radius $\delta = 0.005$, the values of the function can deviate from $f_0 = 0$ no more than 10^{-6}. □

Remark. Note that δ depends on ε and, in general, on the limit point $\mathbf{r}_0$.

Remark. The definition of the limit guarantees that if the limit exists, then *it does depend on a path along which the limit point may be approached.* Indeed, take any curve that ends at the limit point $\mathbf{r}_0$ and fix $\varepsilon > 0$. Then, by the existence of the limit f_0, there is a ball of radius $\delta = \delta(\varepsilon, \mathbf{r}_0) > 0$ centered at $\mathbf{r}_0$ such that the values of f lie in the interval $f_0 - \varepsilon < f(\mathbf{r}_0) < f_0 + \varepsilon$ for all points $\mathbf{r}$ in the ball and hence for all points of the portion of the curve in the ball. Since ε can be chosen arbitrarily small, the limit along any path leading to $\mathbf{r}_0$ must be f_0. This is to be compared with the one-dimensional analog: if the limit of $f(x)$ exists as $x \to x_0$, then the right $x \to x_0^+$ and left $x \to x_0^-$ limits exist and are equal (and vice versa).

Properties of the Limit. The basic properties of limits of functions of one variable discussed in Calculus I are extended to the case of functions of several variables.

Theorem 13.2. (Properties of the Limit).
Let f and g be functions of several variables that have a common domain. Let c be a number. Suppose that $\lim_{\mathbf{r}\to\mathbf{r}_0} f(\mathbf{r}) = f_0$ and $\lim_{\mathbf{r}\to\mathbf{r}_0} g(\mathbf{r}) = g_0$. Then the following properties hold:

$$\begin{aligned}
\lim_{\mathbf{r}\to\mathbf{r}_0}(cf(\mathbf{r})) &= c\lim_{\mathbf{r}\to\mathbf{r}_0} f(\mathbf{r}) = cf_0,\\
\lim_{\mathbf{r}\to\mathbf{r}_0}(g(\mathbf{r}) + f(\mathbf{r})) &= \lim_{\mathbf{r}\to\mathbf{r}_0} g(\mathbf{r}) + \lim_{\mathbf{r}\to\mathbf{r}_0} f(\mathbf{r}) = g_0 + f_0,\\
\lim_{\mathbf{r}\to\mathbf{r}_0}(g(\mathbf{r})f(\mathbf{r})) &= \lim_{\mathbf{r}\to\mathbf{r}_0} g(\mathbf{r}) \lim_{\mathbf{r}\to\mathbf{r}_0} f(\mathbf{r}) = g_0 f_0,\\
\lim_{\mathbf{r}\to\mathbf{r}_0}\frac{g(\mathbf{r})}{f(\mathbf{r})} &= \frac{\lim_{\mathbf{r}\to\mathbf{r}_0} g(\mathbf{r})}{\lim_{\mathbf{r}\to\mathbf{r}_0} f(\mathbf{r})} = \frac{g_0}{f_0}, \quad \text{if} \quad f_0 \neq 0.
\end{aligned}$$

The proof of these properties follows the same line of reasoning as in the case of functions of one variable and is left to the reader as an exercise.

Squeeze Principle. The solution to Example 13.5 employs a rather general strategy to verify whether a particular number f_0 is the limit of $f(\mathbf{r})$ as $\mathbf{r} \to \mathbf{r}_0$.

Corollary 13.1. (Simplified Squeeze Principle).
If there exists a continuous function h of one variable such that

$$|f(\mathbf{r}) - f_0| \leq h(R) \to 0 \quad \text{as} \quad \|\mathbf{r} - \mathbf{r}_0\| = R \to 0^+,$$

then $\lim_{\mathbf{r}\to\mathbf{r}_0} f(\mathbf{r}) = f_0$.

In Example 13.5, $h(R) = 8R^3$. In general, the condition $h(R) \to 0$ as $R \to 0^+$ implies that, for any $\varepsilon > 0$, there is an interval $0 < R < \delta(\varepsilon)$ in which $h(R) < \varepsilon$, where the number δ can be found by solving the equation $h(\delta) = \varepsilon$. Hence, $|f(\mathbf{r}) - f_0| < \varepsilon$ whenever $\|\mathbf{r} - \mathbf{r}_0\| = R < \delta(\varepsilon)$. The corollary is also a consequence of a more general result called the *squeeze principle.*

Theorem 13.3. (Squeeze Principle).
Let the functions of several variables g, f, and h have a common domain D such that $g(\mathbf{r}) \le f(\mathbf{r}) \le h(\mathbf{r})$ for any $\mathbf{r} \in D$. If the limits of $g(\mathbf{r})$ and $h(\mathbf{r})$ as $\mathbf{r} \to \mathbf{r}_0$ exist and equal a number f_0, then the limit of $f(\mathbf{r})$ as $\mathbf{r} \to \mathbf{r}_0$ exists and equals f_0, that is,

$$g(\mathbf{r}) \le f(\mathbf{r}) \le h(\mathbf{r}) \text{ and } \lim_{\mathbf{r}\to\mathbf{r}_0} g(\mathbf{r}) = \lim_{\mathbf{r}\to\mathbf{r}_0} h(\mathbf{r}) = f_0 \;\Rightarrow\; \lim_{\mathbf{r}\to\mathbf{r}_0} f(\mathbf{r}) = f_0.$$

Proof. From the condition of the theorem, it follows that $0 \le f(\mathbf{r}) - g(\mathbf{r}) \le h(\mathbf{r}) - g(\mathbf{r})$. Put $F(\mathbf{r}) = f(\mathbf{r}) - g(\mathbf{r})$ and $H(\mathbf{r}) = h(\mathbf{r}) - g(\mathbf{r})$. Then $0 \le F(\mathbf{r}) \le H(\mathbf{r})$ implies $|F(\mathbf{r})| \le |H(\mathbf{r})|$ (the positivity of F is essential for this conclusion). By the basic properties of the limit, $H(\mathbf{r}) \to 0$ as $\mathbf{r} \to \mathbf{r}_0$. Hence, for any $\varepsilon > 0$, there is a corresponding number δ such that $0 \le |F(\mathbf{r})| \le |H(\mathbf{r})| < \varepsilon$ whenever $0 < \|\mathbf{r} - \mathbf{r}_0\| < \delta$. This inequality also implies that $\lim_{\mathbf{r}\to\mathbf{r}_0} F(\mathbf{r}) = 0$. By the basic properties of the limit, it is then concluded that $f(\mathbf{r}) = F(\mathbf{r}) + g(\mathbf{r}) \to 0 + f_0 = f_0$ as $\mathbf{r} \to \mathbf{r}_0$. □

The simplified squeeze principle is a particular case of this theorem because the condition $|f(\mathbf{r}) - f_0| \le h(R)$ is equivalent to $f_0 - h(R) \le f(\mathbf{r}) \le f_0 + h(R)$.

Example 13.6. *Show that*

$$\lim_{(x,y)\to(0,0)} f(x,y) = 0, \quad \text{where} \quad f(x,y) = \frac{x^3y - 3x^2y^2}{x^2 + y^2 + x^4}.$$

Solution: Let $R = \sqrt{x^2 + y^2}$ (the distance from the limit point $(0,0)$). Then $|x| \le R$ and $|y| \le R$. Therefore,

$$\frac{|x^3y - 3x^2y^2|}{x^2 + y^2 + x^4} \le \frac{|x|^3|y| + 3x^2y^2}{x^2 + y^2 + x^4} \le \frac{4R^4}{R^2 + x^4} \le 4R^2 \frac{1}{1 + (x^4/R^2)} \le 4R^2.$$

It follows from this inequality that $-4(x^2 + y^2) \le f(x,y) \le 4(x^2 + y^2)$, and, by the squeeze principle, $f(x,y)$ must tend to 0 because $\pm 4(x^2 + y^2) = \pm 4R^2 \to 0$ as $R \to 0$. In the definition of the limit, for any $\varepsilon > 0$, the corresponding number δ is $\delta = \sqrt{\varepsilon}/2$. □

86.2. Continuity of Functions of Several Variables.

Definition 13.10. (Continuity).
A function f of several variables with domain D is said to be continuous at a point $\mathbf{r}_0 \in D$ if

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

The function f is said to be continuous on D if it is continuous at every point of D.

Let $f(x,y)=1$ if $y\geq x$ and let $f(x,y)=0$ if $y<x$. The function is continuous at every point (x_0,y_0) if $y_0\neq x_0$. Indeed, if $y_0>x_0$, then $f(x_0,y_0)=1$. On the other hand, for every such point one can find a neighborhood $(x-x_0)+(y-y_0)^2<\delta^2$ (a disk of radius $\delta>0$ centered at (x_0,y_0)) that lies in the region $y>x$. Therefore, $|f(\mathbf{r})-f(\mathbf{r}_0)|=1-1=0<\varepsilon$ for any $\varepsilon>0$ in this disk, that is, $\lim_{\mathbf{r}\to\mathbf{r}_0}f(\mathbf{r})=f(\mathbf{r}_0)=1$. The same line of arguments applies to establish the continuity of f at any point (x_0,y_0), where $y_0<x_0$. If $\mathbf{r}_0=(x_0,x_0)$ that is, the point lies on the line $y=x$), then $f(\mathbf{r}_0)=1$. Any disk centered at such $\mathbf{r}_0$ is split into two parts by the line $y=x$. In one part $(y\geq x)$, $f(\mathbf{r})=1$, whereas in the other part $(y<x)$, $f(\mathbf{r})=0$. So, for $0<\varepsilon<1$, there is no disk of radius $\delta>0$ in which $|f(\mathbf{r})-f(\mathbf{r}_0)|=|f(\mathbf{r})-1|<\varepsilon$ because $|f(\mathbf{r})-1|=1$ for $y<x$ in any such disk. The function is not continuous along the line $y=x$ in its domain.

Theorem 13.4. (Properties of Continuous Functions).
If f and g are continuous on D and c is a number, then $cf(\mathbf{r})$, $f(\mathbf{r})+g(\mathbf{r})$, and $f(\mathbf{r})g(\mathbf{r})$ are continuous on D, and $f(\mathbf{r})/g(\mathbf{r})$ is continuous at any point on D for which $g(\mathbf{r})\neq 0$.

This theorem is a simple consequence of the basic properties of the limit.

The use of the definition to establish the continuity of a function defined by an algebraic rule is not convenient. The following two theorems are helpful when studying the continuity of a given function.

For an ordered n-tuple $\mathbf{r}=(x_1,x_2,...,x_n)$, the function $x_1^{k_1}x_2^{k_2}\cdots x_n^{k_n}$, where $k_1,k_2,...,k_n$ are nonnegative integers, is called a *monomial of degree* $N=k_1+k_2+\cdots+k_n$. For example, for two variables, monomials of degree $N=3$ are x^3, x^2y, xy^2, and y^3. A function f that is a linear combination of monomials is called a *polynomial function*. The ratio of two polynomial functions is called a *rational function*.

Theorem 13.5. (Continuity of Polynomial and Rational Functions).
Let f and g be the polynomial functions of several variables. Then they are continuous everywhere, and the rational function $f(\mathbf{r})/g(\mathbf{r})$ is continuous at any point $\mathbf{r}_0$ if $g(\mathbf{r}_0)\neq 0$.

Proof. A polynomial function in which the argument $(x_1,x_2,...,x_n)$ is changed to $(x_1+a_1,x_2+a_2,...,x_n+a_n)$, where $a_1,a_2,...,a_n$ are constants, is also a polynomial function. So it is sufficient to establish continuity at any particular point, say, the origin. Also, by the basic properties of the limit, the continuity of monomial functions implies the continuity of polynomial functions. The monomial of degree $N=0$ is a constant function that is continuous. For a monomial function f of degree $N>0$

and origin $\mathbf{r}_0$, $f(\mathbf{r}_0) = 0$, one has $|f(\mathbf{r}) - f(\mathbf{r}_0)| = |f(\mathbf{r})| \leq R^N \to 0$ as $R \to 0$ because $|x_i| \leq R = \sqrt{x_1^2 + x_2^2 + \cdots x_n^2}$ for any element of the n-tuple. By the squeeze principle, $f(\mathbf{r}) \to 0 = f(\mathbf{r}_0)$. The rational function $f(\mathbf{r})/g(\mathbf{r})$ is continuous as the ratio of two continuous functions. □

Theorem 13.6. (Continuity of a Composition).
Let $g(u)$ be continuous on the interval $u \in [a, b]$ and let h be a function of several variables that is continuous on D and has the range $[a, b]$. The composition $f(\mathbf{r}) = g(h(\mathbf{r}))$ is continuous on D.

The proof follows the same line of reasoning as in the case of the composition of two functions of one variable in Calculus I and is left to the reader as an exercise.

In particular, some basic functions studied in Calculus I, $\sin u$, $\cos u$, e^u, $\ln u$, and so on, are continuous functions on their domains. If $f(\mathbf{r})$ is a continuous function of several variables, the elementary functions whose argument is replaced by $f(\mathbf{r})$ are continuous functions. In combination with the properties of continuous functions, the composition rule defines a large class of continuous functions of several variables, which is sufficient for many practical applications.

86.3. Exercises. **(1)** Use the definition of the limit to verify each of the following limits (i.e., given $\varepsilon > 0$, find the corresponding $\delta(\varepsilon)$):

$$\text{(i)} \qquad \lim_{\mathbf{r}\to\mathbf{0}} \frac{x^3 - 4y^2x + 5y^3}{x^2 + y^2} = 0$$

$$\text{(ii)} \qquad \lim_{\mathbf{r}\to\mathbf{0}} \frac{x^3 - 4y^2x + 5y^3}{3x^2 + 4y^2} = 0$$

$$\text{(iii)} \qquad \lim_{\mathbf{r}\to\mathbf{0}} \frac{x^3 - 4y^4 + 5y^3x^2}{3x^2 + 4y^2} = 0$$

$$\text{(iv)} \qquad \lim_{\mathbf{r}\to\mathbf{0}} \frac{x^3 - 4y^2x + 5y^3}{3x^2 + 4y^2 + y^4} = 0$$

(2) Verify whether the given function is continuous on its domain:
(i) $f(x, y) = yx/(x^2 + y^2)$ if $(x, y) \neq (0, 0)$ and $f(0, 0) = 1$
(ii) $f(x, y, z) = yxz/(x^2+y^2+z^2)$ if $(x, y, z) \neq (0, 0, 0)$ and $f(0, 0, 0) = 0$
(iii) $f(x, y) = \sin(\sqrt{xy})$
(iv) $f(x, y) = \cos(\sqrt{xyz})/(x^2y^2 + 1)$
(v) $f(x, y) = (x^2 + y^2)\ln(x^2 + y^2)$ if $(x, y) \neq (0, 0)$ and $f(0, 0) = 0$

87. A General Strategy to Study Limits

The definition of the limit gives only the criterion for whether a number f_0 is the limit of $f(\mathbf{r})$ as $\mathbf{r} \to \mathbf{r}_0$. In practice, however, a possible value of the limit is typically unknown. Some studies are needed to make an "educated" guess for a possible value of the limit. Here a procedure to study limits is outlined that might be helpful. In what follows, the limit point is often set to the origin $\mathbf{r}_0 = (0, 0, ..., 0)$. This is not a limitation because one can always translate the origin of the coordinate system to any particular point by shifting the values of the argument, for example,

$$\lim_{(x,y)\to(x_0,y_0)} f(x,y) = \lim_{(x,y)\to(0,0)} f(x + x_0, y + y_0)\,.$$

87.1. Step 1: Continuity Argument.

The simplest scenario in studying the limit happens when the function f in question is continuous at the limit point:

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

For example,

$$\lim_{(x,y)\to(1,2)} \frac{xy}{x^3 - y^2} = -\frac{2}{3}$$

because the function in question is a rational function that is continuous if $x^3 - y^2 \neq 0$. The latter is indeed the case for the limit point $(1, 2)$. If the continuity argument does not apply, then it is helpful to check the following.

87.2. Step 2: Composition Rule.

Theorem 13.7. (Composition Rule for Limits).
Let $g(t)$ be a function of one variable and let h be a continuous function of several variables such that $h(\mathbf{r}) \to t_0 = h(\mathbf{r}_0)$ as $\mathbf{r} \to \mathbf{r}_0$. Suppose that the function f is the composition $f(\mathbf{r}) = g(h(\mathbf{r}))$. Then

$$\lim_{\mathbf{r}\to\mathbf{r}_0} f(\mathbf{r}) = \lim_{t\to t_0} g(t).$$

The proof is omitted as it is similar to the case when h is a continuous function of one variable, which was proved in Calculus I. The significance of this theorem is that, under the conditions of the theorem, a tough problem of studying a multivariable limit is reduced to the problem of the limit of a function of a single argument. The latter problem can be studied, by, for example, l'Hospital's rule. It must be emphasized that *there is no analog of l'Hospital's rule for multivariable functions.*

Example 13.7. *Find*

$$\lim_{(x,y)\to(0,0)} \frac{\cos(xy)-1}{x^2y^2}.$$

Solution: The function in question is $(\cos t - 1)/t^2$, where the argument t is replaced by the function $h(x,y) = xy$. The function h is a polynomial and hence continuous. In particular, $h(x,y) \to h(0,0) = 0$ as $(x,y) \to (0,0)$. Thus, all the hypotheses of the composition rule theorem are fulfilled:

$$\lim_{(x,y)\to(0,0)} \frac{\cos(xy)-1}{x^2y^2} = \lim_{t\to 0}\frac{\cos t - 1}{t^2} = \lim_{t\to 0}\frac{-\sin t}{2t} = \lim_{t\to 0}\frac{-\cos t}{2} = -\frac{1}{2}.$$

where l'Hospital's rule has been used twice to evaluate the single-variable limit. □

It must be stressed that the hypothesis of the continuity of $h(\mathbf{r})$ in the composition rule theorem is crucial. If, for instance, in the above example the argument of $(\cos t - 1)/t^2$ is replaced by $h(x,y) = y/x$, the limit of the resulting function does not exist as $(x,y) \to (0,0)$. The reason is that y/x is not continuous at the origin and its limit does not exist as $(x,y) \to (0,0)$. Simple means to establish the latter fact are provided in the next step.

87.3. Step 3: Limits Along Curves.

Recall the following result about the limit of a function of one variable. The limit of $f(x)$ as $x \to x_0$ exists and equals f_0 if and only if the corresponding right and left limits of $f(x)$ exist and equal f_0:

$$\lim_{x\to x_0^+} f(x) = \lim_{x\to x_0^-} f(x) = f_0 \quad\Longleftrightarrow\quad \lim_{x\to x_0} f(x) = f_0.$$

In other words, if the limit exists, it does not depend on the direction from which the limit point is approached. If the left and right limits exist but do not coincide, then the limit does not exist.

For functions of several variables, there are infinitely many paths along which the limit point can be approached. They include straight lines and paths of any other shape, in contrast to the one-variable case. Nevertheless, a similar result holds for multivariable limits (see the second Remark at the end of Section **86.1**), that is, *if the limit exists, then it should not depend on the curve along which the limit point may be approached. If there are two curves along which the limits do not coincide, then the multivariable limit does not exist.* This result provides a powerful method to investigate the existence of a multivariable limit and to make an "educated" guess about its possible value.

Definition 13.11. (Curve in a Euclidean Space).
A curve in a Euclidean space is a set of points $\mathbf{r}(t) = (x_1(t), x_2(t), ..., x_n(t))$*, where* $x_i(t)$*,* $i = 1, 2, ..., n$*, are continuous functions of a variable* $t \in [a, b]$*.*

This is a natural generalization of the concept of a curve in a plane or space as a vector function defined by the parametric equations $x_i = x_i(t)$, $i = 1, 2, ..., n$.

Definition 13.12. (Limit Along a Curve).
Let $\mathbf{r}_0$ *be a limit point of the domain* D *of a function* f*. Let* $\mathbf{r}(t) = (x_1(t), x_2(t), ..., x_n(t))$ *be a curve* C *in* D *such that* $\mathbf{r}(t) \to \mathbf{r}_0$ *as* $t \to t_0^+$*. The function* $F(t) = f(\mathbf{r}(t))$ *defines the values of* f *on the curve* C*. The limit*

$$\lim_{t \to t_0^+} F(t) = \lim_{t \to t_0^+} f(x_1(t), x_2(t), ..., x_n(t))$$

is called the limit of f *along the curve* C *if it exists.*

Suppose that the limit of $f(\mathbf{r})$ as $\mathbf{r} \to \mathbf{r}_0$ exists and equals f_0. Let C be a curve such that $\mathbf{r}(t) \to \mathbf{r}_0$ as $t \to t_0^+$. Fix $\varepsilon > 0$. By the existence of the limit, there is a neighborhood $N_\delta(\mathbf{r}_0) = \{\mathbf{r} \,|\, \mathbf{r} \in D, 0 < \|\mathbf{r} - \mathbf{r}_0\| < \delta\}$ in which the values of f deviate from f_0 no more than ε, $|f(\mathbf{r}) - f_0| < \varepsilon$. Since the curve C is continuous and passes through $\mathbf{r}_0$, there should be a portion of it that lies in $N_\delta(\mathbf{r}_0)$; that is, there is a number δ' such that $\|\mathbf{r}(t) - \mathbf{r}_0\| < \delta$ for all $t \in (t_0, t_0 + \delta')$. Hence, for any $\varepsilon > 0$, the deviation of values of f along the curve, $F(t) = f(\mathbf{r}(t))$, does not exceed ε, $|F(t) - f_0| < \varepsilon$ whenever $0 < |t - t_0| < \delta'$. By the definition of the one-variable limit, this implies that $F(t) \to f_0$ as $t \to t_0$ for any curve C through $\mathbf{r}_0$. This proves the following.

Theorem 13.8. (Independence of the Limit from a Curve Through the Limit Point).
If the limit of $f(\mathbf{r})$ *exists as* $\mathbf{r} \to \mathbf{r}_0$*, then the limit of* f *along any curve leading to* $\mathbf{r}_0$ *from within the domain of* f *exists and does not depend on the curve.*

An immediate consequence of this theorem is the following result, which is very useful in many practical applications

Corollary 13.2. (Criterion for Nonexistence of the Limit).
Let f *be a function of several variables on* D*. If there is a curve* $\mathbf{r}(t)$ *in* D *such that* $\mathbf{r}(t) \to \mathbf{r}_0$ *as* $t \to t_0^+$ *and the limit* $\lim_{t \to t_0^+} f(\mathbf{r}(t))$ *does not exist, then the multivariable limit* $\lim_{\mathbf{r} \to \mathbf{r}_0} f(\mathbf{r})$ *does not exist either. If there are two curves in* D *leading to* $\mathbf{r}_0$ *such that the limits of* f *along*

them exist but do not coincide, then the multivariable limit $\lim_{\mathbf{r}\to\mathbf{r}_0} f(\mathbf{r})$ *does not exist.*

87.3.1. Limits Along Straight Lines. Let the limit point be the origin $\mathbf{r}_0 = (0, 0, ..., 0)$. The simplest curve leading to $\mathbf{r}_0$ is a straight line $x_i = v_i t$, where $t \to 0^+$ for some numbers v_i, $i = 1, 2, ..., n$. The limit of a function of several variables f along a straight line is then $\lim_{t\to 0^+} f(v_1 t,\, v_2 t, ..., v_n t)$, should exist and be the same for any choice of numbers v_i. For comparison, recall the vector equation of a straight line in space through the origin: $\mathbf{r} = t\mathbf{v}$, where $\mathbf{v}$ is a vector parallel to the line.

EXAMPLE 13.8. *Investigate the two-variable limit*

$$\lim_{(x,y)\to(0,0)} \frac{xy^3}{x^4 + 2y^4}.$$

SOLUTION: Consider the limits along straight lines $x = t$, $y = at$ (or $y = ax$, where a is the slope) as $t \to 0^+$:

$$\lim_{t\to 0^+} f(t, at) = \lim_{t\to 0^+} \frac{a^3 t^4}{t^4(1 + 2a^4)} = \frac{a^3}{1 + 2a^4}\,.$$

So the limit along a straight line depends on the slope of the line. Therefore, the two-variable does not exist. □

EXAMPLE 13.9. *Investigate the limit*

$$\lim_{(x,y)\to(0,0)} \frac{\sin(\sqrt{xy})}{x + y}.$$

SOLUTION: The domain of the function consists of the first and third quadrants as $xy \geq 0$ except the origin. Lines approaching $(0, 0)$ from within the domain are $x = t$, $y = at$, $a \geq 0$ and $t \to 0$. Note the line $x = 0$, $y = t$ also lies in the domain (the line with an infinite slope). The limit along a straight line approaching the origin from within the first quadrant is

$$\lim_{t\to 0^+} f(t, at) = \lim_{t\to 0^+} \frac{\sin(t\sqrt{a})}{t(1 + a)} = \lim_{t\to 0^+} \frac{\sqrt{a}\cos(t\sqrt{a})}{1 + a} = \frac{\sqrt{a}}{1 + a},$$

where l'Hospital's rule has been used to calculate the limit. The limit depends on the slope of the line, and hence the two-variable limit does not exist. □

87.3.2. Limits Along Power Curves (Optional). If the limit along straight lines exists and is independent of the choice of the line, the numerical value of this limit provides a desired "educated" guess for the actual multivariable limit. However, this has yet to be proved by means of either the definition of the multivariable limit or, for example, the squeeze principle. This comprises the last step of the analysis of limits (Step 4; see below).

The following should be stressed. *If the limits along all straight lines happen to be the same number, this does not mean that the multivariable limit exists and equals that number because there might exist other curves through the limit point along which the limit attains a different value or does not even exist.*

Example 13.10. *Investigate the limit*

$$\lim_{(x,y)\to(0,0)} \frac{y^3}{x}.$$

Solution: The domain of the function is the whole plane with the y axis removed ($x \neq 0$). The limit along a straight line

$$\lim_{t\to 0^+} f(t, at) = \lim_{t\to 0^+} \frac{a^3t^3}{t} = a^3 \lim_{t\to 0^+} t^2 = 0$$

vanishes for any slope; that is, it is independent of the choice of the line. However, the two-variable does not exist! Consider the power curve $x = t$, $y = at^{1/3}$ approaching the origin as $t \to 0^+$. The limit along this curve can attain any value by varying the parameter a:

$$\lim_{t\to 0^+} f(t, at^{1/3}) = \lim_{t\to 0^+} \frac{a^3t}{t} = a^3.$$

Thus, the multivariable limit does not exist. □

In general, limits along power curves are convenient for studying limits of rational functions because the values of a rational function of several variables on a power curve are given by a rational function of the curve parameter t. One can then adjust, if possible, the power parameter of the curve so that the leading terms of the top and bottom power functions match in the limit $t \to 0^+$. For instance, in the example considered, put $x = t$ and $y = at^n$. Then $f(t, at^n) = (a^3t^{3n})/t$. The powers of the top and bottom functions in this ratio match if $3n = 1$; hence, for $n = 1/3$, the limit along the power path depends on the parameter a and can be any number.

87.4. Step 4: Using the Squeeze Principle. If Steps 1 and 2 do not apply to the multivariable limit in question, then an "educated" guess for a possible value of the limit is helpful. This is the outcome of Step 3. If limits along a family of curves (e.g., straight lines) happen to be the same number f_0, then this number is the sought-for "educated" guess. The definition of the multivariable limit or the squeeze principle can be used to prove or disprove that f_0 is the multivariable limit.

EXAMPLE 13.11. *Find the limit or prove that it does not exist:*

$$\lim_{(x,y)\to(0,0)} \frac{\sin(xy^2)}{x^2+y^2}.$$

SOLUTION:
Step 1. The function is not defined at the origin. The continuity argument does not apply.
Step 2. No substitution exists to transform the two-variable limit to a one-variable limit.
Step 3. Put $(x,y) = (t, at)$, where $t \to 0^+$. The limit along straight lines

$$\begin{aligned}\lim_{t\to 0^+} f(t,at) &= \lim_{t\to 0^+} \frac{\sin(a^2t^3)}{t^2} = \lim_{u\to 0^+} \frac{\sin(a^2u^{3/2})}{u}\\ &= \lim_{u\to 0^+} \frac{(3/2)a^2u^{1/2}\cos(a^2u^{3/2})}{1} = 0\end{aligned}$$

vanishes (here the substitution $u = t^2$ and l'Hospital's rule have been used to calculate the limit).
Step 4. If the two-variable limit exists, then it must be equal to 0. This can be verified by means of the simplified squeeze principle; that is, one has to verify that there exists $h(R)$ such that $|f(x,y) - f_0| = |f(x,y)| \le h(R) \to 0$ as $R = \sqrt{x^2+y^2} \to 0$. A key technical trick here is the inequality $|\sin u| \le |u|$, which holds for any real u. One has

$$|f(x,y) - 0| = \frac{|\sin(xy^2)|}{x^2+y^2} \le \frac{|xy^2|}{x^2+y^2} \le \frac{R^3}{R^2} = R \to 0,$$

where the inequalities $|x| \le R$ and $|y| \le R$ have been used. Thus, the two-variable limit exists and equals 0. □

For two-variable limits, it is sometimes convenient to use polar coordinates centered at the limit point $x - x_0 = R\cos\theta$, $y - y_0 = R\sin\theta$. The idea is to find out whether the deviation of the function $f(x,y)$ from f_0 (the "educated" guess from Step 3) can be bounded by $h(R)$ uniformly for all $\theta \in [0, 2\pi]$:

$$|f(x,y) - f_0| = |f(x_0 + R\cos\theta, y_0 + R\sin\theta) - f_0| \le h(R) \to 0 \quad R \to 0.$$

This technical task can be accomplished with the help of the basic properties of trigonometric functions, for example, $|\sin\theta| \leq 1$, $|\cos\theta| \leq 1$, and so on.

In Example 13.10, Step 3 gives $f_0 = 0$ if only the limits along straight lines have been studied. Then $|f(R\cos\theta, R\sin\theta)| = R^2\sin^2(\theta)$ $|\tan\theta|$. Despite that the deviation is proportional to $R^2 \to 0$ as $R \to 0$, it cannot be made as small as desired by decreasing R because $\tan\theta$ is not a bounded function. There is a sector in the plane corresponding to angles near $\theta = \pi/2$ where $\tan\theta$ can be any large number whereas $\sin^2\theta$ is *strictly* positive in it so that the deviation of f from 0 can be as large as desired no matter how small R is. So, for any $\varepsilon > 0$, the inequality $|f(\mathbf{r}) - f_0| < \varepsilon$ is violated in that sector of any disk $\|\mathbf{r} - \mathbf{r}_0\| < \delta$, and hence the limit does not exist.

Remark. For multivariable limits with $n > 2$, a similar approach exists. If, for simplicity, $\mathbf{r}_0 = (0, 0, ..., 0)$. Then put $x_i = Ru_i$, where the variables u_i satisfy the condition $u_1^2 + u_2^2 + \cdots + u_n^2 = 1$. For $n = 2$, $u_1 = \cos\theta$ and $u_2 = \sin\theta$. For $n \geq 3$, the variables u_i can be viewed as the directional cosines, that is, the cosines of the angles between $\mathbf{r}$ and $\hat{\mathbf{e}}_i$, $u_i = \mathbf{r} \cdot \mathbf{e}_i/\|\mathbf{r}\|$. Then one has to investigate whether there is $h(R)$ such that

$$|f(Ru_1, Ru_2, ..., Ru_n) - f_0| \leq h(R) \to 0\,, \quad R \to 0\,.$$

This technical, often rather difficult, task may be accomplished using the inequalities $|u_i| \leq 1$ and some specific properties of the function f. As noted, the variables u_i are the directional cosines. They can also be trigonometric functions of the angles in the spherical coordinate system in an n-dimensional Euclidean space. The problem of the existence or non-existence of the limit amounts to studying the behavior of some trigonometric functions.

87.5. Study Problems.

Problem 13.1. *Find the limit* $\lim_{\mathbf{r}\to\mathbf{r}_0} f(\mathbf{r})$ *or show that it does not exist, where*

$$f(\mathbf{r}) = f(x, y, z) = (x^2 + 2y^2 + 4z^2)\ln(x^4 + y^4 + z^4)\,, \quad \mathbf{r}_0 = (0, 0, 0).$$

SOLUTION:

Step 1. The continuity argument does not apply because f is not defined at $\mathbf{r}_0$.

Step 2. No substitution is possible to transform the limit to a one-variable limit.

Step 3. Put $\mathbf{r}(t) = (t, at, bt)$ for some constants a and b that define the direction of the line. Then $f(\mathbf{r}(t)) = At^2 \ln(Bt^4) = 4At^2 \ln(t) + A\ln(B)t^2 \to 0$ as $t \to 0^+$, where $A = 1 + 2a^2 + 4b^2$ and $B = 1 + a^4 + b^4$ (recall that by l'Hospital's rule $t\ln(t) = \ln(t)/t^{-1} \to 0$ as $t \to 0^+$). So, if the limit exists, then it must be equal to 0.
Step 4. Put $R^2 = x^2+y^2+z^2$. By making use of the inequalities $|x| \le R$, $|y| \le R$, $|z| \le R$, one has $x^2 + 2y^2 + 4z^2 \le 7R^2$ and $x^4 + y^4 + z^4 \le 3R^4$. Hence, by the *monotonicity* of the logarithm function,

$$|f(\mathbf{r}) - 0| \le 7R^2 \ln(3R^4) = 7R^2(4\ln(R) + \ln(3)) \to 0 \quad \text{as} \quad R \to 0^+.$$

By the squeeze principle, the limit exists and equals 0. □

Problem 13.2. *Prove that the limit* $\lim_{\mathbf{r}\to\mathbf{r}_0} f(\mathbf{r})$ *exists, where*

$$f(\mathbf{r}) = f(x, y) = \frac{1 - \cos(x^2y)}{x^2 + 2y^2}\,, \quad \mathbf{r}_0 = (0, 0)\,,$$

and find a disk centered at $\mathbf{r}_0$ *in which values of* f *deviate from the limit no more than* $\varepsilon = 0.5 \times 10^{-4}$.

Solution:
Step 1. The continuity argument does not apply because f is not defined at $\mathbf{r}_0$.
Step 2. No substitution is possible to transform the limit to a one-variable limit.
Step 3. Put $\mathbf{r}(t) = (t, at)$. Then

$$\lim_{t\to 0^+} f(\mathbf{r}(t)) = \lim_{t\to 0^+} \frac{1 - \cos(at^3)}{t^2(1 + 2a^2)} = \frac{1}{1 + 2a^2} \lim_{u\to 0^+} \frac{1 - \cos(au^{3/2})}{u}$$
$$= \frac{1}{1 + 2a^2} \lim_{u\to 0^+} \frac{au^{1/2}\sin(au^{3/2})}{1} = 0,$$

where the substitution $u = t^2$ and l'Hospital's rule have been used to evaluate the limit. Therefore, if the limit exists, it must be equal to 0.
Step 4. Note first that $1 - \cos u = 2\sin^2(u/2) \le u^2/2$, where the inequality $|\sin x| \le |x|$ has been used. Put $R^2 = x^2 + y^2$. Then, by making use of the above inequality with $u = x^2y$ together with $|x| \le R$ and $|y| \le R$, the following chain of inequalities is obtained:

$$|f(\mathbf{r}) - 0| \le \frac{(x^2y)^2/2}{x^2 + 2y^2} = \frac{(x^2y)^2/2}{R^2 + y^2} \le \frac{(x^2y)^2/2}{R^2} \le \frac{1}{2}\frac{R^6}{R^2} = \frac{R^4}{2} \to 0$$

as $R \to 0^+$. By the squeeze principle, the limit exists and equals 0. From the above inequality, it follows that $|f(\mathbf{r})| < \varepsilon$ if $R^4/2 < \varepsilon$ and hence $\|\mathbf{r} - \mathbf{r}_0\| = R < \delta(\varepsilon) = (2\varepsilon)^{1/4} = 0.1$. □

Problem 13.3. *Find the limit* $\lim_{\mathbf{r}\to\mathbf{r}_0} f(\mathbf{r})$ *or show that it does not exist, where*

$$f(\mathbf{r}) = f(x,y) = \frac{x^2y}{x^2-y^2}\,, \quad \mathbf{r}_0 = (0,0).$$

SOLUTION:
Step 1. The continuity argument does not apply because f is not defined at $\mathbf{r}_0$.
Step 2. No substitution is possible to transform the limit to a one-variable limit.
Step 3. The domain D of the function is the whole plane with the lines $y = \pm x$ excluded. So put $\mathbf{r}(t) = (t, at)$, where $a \neq \pm 1$. Then $f(\mathbf{r}(t)) = at^3/t^2(1-a^2) = a(1-a^2)^{-1}t \to 0$ as $t \to 0^+$. So, if the limit exists, then it must be equal to 0.
Step 4. In polar coordinates, $x = R\cos\theta$ and $y = R\sin\theta$, where $\|\mathbf{r}-\mathbf{r}_0\| = R$,

$$f(\mathbf{r}) = \frac{R^3\cos^2\theta\sin\theta}{R^2(\cos^2\theta-\sin^2\theta)} = \frac{1}{2}\,\frac{R\cos\theta\sin(2\theta)}{\cos(2\theta)} = \frac{R\cos\theta}{2}\,\tan(2\theta).$$

Therefore, in any disk $0 < \|\mathbf{r}-\mathbf{r}_0\| < R$, there is a sector corresponding to the polar angle $\pi/4 < \theta < \pi/4 + \Delta\theta$ in which the deviation $|f(\mathbf{r}) - 0|$ can be made larger that any positive number by taking $\Delta\theta > 0$ small enough because $\tan(2\theta)$ is not bounded in this interval. Hence, for any $\varepsilon > 0$, there is no $\delta > 0$ such that $|f(\mathbf{r})| < \varepsilon$ whenever $\mathbf{r} \in D$ lies in the disk $\|\mathbf{r}-\mathbf{r}_0\| < \delta$. Thus, the limit does not exist.
Step 3 (Optional). The nonexistence of the limit established in Step 4 implies that there should exist curves along which the limit differs from 0. It is instructive to demonstrate this explicitly. Any such curve should approach the origin from within one of the narrow sectors containing the lines $y = \pm x$ (where $\tan(2\theta)$ takes large values). So put, for example, $\mathbf{r}(t) = (t, t - at^n)$, where $n > 1$ and $a \neq 0$ is a number. Then $f(\mathbf{r}(t)) = (t^3 + at^{n+2})/(2at^{n+1} - a^2t^{2n})$. This function tends to a number as $t \to 0^+$ if n is chosen to match the leading (smallest) powers of the top and bottom of the ratio in this limit (i.e., $3 = n+1$ or $n = 2$). Thus, for $n = 2$, $f(\mathbf{r}(t)) \to 1/(2a)$ as $t \to 0^+$ and $f(\mathbf{r}(t))$ diverges for $n > 2$ in this limit. □

87.6. Exercises. **(1)** Find each of the following limits or show that it does not exist:

(i) $\frac{\cos(xy+z)}{x^4+y^2z^2+4}$

(ii) $\lim_{\mathbf{r}\to\mathbf{0}} \frac{\cos^2(xy)-1}{xy}$

(iii) $\lim_{\mathbf{r}\to\mathbf{0}} \frac{\sqrt{xy^2+1}-1}{xy^2}$

(iv) $\lim_{\mathbf{r}\to\mathbf{0}} \frac{\sin(xy^3)}{x^2}$

(v) $\lim_{\mathbf{r}\to\mathbf{0}} \frac{x^3+y^5}{x^2+2y^2}$

(vi) $\lim_{\mathbf{r}\to\mathbf{0}} \frac{e^{\|\mathbf{r}\|}-1-\|\mathbf{r}\|}{\|\mathbf{r}\|^2}$

(vii) $\lim_{\mathbf{r}\to\mathbf{0}} \frac{x^2+\sin^2 y}{x^2+2y^2}$

(viii) $\lim_{\mathbf{r}\to\mathbf{0}} \frac{xy^2+x\sin(xy)}{x^2+2y^2}$

88. Partial Derivatives

The derivative $f'(x_0)$ of a function $f(x)$ at $x = x_0$ contains important information about the local behavior of the function near $x = x_0$. It defines the slope of the tangent line $L(x) = f(x_0) + f'(x_0)(x - x_0)$, and, for x close enough to x_0, values of f can be well approximated by the linearization $L(x)$, that is, $f(x) \approx L(x)$. In particular, if $f'(x_0) > 0$, f increases near x_0, and, if $f'(x_0) < 0$, f decreases near x_0. Furthermore, the second derivative $f''(x_0)$ supplies more information about f near x_0, namely, its concavity.

It is therefore important to develop a similar concept for functions of several variables in order to study their local behavior. A significant difference is that, given a point in the domain, the rate of change is going to depend on the direction in which it is measured. For example, if $f(\mathbf{r})$ is the height of a hill as a function of position $\mathbf{r}$, then the slopes from west to east and from south to north may be different. This observation leads to the concept of partial derivatives. If x and y are the coordinates from west to east and from south to north, respectively, then the graph of f is the surface $z = f(x, y)$. At a fixed point $\mathbf{r}_0 = (x_0, y_0)$, the height changes as $h(x) = f(x, y_0)$ along the west–east direction, and as $g(y) = f(x_0, y)$ along the south–north direction. Their graphs are intersections of the surface $z = f(x, y)$ with the coordinate planes $x = x_0$ and $y = y_0$, that is, $z = f(x_0, y) = g(y)$ and $z = f(x, y_0) = h(x)$. The slope along the west–east direction is then $h'(x_0)$, and along the south–north direction, is $g'(y_0)$. These slopes are called *partial derivatives* of f and denoted as

$$\frac{\partial f}{\partial x}(x_0, y_0) = \frac{d}{dx} f(x, y_0)\Big|_{x=x_0},$$
$$\frac{\partial f}{\partial y}(x_0, y_0) = \frac{d}{dy} f(x_0, y)\Big|_{y=y_0}.$$

The partial derivatives are often denoted as

$$\frac{\partial f}{\partial x}(x_0, y_0) = f'_x(x_0, y_0)\,, \quad \frac{\partial f}{\partial y}(x_0, y_0) = f'_y(x_0, y_0).$$

The subscript of f' indicates the variable with respect to which the derivative is calculated. The concept of partial derivatives can easily be extended to functions of more than two variables.

88.1. Partial Derivatives of a Function of Several Variables. Let D be a subset of an n-dimensional Euclidean space.

Definition 13.13. (Interior Point of a Set).
A point $\mathbf{r}_0$ is said to be an interior point of D if there is an open ball $B_\delta(\mathbf{r}_0) = \{\mathbf{r}\,|\,\|\mathbf{r} - \mathbf{r}_0\| < \delta\}$ of radius δ that lies in D (i.e., $B_\delta(\mathbf{r}) \subset D$).

In other words, $\mathbf{r}_0$ is an interior point of D if there is a positive number $\delta > 0$ such that all points whose distance from $\mathbf{r}_0$ is less than δ also lie in D. For example, if D is a set points in a plane whose coordinates are integers, then D has no interior points at all because the points of a disk of radius $0 < a < 1$ centered at any point $\mathbf{r}_0$ of D do not belong to D except $\mathbf{r}_0$. If $D = \{(x, y)\,|\,x^2 + y^2 \leq 1\}$, then any point of D that does not lie on the circle $x^2 + y^2 = 1$ is an interior point.

Definition 13.14. (Open Sets).
A set D in a Euclidean space is said to be open if all points of D are interior points of D.

An open set is an extension of the notion of an open interval (a, b) to the multivariable case. In particular, the whole Euclidean space is open.

Recall that any vector in space may be written as a linear combination of three unit vectors, $\mathbf{r} = (x, y, z) = x\hat{\mathbf{e}}_1 + y\hat{\mathbf{e}}_2 + z\hat{\mathbf{e}}_3$, where $\hat{\mathbf{e}}_1 = (1, 0, 0)$, $\hat{\mathbf{e}}_2 = (0, 1, 0)$, and $\hat{\mathbf{e}}_3 = (0, 0, 1)$. Similarly, using the rules for adding n-tuples and multiplying them by real numbers, one can write

$$\mathbf{r} = (x_1, x_2, ..., x_n) = x_1\hat{\mathbf{e}}_1 + x_2\hat{\mathbf{e}}_2 + \cdots + x_n\hat{\mathbf{e}}_n,$$

where $\hat{\mathbf{e}}_i$ is the n-tuple whose components are zeros except the ith one, which is equal to 1. Obviously, $\|\hat{\mathbf{e}}_i\| = 1$, $i = 1, 2, ..., n$.

Definition 13.15. (Partial Derivatives at a Point).
Let f be a function of several variables $(x_1, x_2, ..., x_n)$. Let D be the

domain of f and let $\mathbf{r}_0$ be an interior point of D. If the limit

$$f'_{x_i}(\mathbf{r}_0) = \lim_{h\to 0} \frac{f(\mathbf{r}_0 + h\hat{\mathbf{e}}_i) - f(\mathbf{r}_0)}{h}$$

exists, then it is called the partial derivative of f with respect to x_i at $\mathbf{r}_0$.

The reason the point $\mathbf{r}_0$ needs to be an interior point is simple. By the definition of the one-variable limit, h can be negative or positive. So the points $\mathbf{r}_0 + h\hat{\mathbf{e}}_i$, $i = 1, 2, ..., n$, must be in the domain of the function because otherwise $f(\mathbf{r}_0 + h\hat{\mathbf{e}}_i)$ is not even defined. This is guaranteed if $\mathbf{r}_0$ is an interior point because all points $\mathbf{r}$ in the ball $B_a(\mathbf{r}_0)$ of sufficiently small radius $a = |h|$ are in D.

Let $\mathbf{r}_0 = (a_1, a_2, ..., a_n)$, where a_i are fixed numbers. Consider the function $F(x_i)$ of one variable x_i (i is fixed), which is obtained from $f(\mathbf{r})$ by fixing all the variables $x_j = a_j$ except the ith one (i.e., $x_j = a_j$ for all $j \neq i$). By the definition of the ordinary derivative, the partial derivative $f'_{x_i}(\mathbf{r}_0)$ exists if and only if the derivative $F'(a_i)$ exists because

$$(13.1) \qquad f'_{x_i}(\mathbf{r}_0) = \lim_{h\to 0} \frac{F(a_i + h) - F(a_i)}{h} = \frac{dF(x_i)}{dx_i}\Big|_{x_i=a_i}$$

just like in the case of two variables discussed at the beginning of this section. This rule is practical for calculating partial derivatives as it reduces the problem to computing ordinary derivatives.

Example 13.12. *Find the partial derivatives of $f(x, y, z) = x^3 - y^2 z$ at the point* $(1, 2, 3)$.

Solution: By the rule (13.1),

$$\begin{aligned} f'_x(1,2,3) &= \frac{d}{dx} f(x,2,3)\Big|_{x=1} = \frac{d}{dx}(x^3 - 12)\Big|_{x=1} = 3, \\ f'_y(1,2,3) &= \frac{d}{dy} f(1,y,3)\Big|_{y=2} = \frac{d}{dy}(1 - 3y^2)\Big|_{y=2} = -12, \\ f'_z(1,2,3) &= \frac{d}{dz} f(1,2,z)\Big|_{z=3} = \frac{d}{dz}(1 - 4z)\Big|_{z=3} = -4. \end{aligned}$$

□

88.1.1. Geometrical Significance of Partial Derivatives. From the rule (13.1), it follows that *the partial derivative $f'_{x_i}(\mathbf{r}_0)$ defines the rate of change of the function f when only the variable x_i changes while the other variables are kept fixed.* If, for instance, the function f in Example 13.12 defines the temperature in degrees Celsius as a function of position whose coordinates are given in meters, then, at the point

$(1, 2, 3)$, the temperature increases at the rate 4 degrees Celsius per meter in the direction of the x axis, and it decreases at the rates -12 and -4 degrees Celsius per meter in the direction of the y and z axes, respectively.

88.2. Partial Derivatives as Functions. Suppose that the partial derivatives of f exist at all points of a set D (which is a subset of the domain of f). Then each partial derivative can be viewed as a function of several variables on D. These functions are denoted as $f'_{x_i}(\mathbf{r})$, where $\mathbf{r} \in D$. They can be found by the same rule (13.1) if, when differentiating with respect to x_i, all other variables are not set to any specific values but rather viewed as independent of x_i (i.e., $dx_j/dx_i = 0$ for all $j \neq i$). This agreement is reflected by the notation

$$f'_{x_i}(x_1, x_2, ..., x_n) = \frac{\partial}{\partial x_i} f(x_1, x_2, ..., x_n);$$

that is, the symbol $\partial/\partial x_i$ means differentiation with respect to x_i while regarding all other variables as numerical parameters independent of x_i.

Example 13.13. *Find $f'_x(x, y)$ and $f'_y(x, y)$ if $f(x, y) = x\sin(xy)$.*

Solution: Assuming first that y is a numerical parameter independent of x, one obtains

$$\begin{aligned} f'_x(x, y) = \frac{\partial}{\partial x} f(x, y) &= \Big(\frac{\partial}{\partial x} x\Big)\sin(xy) + x\frac{\partial}{\partial x}\sin(xy) \\ &= \sin(xy) + xy\cos(xy) \end{aligned}$$

by the product rule for the derivative. If now the variable x is viewed as a numerical parameter independent of y, one obtains

$$f'_y(x, y) = \frac{\partial}{\partial y} f(x, y) = x\frac{\partial}{\partial y}\sin(xy) = x^2\cos(xy).$$

□

88.3. Basic Rules of Differentiation. Since a partial derivative is just an ordinary derivative with one additional agreement that all other variables are viewed as numerical parameters, the basic rules of differentiation apply to partial derivatives. Let f and g be functions of several variables and let c be a number. Then

$$\frac{\partial}{\partial x_i}(cf) = c\,\frac{\partial f}{\partial x_i}\,, \qquad \frac{\partial}{\partial x_i}(f + g) = \frac{\partial f}{\partial x_i} + \frac{\partial g}{\partial x_i},$$

$$\frac{\partial}{\partial x_i}(fg) = \frac{\partial f}{\partial x_i}\,g + f\,\frac{\partial g}{\partial x_i}\,, \qquad \frac{\partial}{\partial x_i}\Big(\frac{f}{g}\Big) = \frac{\frac{\partial f}{\partial x_i}\,g - f\,\frac{\partial g}{\partial x_i}}{g^2}.$$

Let $h(u)$ be a differentiable function of one variable and let $g(\mathbf{r})$ be a function of several variables whose range lies in the domain of f. Then one can define the composition $f(\mathbf{r}) = h(g(\mathbf{r}))$. Assuming that the partial derivatives of g exist, the chain rule holds

$$\frac{\partial f}{\partial x_i} = h'(g)\,\frac{\partial g}{\partial x_i}. \tag{13.2}$$

Example 13.14. *Find the partial derivatives of the function $f(\mathbf{r}) = \|\mathbf{r}\|^{-1}$, where $\mathbf{r} = (x_1, x_2, ..., x_n)$.*

Solution: Put $h(u) = u^{-1/2}$ and $g(\mathbf{r}) = x_1^2 + x_2^2 + \cdots + x_n^2 = \|\mathbf{r}\|^2$. Then $f(\mathbf{r}) = h(g(\mathbf{r}))$. Since $h'(u) = (-1/2)u^{-3/2}$ and $\partial g/\partial x_i = 2x_i$, the chain rule gives

$$\frac{\partial}{\partial x_i}\,\|\mathbf{r}\|^{-1} = -\frac{x_i}{\|\mathbf{r}\|^3}.$$

□

88.4. Exercises. **(1)** Find the specified partial derivatives of each of the following functions:
(i) $f(x,y) = (x-y)/(x+y)$, $\quad f'_x(1,2)$, $f'_y(1,2)$
(ii) $f(x,y,z) = (xy+z)/(z+y)$, $\quad f'_x(1,2,3)$, $f'_y(1,2,3)$, $f'_z(1,2,3)$
(iii) $f(\mathbf{r}) = (x_1 + 2x_2 + \cdots + nx_n)/(1 + \|\mathbf{r}\|^2)$, $\quad f'_{x_i}(\mathbf{0})$, $i = 1, 2, ..., n$
(iv) $f(x,y,z) = x\sin(yz)$, $\quad f'_x(1,2,\pi/2)$, $f'_y(1,2,\pi/2)$, $f'_z(1,2,\pi/2)$

(2) Find the specified partial derivatives of each of the following functions:
(i) $f(x,y) = (x+y^2)^n$, $\quad f'_x(x,y)$, $f'_y(x,y)$
(ii) $f(x,y) = x^y$, $\quad f'_x(x,y)$, $f'_y(x,y)$
(iii) $f(x,y) = xe^{(x+2y)^2}$, $\quad f'_x(x,y)$, $f'_y(x,y)$
(iv) $f(x,y) = \sin(xy)\cos(x^2+y^2)$, $\quad f'_x(x,y)$, $f'_y(x,y)$
(v) $f(x,y,z) = \ln(x+y^2+z^3)$, $\quad f'_x(x,y,z)$, $f'_y(x,y,z)$, $f'_z(x,y,z)$
(v) $f(x,y,z) = xy^2\cos(z^2x)$, $\quad f'_x(x,y,z)$, $f'_y(x,y,z)$, $f'_z(x,y,z)$
(vi) $f(\mathbf{r}) = (a_1x_1 + a_2x_2 + \cdots + a_nx_n)^m = (\mathbf{a}\cdot\mathbf{r})^m$, $\quad f'_{x_i}(\mathbf{r})$, $i = 1, 2, ..., n$

(3) Determine whether the function $f(x,y)$ increases or decreases when x increases, while y is fixed, and when y increases, while x is fixed at a specified point P_0:
(i) $f(x,y) = xy/(x+y)$, $\quad P_0(1,2)$
(ii) $f(x,y) = (x^2 - 2y^2)^{1/3}$, $\quad P_0(1,1)$
(iii) $f(x,y) = x^2\sin(xy)$, $\quad P_0(-1,\pi)$

89. Higher-Order Partial Derivatives

Since partial derivatives of a function are also functions of several variables, they can be differentiated with respect to any variable. For example, for a function of two variables, all possible second derivatives are

$$\begin{aligned}\frac{\partial f}{\partial x} \quad &\longmapsto \quad \frac{\partial}{\partial x}\frac{\partial f}{\partial x} = \frac{\partial^2 f}{\partial x^2}\,, \quad \frac{\partial}{\partial y}\frac{\partial f}{\partial x} = \frac{\partial^2 f}{\partial y\,\partial x}\,,\\ \frac{\partial f}{\partial y} \quad &\longmapsto \quad \frac{\partial}{\partial x}\frac{\partial f}{\partial y} = \frac{\partial^2 f}{\partial x\,\partial y}\,, \quad \frac{\partial}{\partial y}\frac{\partial f}{\partial y} = \frac{\partial^2 f}{\partial y^2}\,.\end{aligned}$$

Throughout the text, brief notations for higher-order derivatives will also be used. For example,

$$\frac{\partial^2 f}{\partial x^2} = (f'_x)'_x = f''_{xx}\,, \quad \frac{\partial^2 f}{\partial x\,\partial y} = (f'_y)'_x = f''_{yx}$$

and similarly for f''_{yy} and f''_{xy}. Derivatives of the third order are defined as derivatives of second-order derivatives, and so on.

EXAMPLE 13.15. *For the function $f(x, y) = x^4 - x^2y + y^2$, find all second- and third-order derivatives.*

SOLUTION: The first derivatives are $f'_x = 4x^3 - 2xy$ and $f'_y = -x^2 + 2y$. Then the second derivatives are

$$\begin{aligned}f''_{xx} &= (4x^3 - 2xy)'_x = 12x^2 - 2y\,, \quad f''_{yy} = (-x^2 + 2y)'_y = 2,\\ f''_{xy} &= (4x^3 - 2xy)'_y = -2x\,, \quad f''_{yx} = (-x^2 + 2y)'_x = -2x.\end{aligned}$$

The third derivatives are found similarly:

$$\begin{aligned}f'''_{xxx} &= (12x^2 - 2y)'_x = 24x\,, &\quad f'''_{yyy} &= (2)'_y = 0,\\ f'''_{xxy} &= (12x^2 - 2y)'_y = -2\,, &\quad f'''_{xyx} &= f'''_{yxx} = (-2x)'_x = -2,\\ f'''_{yyx} &= (2)'_x = 0\,, &\quad f'''_{yxy} &= f'''_{xyy} = (-2x)'_y = 0.\end{aligned}$$

□

In contrast to the one-variable case, there are higher-order derivatives of a new type that are obtained by differentiating with respect to different variables in different orders, like f''_{xy} and f''_{yx}. In the above example, it has been found that

$$\begin{aligned}f''_{xy} &= f''_{yx},\\ f'''_{xxy} &= f'''_{xyx} = f'''_{yxx},\\ f'''_{xyy} &= f'''_{yyx} = f'''_{yxy};\end{aligned}$$

that is, the result of differentiation is *independent of the order in which the derivatives have been taken*. Is this a peculiarity of the function considered or a general feature of higher-order derivatives? The following theorem answers this question.

THEOREM 13.9. (Clairaut's or Schwarz's Theorem).
Let f be a function of several variables $(x_1, x_2, ..., x_n)$ that is defined on an open ball D in a Euclidean space. If the second derivatives $f''_{x_ix_j}$ and $f''_{x_jx_i}$, where $j \neq i$, are continuous functions on D, then $f''_{x_ix_j} = f''_{x_jx_i}$ at any point of D.

A consequence of Clairaut's theorem can be proved. It asserts that, if higher-order derivatives are continuous functions, then the result of differentiation is independent of the order in which the derivatives have been taken. In many practical applications, it is not necessary to calculate higher-order derivatives in all possible orders to verify the hypothesis of Clairaut's theorem (i.e., the continuity of the derivatives). Derivatives of polynomials are polynomials and hence continuous. Derivatives of basic elementary functions like the sine and cosine and exponential functions are continuous. So compositions of these functions with multivariable polynomials have continuous derivatives of any order. In other words, the continuity of the derivatives can often be established by different, simpler means.

EXAMPLE 13.16. *Find the third derivatives f'''_{xyz}, f'''_{yzx}, f'''_{zxy}, and so on, for all permutations of x, y, and z, if $f(x, y, z) = \sin(x^2 + yz)$.*

SOLUTION: The sine and cosine functions are continuously differentiable as many times as desired. The argument of the sine function is a multivariable polynomial. By the composition rule $(\sin g)'_x = g'_x \cos g$ and similarly for the other derivatives, partial derivatives of any order must be products of polynomials and the sine and cosine functions whose argument is a polynomial. Therefore, they are continuous in the entire space. The hypothesis of Clairaut's theorem is satisfied, and hence all the derivatives in question coincide and are equal to

$$\begin{aligned} f'''_{xyz} &= (f'_x)''_{yz} = (2x\cos(x^2 + yz))''_{yz} = (-2xz\sin(x^2 + yz))'_z \\ &= -2x\sin(x^2 + yz) - 2xyz\cos(x^2 + yz)\,. \end{aligned}$$

□

89.1. Reconstruction of a Function from Its Derivatives. One of the standard problems in calculus is finding a function $f(x)$ if its derivative $f'(x) = F(x)$ is known. A sufficient condition for the existence of a solution is the continuity of $F(x)$. In this case,

$$f'(x) = F(x) \quad \Longrightarrow \quad f(x) = \int F(x)\,dx + c,$$

where c is a constant. A similar problem can be posed for a function of several variables. Given the first partial derivatives

$$f'_{x_i}(\mathbf{r}) = F_i(\mathbf{r})\,, \quad i = 1, 2, ..., n, \tag{13.3}$$

find $f(\mathbf{r})$ if it exists. The existence of such f is a more subtle question in the case of several variables. Suppose partial derivatives $\partial F_i/\partial x_j$ are continuous functions in an open ball. Then taking the derivative $\partial/\partial x_j$ of both sides of (13.3) and applying Clairaut's theorem, one infers that

$$f''_{x_i x_j} = f''_{x_j x_i} \quad \Longrightarrow \quad \frac{\partial F_i}{\partial x_j} = \frac{\partial F_j}{\partial x_i}. \tag{13.4}$$

Thus, the conditions (13.4) on the functions F_i must be fulfilled; otherwise, f satisfying (13.3) does not exist. The conditions (13.4) are called *integrability conditions* for the system of equations (13.3).

Example 13.17. *Suppose that $f'_x(x, y) = 2x + y$ and $f'_y(x, y) = 2y - x$. Does such a function f exist?*

Solution: The first partial derivatives of f, $F_1(x, y) = 2x + y$ and $F_2(x, y) = 2y - x$, are polynomials, and hence their derivatives are continuous in the entire plane. In order for f to exist, the integrability condition $\partial F_1/\partial y = \partial F_2/\partial x$ must hold in the entire plane. This is not so because $\partial F_1/\partial y = 1$, whereas $\partial F_2/\partial x = -1$. Thus, no such f exists. □

Suppose now that the integrability conditions (13.4) are satisfied. How is a solution f to (13.3) to be found? Evidently, one has to calculate an antiderivative of the partial derivative. In the one-variable case, an antiderivative is defined up to an additive constant. This is not so in the multivariable case. For example, let $f'_x(x, y) = 3x^2y$. An antiderivative of f'_x is a function whose *partial* derivative with respect to x is $3x^2y$. It is easy to verify that x^3y satisfy this requirement. It is obtained by integrating $3x^2y$ with respect to x while viewing y as a numerical parameter independent of x. Just like in the one-variable case, one can always add a constant to the integral, $x^3y + c$ and obtain another solution. The key point to observe is that the integration constant may be a function of y! Indeed, $(x^3y + g(y))'_x = 3x^2y$. Thus, the general solution of $f'_x(x, y) = 3x^2y$ is $f(x, y) = x^3y + g(y)$, where $g(y)$ is arbitrary.

If, in addition, the other partial derivative f'_y is given, then an explicit form of $g(y)$ can be found. Put, for example, $f'_y(x, y) = x^3 + 2y$. The integrability conditions are fulfilled: $(f'_x)'y = (3x^2y)'_y = 3x^2$ and $(f'_y)'_x = (x^3 + 2y)'_x = 3x^2$. So a function with the said partial derivatives does exist. The substitution of $f(x, y) = x^3y + g(y)$ into the equation $f'_y = x^3 + 2y$ yields $x^3 + g'(y) = x^3 + 2y$ or $g'(y) = 2y$ and hence $g(y) = y^2 + c$. Note the cancellation of the x^3 term. This is a direct consequence of the fulfilled integrability condition. Had one tried to apply this procedure without checking the integrability conditions, one could have found that, in general, no such $g(y)$ exists. In Example 13.17, the equation $f'_x = 2x + y$ has a general solution $f(x, y) = x^2 + yx + g(y)$. Its substitution into the second equation $f'_y = 2y - x$ yields $x + g'(y) = 2y - x$ or $g'(y) = 2y - 2x$. The derivative of $g(y)$ cannot depend on x and hence no such $g(y)$ exists.

Example 13.18. *Find $f(x, y, z)$ if $f'_x = yz + 2x = F_1$, $f'_y = xz + 3y^2 = F_2$, and $f'_z = xy + 4z^3 = F_3$ or show that it does not exist.*

Solution: The integrability conditions $(F_1)'_y = (F_2)'_x$, $(F_1)'_z = (F_3)'_x$, and $(F_2)'_z = (F_3)'_z$ are satisfied (their verification is left to the reader). So f exists. Taking the antiderivative with respect to x in the first equation, one finds

$$f'_x = yz + 2x \quad \Longrightarrow \quad f(x, y, z) = \int (yz + 2x)\, dx = xyz + x^2 + g(y, z),$$

where $g(y, z)$ is arbitrary. The substitution of f into the second equations yields

$$\begin{aligned} f'_y = xz + 3y^2 &\Longrightarrow xz + g'_y(y, z) = xz + 3y^2 \\ &\Longrightarrow g'_y(y, z) = 3y^2 \\ &\Longrightarrow g(y, z) = \int 3y^2\, dy = y^3 + h(z) \\ &\Longrightarrow f(x, y, z) = xyz + x^2 + y^3 + h(z), \end{aligned}$$

where $h(z)$ is arbitrary. The substitution of f into the third equation yields

$$\begin{aligned} f'_z = xy + 4z^3 &\Longrightarrow xy + h'(z) = xy + 4z^3 \\ &\Longrightarrow h'(z) = 4z^3 \\ &\Longrightarrow h(z) = z^4 + c \\ &\Longrightarrow f(x, y, z) = xyz + x^2 + y^3 + z^4 + c, \end{aligned}$$

where c is a constant. □

The procedure of reconstructing f from its first partial derivatives as well as the integrability conditions (13.4) will be important when discussing *conservative vector fields* and the *potential* of a conservative vector field.

89.2. Partial Differential Equations. The relation between a function of several variables and its partial derivatives (of any order) is called a *partial differential equation*. Partial differential equations are a key tool to study various phenomena in nature. Many fundamental laws of nature can be state in the form of partial differential equations.

89.2.1. Diffusion Equation. Let $n(\mathbf{r}, t)$, where $\mathbf{r} = (x, y, z)$ is the position vector in space and t is time, be a concentration of a substance, say, in air or water or even in a solid. Even if there is no macroscopic motion in the medium, the concentration changes with time due to thermal motion of the molecules. This process is known as *diffusion*. In some simple situations, the rate at which the concentration changes with time at a point is

$$n'_t = D(n''_{xx} + n''_{yy} + n''_{zz}),$$

where the parameter D is the diffusion constant. So the concentration as a function of the spatial position and time must satisfy the above partial differential equation.

89.2.2. Wave Equation. Sound in air is propagating disturbances of the air density. If $u(\mathbf{r}, t)$ is the deviation of the air density from its constant (nondisturbed) value u_0 at the spatial point $\mathbf{r} = (x, y, z)$ and at time t, then it can be shown that small disturbances $u/u_0 \ll 1$ satisfy the *wave equation*:

$$u''_{tt} = c^2(u''_{xx} + u''_{yy} + u''_{zz}),$$

where c is the speed of sound in the air. Light is an electromagnetic wave. Its propagation is also described by the wave equation, where c is the speed of light in vacuum (or in a medium, if light goes through a medium) and u is the amplitude of electric or magnetic fields.

89.2.3. Laplace and Poisson Equations. The equation

$$u_{xx} + u_{yy} + u_{zz} = f,$$

where f is a given non-zero function of position $\mathbf{r} = (x, y, z)$ in space, is called the *Poisson equation*. In the special case when $f = 0$, this equation is known as the *Laplace equation*. The Poisson and Laplace equations are used to determine static electromagnetic fields created by static electric charges and currents.

EXAMPLE 13.19. *Let $h(q)$ be a twice-differentiable function of a variable q. Show that $u(\mathbf{r}, t) = h(ct - \hat{\mathbf{n}} \cdot \mathbf{r})$ is a solution of the wave equation for any fixed unit vector $\hat{\mathbf{n}}$.*

SOLUTION: Let $\hat{\mathbf{n}} = (n_1, n_2, n_3)$, where $n_1^2 + n_2^2 + n_3^2 = 1$ as $\hat{\mathbf{n}}$ is the unit vector. Put $q = ct - \hat{\mathbf{n}} \cdot \mathbf{r} = ct - n_1 x - n_2 y - n_3 z$. By the chain rule (13.2), $u_t' = q_t' h'(q)$ and similarly for the other derivatives, one finds $u_t' = ch'(q)$, $u_{tt}'' = c^2 h''(q)$, $u_x' = -n_1 h'(q)$, $u_{xx}'' = n_1^2 h''(q)$, and, in the same fashion, $u_{yy}'' = n_2^2 h''(q)$, $u_{zz}'' = n_3^2 h''(q)$. Then $u_{xx}'' + u_{yy}'' + u_{zz}'' = (n_1^2 + n_2^2 + n_3^2) h''(q) = h''(q)$, which coincides with u_{tt}''/c^2, meaning that the wave equation is satisfied for any h. □

Consider the level surfaces of the solution of the wave equation discussed in this example. They correspond to a fixed value of $q = q_0$. So, for each moment of time t, the disturbance of the air density $u(\mathbf{r}, t)$ has a constant value $h(q_0)$ in the plane $\hat{\mathbf{n}} \cdot \mathbf{r} = ct - q_0 = d(t)$. All planes with different values of the parameter d are parallel as they have the same normal vector $\hat{\mathbf{n}}$. Since here $d(t)$ is a function of time, the plane on which the air density has a fixed value moves along the vector $\hat{\mathbf{n}}$ at the rate $d'(t) = c$. Thus, a disturbance of the air density propagates with speed c. This is the reason that the constant c in the wave equation is called the *speed of sound*. Evidently, the same line of arguments applies to electromagnetic waves; that is, they move through space at the speed of light. The speed of sound in the air is about 342 meters per second, or about 768 mph. The speed of light is $3 \cdot 10^8$ meters per second, or 186 miles per second. If a lightning strike occurs a mile away during a thunderstorm, it can be seen almost instantaneously, while the thunder will be heard in about 5 seconds later.

89.3. Study Problems.

Problem 13.4. *Find the value of a constant k for which the function*

$$u(\mathbf{r}, t) = t^{-3/2} e^{-kr^2/t}, \quad r = \|\mathbf{r}\|,$$

satisfies the diffusion equation for all $t > 0$.

SOLUTION: Note that u depends on the combination $r^2 = x^2 + y^2 + z^2$. To find the partial derivatives of u, it is convenient to use the chain rule:

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial r^2} \frac{\partial r^2}{\partial x} = 2x \frac{\partial u}{\partial r^2} = -\frac{2kx}{t} u,$$

$$u_{xx}'' = \frac{\partial}{\partial x}\Big(\frac{\partial u}{\partial x}\Big) = -\frac{2k}{t} u - \frac{2kx}{t} \frac{\partial u}{\partial x} = \Big(-\frac{2k}{t} + \frac{4k^2 x^2}{t^2}\Big) u.$$

To obtain u_{yy}'' and u_{zz}'', note that r^2 is symmetric with respect to permutations of x, y, and z. Therefore, u_{yy}'' and u_{zz}'' are obtained from u_{xx}'' by replacing, in the latter, x by y and x by z, respectively. Hence, the right side of the diffusion equation reads

$$D(u''_{xx} + u''_{yy} + u''_{zz}) = \Big(-\frac{6Dk}{t} + \frac{4Dk^2r^2}{t^2}\Big)u.$$

Using the product rule to calculate the time derivative, one finds for the left side

$$u'_t = -\frac{3}{2}t^{-5/2}e^{-kr^2/t} + t^{-3/2}e^{-kr^2/t}\frac{kr^2}{t^2} = \Big(-\frac{3}{2t} + \frac{kr^2}{t^2}\Big)u.$$

Since both sides must be equal for *all* values of $t > 0$ and r^2, the comparison of the last two expressions yields *two* conditions: $6Dk = 3/2$ (as the equality of the coefficients at $1/t$) and $k = 4Dk^2$ (as the equality of the coefficients at r^2/t^2). The only common solution of these conditions is $k = 1/(4D)$. □

Problem 13.5. *Consider the function*

$$f(x,y) = \frac{x^3y - xy^3}{x^2+y^2} \quad \text{if} \quad (x,y) \neq (0,0) \quad \text{and} \quad f(0,0) = 0.$$

Find $f'_x(x,y)$ and $f'_y(x,y)$ for $(x,y) \neq (0,0)$. Use the rule (13.1) to find $f'_x(0,0)$ and $f'_y(0,0)$ and, thereby, to establish that f'_x and f'_y exist everywhere. Use the rule (13.1) again to show that $f''_{xy}(0,0) = -1$ and $f''_{yx}(0,0) = 1$, that is, $f''_{xy}(0,0) \neq f''_{yx}(0,0)$. Does this result contradict Clairaut's theorem?

Solution: Using the ratio rule for differentiation, one finds

$$f'_x(x,y) = \frac{x^4y + 4x^2y^3 - y^5}{(x^2+y^2)^2}\,, \quad f'_y(x,y) = \frac{x^5 - 4x^3y^2 - xy^4}{(x^2+y^2)^2}$$

if $(x,y) \neq (0,0)$. Note that, owing to the symmetry $f(x,y) = -f(y,x)$, the derivative f'_y is obtained from f'_x by changing the sign of the latter and swapping x and y. The derivatives at $(0,0)$ are found by the rule (13.1):

$$f'_x(0,0) = \frac{d}{dx}f(x,0)\Big|_{x=0} = 0\,, \quad f'_y(0,0) = \frac{d}{dy}f(0,y)\Big|_{y=0} = 0.$$

The derivatives are continuous functions (the proof is left to the reader as an exercise). Next, one has

$$\begin{aligned} f''_{xy}(0,0) &= \frac{d}{dy}f'_x(0,y)\Big|_{y=0} = \lim_{h\to 0}\frac{f'_x(0,h) - f'_x(0,0)}{h} \\ &= \lim_{h\to 0}\frac{-h-0}{h} = -1, \end{aligned}$$

$$\begin{aligned} f''_{yx}(0,0) &= \frac{d}{dx} f'_y(x,0)\Big|_{x=0} = \lim_{h\to 0} \frac{f'_y(h,0) - f'_y(0,0)}{h} \\ &= \lim_{h\to 0} \frac{h-0}{h} = 1. \end{aligned}$$

The result does not contradict Clairaut's theorem because $f''_{xy}(x,y)$ and $f''_{yx}(x,y)$ are not continuous at $(0,0)$. By using the ratio rule to differentiate $f'_x(x,y)$ with respect to y, an explicit form of $f''_{xy}(x,y)$ for $(x,y) \neq (0,0)$ can be obtained. By taking the limit of $f''_{xy}(x,y)$ as $(x,y) \to (0,0)$ along the straight line $(x,y) = (t, at)$, $t \to 0$, one infers that the limit depends on the slope a and hence the two-dimensional limit does not exist, that is, $\lim_{(x,y)\to(0,0)} f''_{xy}(x,y) \neq f''_{xy}(0,0) = -1$ and f''_{xy} is not continuous at $(0,0)$. The technical details are left to the reader. □

89.4. Exercises. **(1)** Find all second partial derivatives of each of the following functions and verify Clairaut's theorem:
(i) $f(x,y) = \tan^{-1} xy$
(ii) $f(x,y,z) = x\sin(zy^2)$
(iii) $f(x,y,z) = x^3 + zy + z^2$
(iv) $f(x,y,z) = (x+y)/(x+2z)$

(2) Find the indicated partial derivatives of each of the following functions:
(i) $f(x,y) = x^n + xy + y^m$, f'''_{xxy}, f'''_{xyx}, f'''_{yyx}, f'''_{xyy}
(ii) $f(x,y,z) = x\cos(yx) + z^3$, f'''_{xyz}, f'''_{xxz}, f'''_{yyz}
(iii) $f(x,y,z) = \sin(xy)e^z$, $\partial f^5/\partial z^5$, $f^{(4)}_{xyzz}$, $f^{(4)}_{zyxz}$, $f^{(4)}_{zxzy}$

(3) Given partial derivatives, find the function or show that it does not exist:
(i) $f'_x = 3x^2 y$, $f'_y = x^3 + 3y^2$
(ii) $f'_x = yz + 3x^2$, $f'_y = xz + 4y$, $f'_z = xy + 1$
(iii) $f'_{x_k} = kx_k$, $k = 1, 2, ..., n$
(iv) $f'_x = xy + z$, $f'_y = x^2/2$, $f'_z = x + y$
(v) $f'_x = \sin(xy) + xy\cos(xy)$, $f'_y = x^2\cos(xy) + 1$

(4) Verify that a given function is a solution of the indicated differential equation:
(i) $f(t,x) = A\sin(ct - x) + B\cos(ct + x)$, $c^{-2} f''_{tt} - f''_{xx} = 0$
(ii) $f(x,y) = \ln(x^2 + y^2)$, $f''_{xx} + f''_{yy} = 0$
(iii) $f(x,y) = \ln(e^x + e^y)$, $f'_x + f'_y = 1$ and $f''_{xx} f''_{yy} - (f''_{xy})^2 = 0$
(iv) $f(\mathbf{r}) = \exp(\mathbf{a}\cdot\mathbf{r})$, where $\mathbf{a}\cdot\mathbf{a} = 1$, $f''_{x_1x_1} + f''_{x_2x_2} + \cdots + f''_{x_nx_n} = f$

90. Chain Rules and Implicit Differentiation

90.1. Chain Rules. Consider the function $f(x,y) = x^3 + xy^2$ whose domain is the entire plane. Points of the plane can be labeled in a different way. For example, the polar coordinates $x = r\cos\theta$, $y = r\sin\theta$ may be viewed as a rule that assigns an ordered pair (x,y) to an ordered pair (r,θ). Using this rule, the function can be expressed in the new variables as $f(r\cos\theta, r\sin\theta) = r^3\sin\theta = F(r,\theta)$. One can compute the rates of change of f with respect to the new variables:

$$\frac{\partial f}{\partial r} = \frac{\partial F}{\partial r} = 3r^2\sin\theta\,, \quad \frac{\partial f}{\partial \theta} = \frac{\partial F}{\partial \theta} = r^3\cos\theta.$$

Alternatively, these rates can be computed as

$$\begin{aligned}\frac{\partial f}{\partial r} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial r} = (3x^2 + y^2)\cos\theta + 2xy\sin\theta = 3r^2\sin\theta,\\ \frac{\partial f}{\partial \theta} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial \theta} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial \theta} = -(3x^2 + y^2)r\sin\theta + 2xyr\cos\theta = r^3\cos\theta,\end{aligned}$$

where x and y have been expressed in the polar coordinates to obtain the final expressions. The difference between the two approaches is that in the second one *an explicit form of the function in the new variables is not required to find its rates with respect to the new variables.*

Furthermore, consider the values of this function along the curve $x = t$, $y = t^2$, $f(t,t^2) = t^3 + t^5 = F(t)$. The rate of change of f with respect to the curve parameter is

$$\frac{df}{dt} = \frac{dF}{dt} = 3t^2 + 5t^4\,.$$

It can also be obtained without calculating first the explicit form of the function f as a function of the curve parameter in much the same fashion as in the case of polar coordinates:

$$\frac{df}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} = (3x^2 + y^2) + (2xy)(2t) = 3t^2 + 5t^4\,.$$

Qualitatively, this expression of the rate df/dt is rather natural. If only x depends on t while y does not (i.e., $y = \text{const}$), then the rate df/dt is determined by an ordinary chain rule $df/dt = f'_x x'(t)$; the y variable is merely a numerical parameter. If $y = y(t)$ and $x = \text{const}$, the chain rule for one-variable functions gives $df/dt = f'_y y'(t)$. When both x and y depend on t, then df/dt becomes the sum of these two terms as both rates $x'(t)$ and $y'(t)$ should contribute to df/dt.

The examples considered above illustrate a general rule of differentiation called the *chain rule.*

THEOREM 13.10. (Chain Rule).
Let f be a function of n variables $\mathbf{r} = (x_1, x_2, ..., x_n)$ such that all its partial derivatives exist. Suppose that each variable x_i is, in turn, a function of m variables $\mathbf{u} = (u_1, u_2, ..., u_m)$ such that all its partial derivatives exist. The composition of $x_i = x_i(\mathbf{u})$ with $f(\mathbf{r})$ defines f as a function of $\mathbf{u}$. Then its rate of change with respect to u_j, $j = 1, 2, ..., m$, reads

$$\frac{\partial f}{\partial u_j} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial u_j} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial u_j} + \cdots + \frac{\partial f}{\partial x_n}\frac{\partial x_n}{\partial u_j} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\frac{\partial x_i}{\partial u_j}.$$

The proof of this theorem is rather technical and is omitted.

For $n = m = 1$, this is the familiar chain rule for functions of one variable $df/du = f'(x)x'(u)$. If $n = 1$ and $m > 1$, it is the chain rule (13.2) established earlier. The example of polar coordinates corresponds to the case $n = m = 2$, where $\mathbf{r} = (x, y)$ and $\mathbf{u} = (r, \theta)$. The rate of change along a curve in a plane is the case $n = 2$, $m = 1$, where $\mathbf{r} = (x, y)$ and $u = t$. Note also that some of the functions $x_i(\mathbf{u})$ may not depend on all variables u_j, and the corresponding partial derivatives in the chain rule vanish.

EXAMPLE 13.20. *A function $f(x, y, z)$ has the following rates of change at the point $\mathbf{r}_0 = (1, 2, 3)$, $f'_x(\mathbf{r}_0) = 1$, $f'_y(\mathbf{r}_0) = 2$, and $f'_z(\mathbf{r}_0) = -2$. Suppose that $x = x(t, s) = t^2 s$, $y = y(t, s) = s + t$, and $z = z(t, s) = 3s$. Find the rates of change f with respect to t and s at the point $\mathbf{r}_0$.*

SOLUTION: In the chain rule, put $\mathbf{r} = (x, y, z)$ and $\mathbf{u} = (t, s)$. The point $\mathbf{r}_0 = (1, 2, 3)$ corresponds to the point $\mathbf{u}_0 = (1, 1)$ in the new variables. Note that $z = 3$ gives $3s = 3$ and hence $s = 1$. Then, from $y = 2$, it follows that $s + t = 2$ or $1 + t = 2$ or $t = 1$. Also, $x(1, 1) = 1$ as required. The partial derivatives of the old variables with respect to the new ones are $x'_t = 2ts$, $y'_t = 1$, $z'_t = 0$, $x'_s = t^2$, $y'_s = 1$, and $z'_s = 3$. By the chain rule,

$$\begin{aligned} f'_t(\mathbf{r}_0) &= f'_x(\mathbf{r}_0)x'_t(\mathbf{u}_0) + f'_y(\mathbf{r}_0)y'_t(\mathbf{u}_0) + f'_z(\mathbf{r}_0)z'_t(\mathbf{u}_0) \\ &= 2 + 2 + 0 = 4, \\ f'_s(\mathbf{r}_0) &= f'_x(\mathbf{r}_0)x'_s(\mathbf{u}_0) + f'_y(\mathbf{r}_0)y'_s(\mathbf{u}_0) + f'_z(\mathbf{r}_0)z'_s(\mathbf{u}_0) \\ &= 1 + 2 - 6 = -3. \end{aligned}$$

□

90.2. Implicit Differentiation. Consider the function of three variables, $F(x, y, z) = x^2 + y^4 - z$. The equation $F(x, y, z) = 0$ can be solved for one of the variables, say, z to obtain z as a function of two variables:

$$F(x,y,z)=0 \quad \Longrightarrow \quad z=z(x,y)=x^2+y^4;$$

that is, the function $z(x,y)$ is defined as a root of $F(x,y,z)$ and has the characteristic property that

$$F(x,y,z(x,y))=0 \quad \text{for all } (x,y). \tag{13.5}$$

In the example considered, the equation $F(x,y,z)=0$ can be solved analytically, and an *explicit* form of its root as a function of (x,y) can be found.

In general, given a function $F(x,y,z)$, an explicit form of a solution to the equation $F(x,y,z)=0$ is not always possible to find. Putting aside the question about the very existence of a solution and its uniqueness, suppose that this equation is proved to have a unique solution when $(x,y)\in D$. In this case, the function $z(x,y)$ with the property (13.5) for all $(x,y)\in D$ is said to be defined *implicitly* on D.

Although an analytic form of an implicitly defined function is unknown, its rates of change can be found and provide important information about its local behavior. Suppose that partial derivatives of F exist. Furthermore, the rates $z'_x(x,y)$ and $z'_y(x,y)$ are also assumed to exist on an open disk D in the plane. Since relation (13.5) holds for all $(x,y)\in D$, the partial derivatives of its left side must also vanish in D. The derivatives can be computed by the chain rule, $n=3$, $m=2$, $\mathbf{r}=(x,y,z)$, and $\mathbf{u}=(u,v)$, where the relations between old and new variables are $x=u$, $y=v$, and $z=z(u,v)$. One has

$$\frac{\partial}{\partial u}F(x,y,z(x,y))=\frac{\partial F}{\partial x}+\frac{\partial F}{\partial z}\frac{\partial z}{\partial x}=0 \quad \Longrightarrow \quad z'_x=-\frac{F'_x}{F'_z},$$

$$\frac{\partial}{\partial v}F(x,y,z(x,y))=\frac{\partial F}{\partial y}+\frac{\partial F}{\partial z}\frac{\partial z}{\partial y}=0 \quad \Longrightarrow \quad z'_y=-\frac{F'_y}{F'_z},$$

where $z'_u(u,v)=z'_x(x,y)$ and $z'_v(u,v)=z'_y(x,y)$ because $x=u$ and $y=v$. These equations determine the rates of change of an implicitly defined function of two variables. Note that in order for these equations to make sense, the condition $F'_z\neq 0$ must be imposed. Several questions about the very existence and uniqueness of $z(x,y)$ for a given $F(x,y,z)$ and the existence of derivatives of $z(x,y)$ have been left unanswered in the above analysis. The following theorem addresses them all.

Theorem 13.11. (Implicit Function Theorem).
Let F be a function of $n+1$ variables, $F(\mathbf{r},z)$, where $\mathbf{r}=(x_1,x_2,...,x_n)$ and z is real, such that all its partial derivatives are continuous in an

open ball B. Suppose that there exists a point $(\mathbf{r}_0, z_0) \in B$ such that $F(\mathbf{r}_0, z_0) = 0$ and $F'_z(\mathbf{r}_0, z_0) \neq 0$. Then there exist a neighborhood D of $\mathbf{r}_0$ and a unique function $z = z(\mathbf{r})$ with continuous derivatives on D such that

$$F(\mathbf{r}, z(\mathbf{r})) = 0 \quad \text{and} \quad z'_{x_i}(\mathbf{r}) = -\frac{F'_{x_i}(\mathbf{r}, z(\mathbf{r}))}{F'_z(\mathbf{r}, z(\mathbf{r}))}.$$

for all $\mathbf{r}$ in D.

The proof of this theorem goes beyond the scope of this course. It includes proofs of the existence and uniqueness of $z(\mathbf{r})$ and the existence of its derivatives. Once these facts are established, a derivation of the implicit differentiation formula follows the same way as in the $n = 2$ case:

$$\frac{\partial F}{\partial x_i} + \frac{\partial F}{\partial z}\frac{\partial z}{\partial x_i} = 0 \quad \Longrightarrow \quad z'_{x_i}(\mathbf{r}) = -\frac{F'_{x_i}(\mathbf{r}, z(\mathbf{r}))}{F'_z(\mathbf{r}, z(\mathbf{r}))}.$$

EXAMPLE 13.21. *Show that the equation $z(3x - y) = \pi \sin(xyz)$ has a unique solution $z = z(x, y)$ in a neighborhood of $(1, 1)$ such that $z(1, 1) = \pi/2$ and find the rates of change $z'_x(1, 1)$ and $z'_y(1, 1)$.*

SOLUTION: Put $F(x, y, z) = \pi \sin(xyz) - z(3x - y)$. Then the existence and uniqueness of the solution can be established by verifying the hypotheses of the implicit function theorem in which $\mathbf{r} = (x, y)$, $\mathbf{r}_0 = (1, 1)$, and $z_0 = \pi/2$. First, note that the function F is the sum of a polynomial and the sine function of a polynomial. So its derivatives

$$\begin{aligned} F'_x &= \pi yz \cos(xyz) - 3x\,, \quad F'_y = \pi xz \cos(xyz) + z\,, \\ F'_z &= \pi xy \cos(xyz) - 3x + y \end{aligned}$$

are continuous for all (x, y, z). Next, $F(1, 1, \pi/2) = 0$ as required. Finally, $F'_z(1, 1, \pi/2) = -2 \neq 0$. Therefore, by the implicit function theorem, there is an open disk in the xy plane containing the point $(1, 1)$ in which the equation has a unique solution $z = z(x, y)$. By the implicit differentiation formulas,

$$z'_x(1, 1) = -\frac{F'_x(1, 1, \pi/2)}{F'_z(1, 1, \pi/2)} = -\frac{3\pi}{4}\,, \qquad z'_y(1, 1) = -\frac{F'_y(1, 1, \pi/2)}{F'_z(1, 1, \pi/2)} = \frac{\pi}{4}.$$

In particular, this result implies that, near the point $(1, 1)$, the root $z(x, y)$ decreases in the direction of the x axis and increases in the direction of the y axis. It should be noted that the numerical values of the derivatives can be used to accurately approximate the root $z(x, y)$ of a nonlinear equation in a neighborhood of $(1, 1)$ by invoking the concept of *linearization* discussed below (see also Study Problem 13.6). □

90.3. Exercises. **(1)** Use the chain rule to find dz/dt if $z = \sqrt{1+x^2+2y^2}$ and $x = 2t^3$, $y = \ln t$.

(2) Use the chain rule to find $\partial z/\partial s$ and $\partial z/\partial t$ if $z = e^{-x}\sin(xy)$ and $x = ts$, $y = \sqrt{s^2+t^2}$.

(3) Use the chain rule to write the partial derivatives of F with respect to the new variables:
(i) $F = f(x,y)$, $x = x(u,v,w)$, $y = y(u,v,w)$
(ii) $F = f(x,y,z,t)$, $x = x(u,v)$, $y = y(u,v)$, $z = z(w,s)$, $t = t(w,s)$

(4) Find the rates of change $\partial z/\partial u$, $\partial z/\partial v$, $\partial z/\partial w$ when $(u,v,w) = (2,1,1)$ if $z = x^2+yx+y^3$ and $x = uv^2+w^3$, $y = u+v\ln w$.

(5) Find the rates of change $\partial f/\partial u$, $\partial f/\partial v$, $\partial f/\partial w$ when $(x,y,z) = (1/3,2,0)$ if $x = 2/u - v + w$, $y = vuw$, $z = e^w$.

(6) Find the partial derivatives of $z = f(x,y)$ defined implicitly by the equation $x - z = \tan^1(yz)$.

(7) Let the temperature of the air at a point (x,y,z) be $T(x,y,z)$ degrees Celsius. An insect flies through the air so that its position as a function of time t in seconds is given by $x = \sqrt{1+t}$, $y = 2t$, $z = t^2-1$. If $T'_x(2,6,8) = 2$, $T'_y(2,6,8) = -1$, and $T'_x(2,6,8) = 1$, how fast is the temperature rising (or decreasing) on the insect's path as it flies through the point $(2,6,8)$?

(8) Let a rectangular box have the dimensions x, y, and z that change with time. Suppose that at a certain instant the dimensions are $x = 1$ m, $y = z = 2$ m, and x and y are increasing at the rate 2 m/s and z is decreasing at the rate 3 m/s. At that instance, find the rates at which the volume, the surface area, and the largest diagonal are changing.

(9) A function is said to be homogeneous of degree n if, for any number t, it has the property $f(tx,ty) = t^n f(x,y)$. Give an example of a polynomial function that is homogeneous of degree n. Show that a homogeneous differentiable function satisfies the equation $xf'_x + yf'_y = nf$. Show also that $f'_x(tx,ty) = t^{n-1}f(x,y)$.

(10) Suppose that the equation $F(x,y,z) = 0$ defines implicitly $z = f(x,y)$, or $y = g(x,z)$, or $x = h(y,z)$. Assuming that the derivatives F'_x, F'_y, and F'_z do not vanish, prove that $(\partial z/\partial x)(\partial x/\partial y)(\partial y/\partial z) = -1$.

91. Linearization of Multivariable Functions

A differentiable one-variable function $f(x)$ can be approximated near $x = x_0$ by its linearization $L(x) = f(x_0) + f'(x_0)(x - x_0)$ or the tangent line. If $x = x_0 + \Delta x$, then

$$\frac{f(x) - L(x)}{\Delta x} = \frac{f(x_0+\Delta x) - f(x_0)}{\Delta x} - f'(x_0) \to 0 \quad \text{as} \quad \Delta x \to 0$$

by the definition of the derivative $f'(x_0)$. This relation implies that the error of the linear approximation goes to 0 faster than the deviation $\Delta x = x - x_0$ of x from x_0, that is,

$$f(x) = L(x) + \varepsilon(\Delta x)\,\Delta x, \quad \text{where} \quad \varepsilon(\Delta x) \to 0 \quad \text{as} \quad \Delta x \to 0\,.$$

For example, if $f(x) = x^2$, then its linearization at $x = 1$ is $L(x) = 1 + 2(x-1)$. It follows that $f(1+\Delta x) - L(1+\Delta x) = \Delta x^2$ or $\varepsilon(\Delta x) = \Delta x$. In the interval $x \in (0.9, 1.1)$, the absolute error of the linear approximation is less than 0.01. So the linearization of f at x_0 provides a good approximation of f in a sufficiently small neighborhood of x_0. Naturally, such a useful tool needs to be extended to multivariable functions.

91.1. Tangent Plane Approximation. Consider first the case of two-variable functions. The graph of $f(x,y)$ is the surface $z = f(x,y)$. Consider the curve of intersection of this surface with the coordinate plane $x = x_0$. Its equation is $z = f(x_0, y)$. The vector function $\mathbf{r}(t) = (x_0, t, f(x_0,t))$ traces out the curve of intersection. The curve goes through the point $\mathbf{r}_0 = (x_0, y_0, z_0)$, where $z_0 = f(x_0,y_0)$, because $\mathbf{r}(y_0) = \mathbf{r}_0$. Its tangent vector at the point $\mathbf{r}_0$ is $\mathbf{v}_1 = \mathbf{r}'(y_0) = (0, 1, f'_y(x_0,y_0))$. The line parallel to $\mathbf{v}_1$ through the point $\mathbf{r}_0$ lies in the plane $x = x_0$ and is tangent to the intersection curve $z = f(x_0, y)$. Similarly, the graph $z = f(x,y)$ intersects the coordinate plane $y = y_0$ along the curve $z = f(x, y_0)$ whose parametric equations are $\mathbf{r}(t) = (t, y_0, f(t, y_0))$. The tangent vector to this curve at the point $\mathbf{r}_0$ is $\mathbf{v}_2 = \mathbf{r}'(x_0) = (1, 0, f'_x(x_0, y_0))$. The line parallel to $\mathbf{v}_2$ through $\mathbf{r}_0$ lies in the plane $y = y_0$ and is tangent to the curve $z = f(x, y_0)$.

Now one can define a plane through the point $\mathbf{r}_0$ of the graph that contains the two tangent lines. This plane is called the *tangent plane* to the graph. Its normal must be perpendicular to both vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ and, by the geometrical properties of the cross product, may be taken as $\mathbf{n} = \mathbf{v}_1 \times \mathbf{v}_2 = (f'_x(x_0,y_0), f'_y(x_0,y_0), -1)$. The standard equation of the plane $\mathbf{n} \cdot (\mathbf{r} - \mathbf{r}_0) = 0$ can then be written in the form

$$z = z_0 + n_1(x - x_0) + n_2(y - y_0)\,, \quad n_1 = f'_x(x_0,y_0), \quad n_2 = f'_y(x_0,y_0).$$

The graph goes through the point $\mathbf{r}_0$ and so does the tangent plane. So the tangent plane is expected to be close to the graph in a neighborhood of $\mathbf{r}_0$. The exact meaning of "close" will be clarified in the following section.

EXAMPLE 13.22. *Find the tangent plane to the paraboloid* $z = x^2 + 3y^2$ *at the point* $(2, 1, 7)$.

Solution: The paraboloid is the graph of the function $f(x, y) = x^2+3y^2$. The components of the normal are $n_1 = f'_x(2, 1) = 2x|_{(2,1)} = 4$, $n_2 = f'_y(2, 1) = 6y|_{(2,1)} = 6$, and $n_3 = -1$. An equation of the tangent plane is $4(x - 2) + 6(y - 1) - (z - 7) = 0$ or $4x + 6y - z = 7$. □

By analogy with the one-variable case, one can define the *linearization* of $f(x, y)$ at a point (x_0, y_0) in the domain of f by the linear function

$$L(x, y) = f(x_0, y_0) + f'_x(x_0, y_0)(x - x_0) + f'_y(x_0, y_0)(y - y_0).$$

The approximation $f(x, y) \approx L(x, y)$ is called a *linear approximation* of f near (x_0, y_0). By analogy, the concepts of linearization and the linear approximation are extended to functions of more than two variables.

Definition 13.16. (Linearization of a Multivariable Function). *Let f be a function of m variables $\mathbf{r} = (x_1, x_2, ..., x_m)$ on D such that its derivatives exist at an interior point $\mathbf{r}_0 = (a_1, a_2, ..., a_m)$ of D. Put $n_i = f'_{x_i}(\mathbf{r}_0)$, $i = 1, 2, ..., m$. The function*

$$L(\mathbf{r}) = f(\mathbf{r}_0) + n_1(x_1 - a_1) + n_2(x_2 - a_2) + \cdots + n_m(x_m - a_m)$$

is called the linearization *of f at $\mathbf{r}_0$, and the approximation $f(\mathbf{r}) \approx L(\mathbf{r})$ is called the* linear approximation *of f near $\mathbf{r}_0$.*

It is convenient to write the linearization in a more compact form

$$L(\mathbf{r}) = f(\mathbf{r}_0) + n_1\Delta x_1 + n_2\Delta x_2 + \cdots + n_m\Delta x_m\,, \quad n_i = f'_{x_i}(\mathbf{r}_0)\,, \tag{13.6}$$

where Δx_i is the deviation of x_i from a_i.

Example 13.23. *Use the linear approximation to estimate the number* $[(2.03)^2 + (1.97)^2 + (0.94)^2]^{1/2}$.

Solution: Consider the function of three variables $f(x, y, z) = [x^2 + y^2 + z^2]^{1/2}$. The number in question is the value of this function at $(x, y, z) = (2.03, 1.97, 0.94)$. This point is close to $\mathbf{r}_0 = (2, 2, 1)$ at which $f(\mathbf{r}_0) = 3$. The deviations are $\Delta x = x - 2 = 0.03$, $\Delta y = y - 2 = -0.03$, and $\Delta z = 0.94 - 1 = -0.06$. The partial derivatives are $f'_x = x/(x^2 + y^2 + z^2)^{1/2}$, $f'_y = y/(x^2 + y^2 + z^2)^{1/2}$, and $f'_z = z/(x^2 + y^2 + z^2)^{1/2}$. Therefore, $n_1 = 2/3$, $n_2 = 2/3$, and $n_3 = 1/3$. The linear approximation gives

$$f(x, y, z) \approx L(x, y, z) = 3 + (2/3)\,\Delta x + (2/3)\,\Delta y + (1/3)\,\Delta z = 2.98.$$

□

91.2. Differentiability of Multivariable Functions. The concepts of linearization and the linear approximation have been formally extended from the one-variable case in which the very existence of the derivative at a point is sufficient for the linear approximation to be good, as argued at the beginning of this section. In the case of two-variable functions, the existence of partial derivatives at a point allows one to define the tangent plane to the graph. Based on this geometrical observation that the plane $z = L(x, y)$ is tangent to the graph $z = f(x, y)$, it has been assumed that the difference $f(x, y) - L(x, y)$ should decrease with the decreasing distance $(\Delta x^2 + \Delta y^2)^{1/2}$ between the points (x, y) and (x_0, y_0). The key difference between one-variable and multivariable cases is that the mere existence of partial derivatives is *not sufficient* to make the linear approximation a good one.

This can be illustrated by the following example. Put

$$f(x,y) = \begin{cases} \frac{xy}{x^2+y^2} & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } (x,y) = (0,0) \end{cases} . \tag{13.7}$$

This function has the property that $f(x, 0) = 0$ for all x, which implies that its rate along the x axis vanishes, $f'_x(x, 0) = 0$. Similarly, the function vanishes on the y axis, $f(0, y) = 0$, and hence has no rate of change along it, $f'_y(0, y) = 0$. In particular, the partial derivatives exist at the origin, $f'_x(0, 0) = f_y(0, 0) = 0$. So the tangent plane should coincide with the xy plane, $z = 0$, and the linearization is a constant (zero) function, $L(x, y) = 0$. Can one say that $f(x, y) - L(x, y) = f(x, y) \to 0$ as $(x, y) \to (0, 0)$, that is, that the linear approximation is a good one? To answer this question, the two-variable limit must be studied. Consider this limit along the straight line $(x, y) = (t, at)$, $t \to 0^+$. One has $f(t, at) = a/(1 + a^2) \neq 0$. Thus, the limit does not exist. In particular, the difference $f - L$ remains a non zero constant as the argument approaches the origin along a straight line. Furthermore, *the function is not even continuous at* $(0, 0)$ *despite that the partial derivatives exist at* $(0, 0)$! This is quite a departure from the one-variable case where the existence of the derivative implies that the function is necessarily continuous. What has to be changed in the multivariable case in order to achieve a similarity with the one-variable case? The question is answered with the concept of *differentiability* of a function of several variables.

Definition 13.17. (Differentiable Functions).
The function f of several variables $\mathbf{r} = (x_1, x_2, ..., x_m)$ on an open set D is said to be differentiable at a point $\mathbf{r}_0 \in D$ if there exists a linear

function $L(\mathbf{r})$ such that

$$\lim_{\mathbf{r}\to\mathbf{r}_0} \frac{f(\mathbf{r}) - L(\mathbf{r})}{\|\mathbf{r} - \mathbf{r}_0\|} = 0. \tag{13.8}$$

If f is differentiable at all points of D, then f is said to be differentiable on D.

This definition demands that $f(\mathbf{r})-L(\mathbf{r}) = \varepsilon(\mathbf{r})\|\mathbf{r}-\mathbf{r}_0\|$ and $\varepsilon(\mathbf{r}) \to 0$ as $\mathbf{r} \to \mathbf{r}_0$; that is, the error of the linear approximation decreases faster than the distance $\|\mathbf{r} - \mathbf{r}_0\|$ between $\mathbf{r}$ and $\mathbf{r}_0$ just like in the one-variable case. The function (13.7) is not differentiable at $(0, 0)$ despite the existence of its partial derivatives because it does not have a good linear approximation at $(0, 0)$ in the sense of (13.8). In general, one can prove the following result.

Theorem 13.12. (Existence of Partial Derivatives).
If f is differentiable at a point $\mathbf{r}_0$, then its partial derivatives exist at $\mathbf{r}_0$.

The converse is not true! This is the reason the linear approximation may be bad despite the existence of partial derivatives. With some additional assumptions about the partial derivatives, the converse statement can be established. It provides a useful criterion for differentiability.

Theorem 13.13. (Differentiability and Partial Derivatives).
Let f be a function on an open set D of a Euclidean space. Then f is differentiable on D if and only if its partial derivatives exist and are continuous functions on D.

Thus, if, in addition to their existence, the partial derivatives happen to be continuous functions, then the linear approximation is always a good one in the sense of (13.8). Conversely, the partial derivatives of a differentiable function are continuous functions. It is straightforward to verify that $f'_x(x, y)$ and $f'_y(x, y)$ for the function (13.7) are not continuous at $(0, 0)$.

91.3. Study Problems.

Problem 13.6. *Show that the equation $z(3x - y) = \pi \sin(xyz)$ has a unique solution $z = z(x, y)$ in a neighborhood of $(1, 1)$ such that $z(1, 1) = \pi/2$. Estimate $z(1.04, 0.96)$.*

Solution: In Example 13.21, the existence and uniqueness of $z(x, y)$ has been established by the implicit function theorem. The partial derivatives have also been evaluated, $z'_x(1,1) = -3\pi/4$ and $z'_y(1,1) = \pi/4$. The linearization of $z(x,y)$ near $(1,1)$ is

$$z(1+\Delta x, 1+\Delta y) \approx z(1,1) + z'_x(1,1)\,\Delta x + z'_y(1,1)$$

$$\Delta y = \frac{\pi}{2}\Big(1 - \frac{3\Delta x}{2} + \frac{\Delta y}{2}\Big).$$

Putting $\Delta x = 0.04$ and $\Delta y = -0.04$, this equation yields the estimate $z(1.04,\, 0.96) \approx 0.45\pi$. Note that the combination of the implicit differentiation and linearization allows one to approximate the root of a nonlinear equation in a small neighborhood of the point where the value of the root is known. This is an extremely useful concept in many practical applications. □

92. The Differential and Taylor Polynomials

Just like in the one-variable case, given variables $\mathbf{r} = (x_1, x_2, ..., x_m)$, one can introduce independent variables $d\mathbf{r} = (dx_1, dx_2, ..., dx_m)$ that are infinitesimal variations of $\mathbf{r}$ and also called *differentials* of $\mathbf{r}$. The word "infinitesimal" means here that powers $(dx_i)^k$ can always be neglected for $k > 1$.

DEFINITION 13.18. (Differential).
Let $f(\mathbf{r})$ be a differentiable function. The function

$$df(\mathbf{r}) = f'_{x_1}(\mathbf{r})\,dx_1 + f'_{x_2}(\mathbf{r})\,dx_2 + \cdots + f'_{x_m}(\mathbf{r})\,dx_m$$

is called the differential *of f.*

Note that the differential is a function of $2m$ *independent* variables $\mathbf{r}$ and $d\mathbf{r}$. The geometrical significance of the differential follows from its relation with the linearization of f at a point $\mathbf{r}_0$:

$$L(\mathbf{r}) = f(\mathbf{r}_0) + df(\mathbf{r}_0)\,, \quad dx_i = \Delta x_i\,, \quad i = 1, 2, ..., m;$$

that is, if the infinitesimal variations (or differentials) $d\mathbf{r}$ are set to be the deviations $\Delta\mathbf{r} = \mathbf{r} - \mathbf{r}_0$ of the variables $\mathbf{r}$ from $\mathbf{r}_0$, then *the differential df at the point $\mathbf{r}_0$ defines the linearization of f at $\mathbf{r}_0$*. This linearization is a good one in the sense (13.8). From an algebraic point of view, the differential determines variations of values of f under infinitesimal independent variations of its arguments such that contributions of powers of the variations $(dx_i)^k$, $k > 1$, can be neglected.

92.1. Error Analysis. The volume of a rectangle with dimensions x, y, and z is the function of three variables $V(x,y,z) = xyz$. In practice, measurements of the dimensions always contain errors; that is, repetitive measurements give the values of x, y, and z from the intervals

$x \in [x_0 - \Delta x, x_0 + \Delta x]$, $y \in [y_0 - \Delta y, y_0 + \Delta y]$, and $z \in [z_0 - \Delta z, z_0 + \Delta z]$, where $\mathbf{r}_0 = (x_0, y_0, z_0)$ are the mean values of the dimensions, while $\Delta\mathbf{r} = (\Delta x, \Delta y, \Delta z)$ are the absolute errors of the measurements. Different methods of the length measurement would have different absolute errors. In other words, the dimensions x, y, and z and the errors Δx, Δy, and Δz are all independent variables. Since the errors should be small (at least, one wishes so), the values of the dimensions obtained in each measurement are $x = x_0 + dx$, $y = y_0 + dy$, and $z = z_0 + dz$, where the differentials or infinitesimal variations can take their values in the intervals $dx \in [-\Delta x, \Delta x] = I_{\Delta x}$ and similarly for dy and dz. The question arises: Given the mean values $\mathbf{r}_0 = (x_0, y_0, z_0)$ and the absolute errors $\Delta\mathbf{r}$, what is the absolute error of the volume value calculated at $\mathbf{r}_0$? For each particular measurement, the error is $V(\mathbf{r}_0 + d\mathbf{r}) - V(\mathbf{r}_0) = dV(\mathbf{r}_0)$ as higher powers of the differentials $d\mathbf{r}$ are irrelevant (small). The components of $d\mathbf{r}$ are independent variables taking their values in the specified intervals. All such triples $d\mathbf{r}$ correspond to points of the error rectangle $R_\Delta = I_{\Delta x} \times I_{\Delta y} \times I_{\Delta z}$. Then the maximal or absolute error is $\Delta V = |\max dV(\mathbf{r}_0)|$, where the maximum is taken over all $d\mathbf{r} \in R_\Delta$. For example, if $\mathbf{r}_0 = (1, 2, 3)$ is in centimeters and $\Delta\mathbf{r} = (1, 1, 1)$ is in millimeters, then the absolute error of the volume is $\Delta V = |\max dV(\mathbf{r}_0)| = \max(y_0 z_0\, dx + x_0 z_0\, dy + x_0 y_0\, dz) = 0.6 + 0.3 + 0.2 = 1.1\,\text{cm}^3$, and $V = 6 \pm 1.1\,\text{cm}^3$. Here the maximum is reached at $dx = dy = dz = 0.1\,\text{cm}$. This concept can be generalized.

Definition 13.19. (Absolute and Relative Errors). *Let f be a quantity that depends on other quantities $\mathbf{r} = (x_1, x_2, ..., x_m)$; that is, $f = f(\mathbf{r})$ is a function of $\mathbf{r}$. Suppose that the values $x_i = a_i$ are known with the absolute errors Δx_i. Put $\mathbf{r}_0 = (a_1, a_2, ..., a_m)$ and $\Delta f = |\max df(\mathbf{r}_0)|$ where the maximum is taken over all $dx_i \in [-\Delta x_i, \Delta x_i]$. The numbers Δf and $\Delta f / |f(\mathbf{r}_0)|$ are called, respectively, the* absolute and relative errors *of the value of f at $\mathbf{r} = \mathbf{r}_0$.*

In the above example, the relative error of the volume measurements is $1.1/6 \approx 0.18$; that is, the accuracy of the measurements is about 18%. In general, since

$$df(\mathbf{r}_0) = \sum_{i=1}^{m} f'_{x_i}(\mathbf{r}_0)\, dx_i$$

is linear in dx_i, the maximum is attained at $dx_i = \Delta x_i$ if the coefficient $f'_{x_i}(\mathbf{r}_0)$ is positive, and at $dx_i = -\Delta x_i$ if the coefficient $f'_{x_i}(\mathbf{r}_0)$ is negative. So the absolute error can be written in the form

$$\Delta f = \sum_{i=1}^{m} |f'_{x_i}(\mathbf{r}_0)|\, \Delta x_i .$$

92.2. Accuracy of the Linear Approximation. In the one-variable case, the Taylor theorem asserts that if $f(x)$ is an n times continuously differentiable function, then there is a constant $M_n > 0$ such that

$$\begin{aligned} f(x) &= T_n(x) + \varepsilon_{n+1}(x)\,, \\ T_n(x) &= f(x_0) + \frac{f'(x_0)}{1!}\Delta + \frac{f''(x_0)}{2!}\Delta^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}\Delta^n\,, \end{aligned}$$

$$|\varepsilon_{n+1}(x)| \le \frac{M_{n+1}}{(n+1)!}|x - x_0|^{n+1}\,, \tag{13.9}$$

where $\Delta = x - x_0$. The polynomial $T_n(x)$ is called the *Taylor polynomial of degree n*. The remainder $\varepsilon_{n+1}(x)$ determines the accuracy of the approximation $f(x) \approx T_n(x)$. If, in addition, the derivative $f^{(n+1)}$ exists, the remainder is proved to have the form

$$\varepsilon_{n+1}(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}.$$

In this case, M_{n+1} is an upper bound of $|f^{(n+1)}(\xi)|$ over the interval between x and x_0.

The first-order Taylor polynomial is the linearization of f at $x = x_0$, $T_1(x) = L(x)$, and the remainder $\varepsilon_2(x)$ determines the accuracy of the linear approximation. Suppose that $f''(x)$ is bounded on an interval (a, b), that is, $|f''(x)| \le M_2$ for all $x \in (a, b)$. Then, for $x_0 \in (a, b)$, the accuracy of the linear approximation is given by

$$|\varepsilon_2(x)| = |f(x) - L(x)| \le \frac{M_2}{2}(b - a)^2\,, \quad x \in (a, b). \tag{13.10}$$

92.3. Two Variable Taylor Polynomials. There is a multivariable extension of the Taylor theorem that can be used to assess the accuracy of the linear approximation as well as to obtain the Taylor polynomial approximation when the linear one is not sufficiently accurate. The two-variable case is discussed first. It is assumed in what follows that functions have continuous partial derivatives of needed orders. The following notation is adopted. Let $f(x, y)$ be a function of two variables. The symbols ∂_x and ∂_y denote the operation of taking partial derivatives with respect to x and y, that is,

$$(a\partial_x + b\partial_y)^2 f = (a^2\partial_x^2 + 2ab\partial_x\partial_y + b^2\partial_y^2)f = a^2 f''_{xx} + 2ab f''_{xy} + b^2 f''_{yy}$$

for any numbers a and b. By an earlier assumption, all the derivatives are continuous, and therefore Clairaut's theorem applies, $\partial_x\partial_y f = \partial_y\partial_x f$; that is, the order of differentiation is irrelevant.

Let $\mathbf{r}_0 = (x_0, y_0)$ be a fixed point and let a function $f(\mathbf{r})$, $\mathbf{r} = (x, y)$, have continuous partial derivatives up to order n in an open disk D containing $\mathbf{r}_0$. Let $\boldsymbol{\Delta} = (\Delta_1, \Delta_2)$ be an ordered pair of numbers. Consider a polynomial in the variables $\boldsymbol{\Delta}$ of degree n defined by

$$\begin{aligned} P_n(\boldsymbol{\Delta}) &= f(\mathbf{r}_0) + \frac{1}{1!}(\Delta_1\partial_x + \Delta_2\partial_y)f(\mathbf{r}_0) \\ &\quad + \frac{1}{2!}(\Delta_1\partial_x + \Delta_2\partial_y)^2 f(\mathbf{r}_0) + \cdots \\ &\quad + \frac{1}{n!}(\Delta_1\partial_x + \Delta_2\partial_y)^n f(\mathbf{r}_0) \\ &= \sum_{k=0}^{n} \frac{1}{k!}(\Delta_1\partial_x + \Delta_2\partial_y)^k f(\mathbf{r}_0), \end{aligned}$$

where the partial derivatives are computed at the point $\mathbf{r}_0$. By construction, these polynomials satisfy the recurrence relation:

$$P_n(\boldsymbol{\Delta}) = P_{n-1}(\boldsymbol{\Delta}) + \frac{1}{n!}(\Delta_1\partial_x + \Delta_2\partial_y)^n f(\mathbf{r}_0). \tag{13.11}$$

Here the last term is computed by means of binomial coefficients $(a + b)^n = \sum_{k=0}^{n} B_n^k a^{n-k} b^k$, where $B_k^n = n!/(k!(n-k)!)$.

Definition 13.20. (Taylor Polynomials).
The polynomial $T_n(\mathbf{r}) = P_n(\boldsymbol{\Delta})$, *where* $\boldsymbol{\Delta} = \mathbf{r} - \mathbf{r}_0 = (x - x_0, y - y_0)$ *is called the* Taylor polynomial *for a function f near the point* $\mathbf{r}_0$. *The approximation* $f(\mathbf{r}) \approx T_n(\mathbf{r})$ *is called the* Taylor polynomial approximation of degree n near $\mathbf{r}_0$.

For example, with $\Delta_1 = x - x_0$ and $\Delta_2 = y - y_0$, the first four Taylor polynomials are

$$\begin{aligned} T_0(\mathbf{r}) &= f(\mathbf{r}_0), \\ T_1(\mathbf{r}) &= f(\mathbf{r}_0) + f'_x(\mathbf{r}_0)\,\Delta_1 + f'_y(\mathbf{r}_0)\,\Delta_2 = L(\mathbf{r}), \\ T_2(\mathbf{r}) &= T_1(\mathbf{r}) + \frac{f''_{xx}(\mathbf{r}_0)}{2}\,\Delta_1^2 + f''_{xy}(\mathbf{r}_0)\,\Delta_1\Delta_2 + \frac{f''_{yy}(\mathbf{r}_0)}{2}\,\Delta_2^2, \\ T_3(\mathbf{r}) &= T_2(\mathbf{r}) + \frac{f'''_{xxx}(\mathbf{r}_0)}{6}\,\Delta_1^3 + \frac{f'''_{xxy}(\mathbf{r}_0)}{2}\,\Delta_1^2\Delta_2 + \frac{f'''_{xyy}(\mathbf{r}_0)}{2}\,\Delta_1\Delta_2^2 \\ &\quad + \frac{f'''_{yyy}(\mathbf{r}_0)}{6}\,\Delta_2^3. \end{aligned}$$

The linear or tangent plane approximation $f(\mathbf{r}) \approx L(\mathbf{r}) = T_1(\mathbf{r})$ is a particular case of the Taylor polynomial approximation of the first degree.

The following theorem assesses the accuracy of the Taylor polynomial approximation.

THEOREM 13.14. (Accuracy of the Taylor Polynomial Approximation). *Let D be an open disk centered at $\mathbf{r}_0$ and let the partial derivatives of a function f be continuous up to order n on D. Suppose that the partial derivatives of order n are also bounded on D; that is, there are numbers M_{nk}, $k = 1, 2, ..., n$, such that $|\partial_x^{n-k}\partial_y^k f(\mathbf{r})| \le M_{nk}$ for all $\mathbf{r} \in D$. Then $f(\mathbf{r}) = T_{n-1}(\mathbf{r}) + \varepsilon_n(\mathbf{r})$ where the remainder ε_n satisfies*

$$|\varepsilon_n(\mathbf{r})| \le \sum_{k=0}^{n} \frac{B_k^n M_{nk}}{n!}\, |x - x_0|^{n-k}|y - y_0|^k$$

for all $(x, y) \in D$, where $B_k^n = n!/(k!(n-k)!)$ are binomial coefficients.

Remark. In fact, the continuity of the nth-order partial derivatives is not necessary; their existence and boundedness are sufficient for the accuracy assessment. The discussion of this remark as well as the proof of the theorem goes beyond the scope of the course.

To make the analogy of this theorem with the one-variable case, note that $|x - x_0| \le \|\mathbf{r} - \mathbf{r}_0\|$ and $|y - y_0| \le \|\mathbf{r} - \mathbf{r}_0\|$ and hence $|x - x_0|^{n-k}|y - y_0|^k \le \|\mathbf{r} - \mathbf{r}_0\|^n$. Making use of this inequality, one infers that

$$|\varepsilon_n(\mathbf{r})| \le \frac{M_n}{n!}\|\mathbf{r} - \mathbf{r}_0\|^n, \tag{13.12}$$

where the constant $M_n = \sum_{k=0}^{n} B_k^n M_{nk}$. In particular, for the linear approximation $n = 2$,

$$\begin{aligned} |f(\mathbf{r}) - L(\mathbf{r})| &\le \frac{M_{20}}{2}(x - x_0)^2 + M_{11}|(x - x_0)(y - y_0)| + \\ &\frac{M_{02}}{2}(y - y_0)^2 \le \frac{M_2}{2}\|\mathbf{r} - \mathbf{r}_0\|^2 \le \frac{M_2}{2}R^2\,, \end{aligned} \tag{13.13}$$

where $M_2 = M_{20} + 2M_{11} + M_{02}$ and R is the radius of the disk D. The results (13.12) and (13.13) are to be compared with the similar results (13.9) and (13.10) in the one-variable case. So, if the second derivatives exist and are bounded, the error of the linear or tangent plane approximation decreases as the squared distance $\|\mathbf{r} - \mathbf{r}_0\|^2$.

EXAMPLE 13.24. *Use the linear approximation or the differential to estimate the amount of aluminum in a closed aluminum can with* *diameter 10 cm and height 10 cm if the aluminum is 0.05 cm thick. Assess the accuracy of the estimate.*

Solution: The volume of a cylinder of radius r and height h is $f(h,r) = \pi h r^2$. The volume of a closed cylindrical shell (or the can) of thickness δ is therefore $V = f(h+2\delta, r+\delta) - f(h,r)$, where h and r are the internal height and radius of the shell. Since the variations $\Delta h = 2\delta = 0.1$ and $\Delta r = \delta = 0.05$ are small, this difference can be estimated by linearizing the function f at $(h,r) = (10,5)$. One has $f'_h = \pi r^2$ and $f'_r = 2\pi h r$; hence, $V \approx V_a = df(10,5) = f'_h(10,5)\,dh + f'_r(10,5)\,dr = 25\pi\,\Delta h + 100\pi\,\Delta r = 7.5\pi$ cm^3, where $dh = \Delta h$ and $dr = \Delta r$.

To assess the accuracy, note that the approximation $V \approx V_a$ is based on the linear approximation of $f(h,r)$, near the point $(10,5)$. Put $f(h,r) = L(h,r) + \varepsilon_2(h,r)$, where $L(h,r)$ is the linearization of f at $(10,5)$ and $\varepsilon_2(h,r)$ is the remainder. Then, for $h = 10+2\delta$ and $r = 5+\delta$, one has $V = f(h,r) - f(10,5) = L(h,r) - f(10,5) + \varepsilon_2(h,r) = V_a + \varepsilon_2(h,r)$. To estimate the remainder, one has to find the upper bounds on the second-order derivatives of f when (h,r) lies in the disk of radius $R = 2\delta$ centered at $(10,5)$ (see (13.13)). One has $f''_{hh} = 0 = M_{20}$, $f''_{hr} = 2\pi r \leq 10.2\pi = M_{11}$, and $f''_{rr} = 2\pi h \leq 20.2\pi = M_{02}$. Therefore, $M_2 = 40.4\pi$, and the absolute error of the estimate is $|V - V_a| = |\varepsilon_2| \leq (M_2/2)R^2 = 80.8\pi\delta^2 = 0.202\pi$ cm^3. The relative error reads $|V - V_a|/V_a \approx 0.026$; that is, it is about 2.6%. □

92.4. Multivariable Taylor Polynomials. For more than two variables, Taylor polynomials are defined similarly. Let $\mathbf{r} = (x_1, x_2, ..., x_m)$ and let $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, ..., \Delta_m)$. Consider polynomials defined by the recurrence relation (13.11), where $\Delta_1\partial_x + \Delta_2\partial_y$ is replaced by $\Delta_1\partial_1 + \Delta_2\partial_2 + \cdots + \Delta_m\partial_m = D_\Delta$ and ∂_i means the operation of taking the derivative with respect to x_i, that is, $\partial_i f = \partial f/\partial x_i$. The Taylor polynomial $T_n(\mathbf{r})$ of degree n near the point $\mathbf{r}_0 = (a_1, a_2, ..., a_m)$ is then obtained by setting $\Delta_i = x_i - a_i$ in the polynomial $P_n(\boldsymbol{\Delta})$; that is, in the compact form the rule reads

$$P_n(\boldsymbol{\Delta}) = \sum_{k=1}^{n} \frac{1}{k!} D_\Delta^k f(\mathbf{r}_0)\,, \qquad T_n(\mathbf{r}) = P_n(\mathbf{r} - \mathbf{r}_0)\,.$$

The accuracy of the approximation $f(\mathbf{r}) \approx T_{n-1}(\mathbf{r})$ can be assessed by putting upper bounds on the nth-order derivatives of f in a ball centered at $\mathbf{r}_0$. Due to excessive technicalities, the details are omitted. It is worth noting that inequality (13.12) holds for any number of variables; that is the difference $|f(\mathbf{r}) - T_{n-1}(\mathbf{r})|$ tends to 0 no slower than $\|\mathbf{r} - \mathbf{r}_0\|^n$ as $\mathbf{r} \to \mathbf{r}_0$.

For practical purposes, if the difference $|T_n(\mathbf{r}) - T_{n-1}(\mathbf{r})|$ is small as compared to $|T_{n-1}(\mathbf{r})|$ in a ball $\|\mathbf{r} - \mathbf{r}_0\| \leq R$, then the approximation $f(\mathbf{r}) \approx T_{n-1}(\mathbf{r})$ is accurate; that is, the approximation by a higher-order polynomial T_n instead of T_{n-1} is not going to significantly improve the accuracy.

92.5. Study Problems.

Problem 13.7. *Calculation of higher-order derivatives to find Taylor polynomials might be a technically tedious problem. In some special cases, however, it can be avoided. Suppose f is a composition of two functions: $f(\mathbf{r}) = g(u)$, where $u = u(\mathbf{r})$. Suppose that $u(\mathbf{r}_0) = 0$, that is, $f(\mathbf{r}_0) = g(0)$. Let $T_n(\mathbf{r}) = P_n(\mathbf{\Delta})$ be a Taylor polynomial for the function u at $\mathbf{r}_0$. It has the property that it has no constant term as $u(\mathbf{r}_0) = 0$. Therefore, its kth power, $(P_n(\mathbf{\Delta}))^k$, is a polynomial of degree nk whose constant term, linear terms, quadratic terms, and so on up to the $(k-1)$th-degree terms vanish. Consider the Taylor polynomial for the function g:*

$$T_n^g(u) = g(0) + g'(0)u + \frac{g''(0)}{2!}u^2 + \cdots + \frac{g^{(n)}(0)}{n!}u^n.$$

If u is approximated by its Taylor polynomial $P_n(\mathbf{\Delta})$ in this expression, then the resulting expression will be a polynomial in $\mathbf{\Delta}$ whose degree is n^2 and whose terms up to degree n coincide with the Taylor polynomial T_n^f for f at $\mathbf{r}_0$. Indeed, any polynomial in $\mathbf{\Delta}$ is uniquely determined by its coefficients, including T_n^f. On the other hand, an approximation of g by a higher-degree polynomial T_{n+1}^g will add the term u^{n+1}, which can only contain terms in $\mathbf{\Delta}$ of degree $n+1$ or higher, and hence will not change the terms of degree less than or equal to n. Use this observation to find T_3 for the function $f(x, y, z) = \sin(xy + z)$ at the origin.

Solution: Put $u = xy + z$. The third-degree Taylor polynomial for $g(u) = \sin u$ at $u = 0$ is $T_3^g(u) = u - \frac{1}{6}u^3$. As u is a polynomial of degree 2, its Taylor polynomial of degree 3 coincides with it, $T_3(\mathbf{r}) = u(\mathbf{r})$. Hence, T_3^f is obtained from T_3^g by omitting all terms of degree higher than 3:

$$T_3^g = (xy+z) - \frac{1}{6}(xy+z)^3 = z + xy - \frac{1}{6}(z^3 + 3(xy)z^2 + 3(xy)^2z + (xy)^3).$$

Therefore, $T_3^f(\mathbf{r}) = z + xy - \frac{1}{6}z^3$. Evidently, the procedure is far simpler than calculating 19 partial derivatives (up to the third order)! □

93. Directional Derivative and the Gradient

93.1. Directional Derivative. Let f be a function of several variables $\mathbf{r} = (x_1, x_2, ..., x_m)$. The partial derivative $f'_{x_i}(\mathbf{r}_0)$ is the rate of change in the direction of the ith coordinate axis. This direction is defined by the unit vector $\hat{\mathbf{e}}_i$ parallel to the corresponding coordinate axis. Let $\hat{\mathbf{u}}$ be a unit vector that does not coincide with any of the vectors $\hat{\mathbf{e}}_i$. What is the rate of change of f at $\mathbf{r}_0$ in the direction of $\hat{\mathbf{u}}$? For example, if $f(x, y)$ is the height of a mountain, where the x and y axes are oriented along the west–east and south–north directions, respectively, then it is reasonable to ask about the slopes, for example, in the south–east or north–west directions. Naturally, these slopes generally differ from the slopes f'_x and f'_y.

To answer the question about the slope in the direction of a unit vector $\hat{\mathbf{u}}$, consider a straight line through $\mathbf{r}_0$ parallel to $\hat{\mathbf{u}}$. Its vector equation is $\mathbf{r}(h) = \mathbf{r}_0 + h\hat{\mathbf{u}}$, where h is a parameter that labels points of the line. The values of f along the line are given by the composition $g(h) = f(\mathbf{r}(h))$. The numbers $g(0)$ and $g(h)$ are the values of f at a given point $\mathbf{r}_0$ and the point $\mathbf{r}(h)$, $h > 0$, that is at the distance h from $\mathbf{r}_0$ in the direction of $\hat{\mathbf{u}}$. So the slope is given by the derivative $g'(0)$. Therefore, the following definition is natural.

Definition 13.21. (Directional Derivative).
Let f be a function on an open set D. The directional derivative of f at $\mathbf{r}_0 \in D$ in the direction of a unit vector $\hat{\mathbf{u}}$ is the limit

$$D_{\mathbf{u}}f(\mathbf{r}_0) = \lim_{h\to 0} \frac{f(\mathbf{r}_0 + h\hat{\mathbf{u}}) - f(\mathbf{r}_0)}{h}$$

if the limit exists.

The number $D_{\mathbf{u}}f(\mathbf{r}_0)$ is the rate of change of f at $\mathbf{r}_0$ in the direction of $\hat{\mathbf{u}}$. By definition, $D_{\mathbf{u}}f(\mathbf{r}_0) = df(\mathbf{r}(h))/dh$ taken at $h = 0$, where $\mathbf{r}(h) = \mathbf{r}_0 + h\hat{\mathbf{u}}$. So, by the chain rule, the directional derivative exists if the partial derivatives of f at $\mathbf{r}_0$ exist:

$$\frac{df(\mathbf{r}(h))}{dh} = f'_{x_1}(\mathbf{r}(h))x'_1(h) + f'_{x_2}(\mathbf{r}(h))x'_2(h) + \cdots + f'_{x_m}(\mathbf{r}(h))x'_m(h)\,.$$

Setting $h = 0$ in this relation and taking into account that $\mathbf{r}'(h) = \hat{\mathbf{u}}$ or $x'_i(h) = u_i$, where $\hat{\mathbf{u}} = (u_1, u_2, ..., u_m)$, one infers that

$$D_{\mathbf{u}}f(\mathbf{r}_0) = f'_{x_1}(\mathbf{r}_0)u_1 + f'_{x_2}(\mathbf{r}_0)u_2 + \cdots + f'_{x_m}(\mathbf{r}_0)u_m\,. \tag{13.14}$$

Equation (13.14) provides a convenient way to compute the directional derivative. Recall also that if the direction is specified by a nonunit vector $\mathbf{u}$, then the corresponding unit vector can be obtained by dividing it by its length $\|\mathbf{u}\|$, that is, $\hat{\mathbf{u}} = \mathbf{u}/\|\mathbf{u}\|$.

Example 13.25. *The height of a hill is* $f(x, y) = (9 - 3x^2 - y^2)^{1/2}$, *where the* x *and* y *axes are directed from west to east and from south to north, respectively. A hiker is at the point* $\mathbf{r}_0 = (1, 2)$. *Suppose the hiker is facing in the north-west direction. What is the slope the hiker sees?*

Solution: A unit vector in the plane can always be written in the form $\hat{\mathbf{u}} = (\cos\varphi, \sin\varphi)$, where the angle φ is counted counterclockwise from the positive x axis; that is, $\varphi = 0$ corresponds to the east direction, $\varphi = \pi/2$ to the north direction, $\varphi = \pi$ to the west direction, and so on. So for the north–west direction $\varphi = 3\pi/2$ and $\hat{\mathbf{u}} = (-1/\sqrt{2}, 1/\sqrt{2}) = (u_1, u_2)$. The partial derivatives are $f'_x = -3x/(9 - 3x^2 - y^2)^{1/2}$ and $f'_y = -y/(9 - 3x^2 - y^2)^{1/2}$. Their values at $\mathbf{r}_0 = (1, 2)$ are $f'_x(1, 2) = -3/\sqrt{2}$ and $f'_y(1, 2) = -2/\sqrt{2}$. By (13.14), the slope is

$$D_{\mathbf{u}} f(\mathbf{r}_0) = f'_x(\mathbf{r}_0)u_1 + f'_y(\mathbf{r}_0)u_2 = 3/2 - 1/2 = 1\,.$$

If the hiker goes north–west, he has to climb at an angle of 45° relative to the horizon. □

Example 13.26. *Find the directional derivative of* $f(x, y, z) = x^2 + 3xz + z^2 y$ *at the point* $(1, 1, -1)$ *in the direction toward the point* $(3, -1, 0)$. *Does the function increase or decrease in this direction?*

Solution: Put $\mathbf{r}_0 = (1, 1, -1)$ and $\mathbf{r}_1 = (3, -1, 0)$. Then the vector $\mathbf{u} = \mathbf{r}_1 - \mathbf{r}_0 = (2, -2, 1)$ points from the point $\mathbf{r}_0$ toward the point $\mathbf{r}_1$ according to the rules of vector algebra. But it is not a unit vector because its length is $\|\mathbf{u}\| = 3$. So the unit vector in the same direction is $\hat{\mathbf{u}} = \mathbf{u}/3 = (2/3, -2/3, 1/3) = (u_1, u_2, u_3)$. The partial derivatives are $f'_x = 2x + 3z$, $f'_y = z^2$, and $f'_z = 3x + 2zy$. Their values at $\mathbf{r}_0$ read $f'_x(\mathbf{r}_0) = -1$, $f'_y(\mathbf{r}_0) = 1$, and $f'_z(\mathbf{r}_0) = 1$. By (13.14), the directional derivative is

$$D_{\mathbf{u}} f(\mathbf{r}_0) = f'_x(\mathbf{r}_0)u_1 + f'_y(\mathbf{r}_0)u_2 + f'_z(\mathbf{r}_0)u_3 = -2/3 - 2/3 + 1/3 = -1\,.$$

Since the directional derivative is negative, the function decreases at $\mathbf{r}_0$ in the direction toward $\mathbf{r}_1$ (the rate of change is negative in that direction). □

93.2. The Gradient and Its Geometrical Significance.

Definition 13.22. (The Gradient).
Let f be a function of several variables $\mathbf{r} = (x_1, x_2, ..., x_m)$ on an open set D and let $\mathbf{r}_0 \in D$. The vector whose components are partial derivatives of f at $\mathbf{r}_0$,

$$\nabla f(\mathbf{r}_0) = (f'_{x_1}(\mathbf{r}_0),\ f'_{x_2}(\mathbf{r}_0),\ \dots\ ,\ f'_{x_m}(\mathbf{r}_0)),$$

is the gradient of f at the point $\mathbf{r}_0$.

So, for two-variable functions $f(x, y)$, the gradient is $\nabla f = (f'_x,\ f'_y)$; for three-variable functions $f(x, y, z)$, the gradient is $\nabla f = (f'_x, f'_y, f'_z)$; and so on. Comparing (13.14) with the definition of the gradient and recalling the definition of the dot product, the directional derivative can now be written in the compact form

$$D_{\mathbf{u}} f(\mathbf{r}_0) = \nabla f(\mathbf{r}_0) \cdot \hat{\mathbf{u}}\,. \tag{13.15}$$

This equation is the most suitable for analyzing the significance of the gradient.

Consider first the cases of two- and three-variable functions. The gradient is either a vector in a plane or space. In Example 13.25, the gradient at $(1, 2)$ is $\nabla f(1, 2) = (-3/\sqrt{2}, -2/\sqrt{2})$. In Example 13.26, the gradient at $(1, 1, -1)$ is $\nabla f(1, 1, -1) = (-1, 1, 1)$. Recall the geometrical property of the dot product $\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\|\|\mathbf{b}\| \cos\theta$, where $\theta \in [0, \pi]$ is the angle between the vectors $\mathbf{a}$ and $\mathbf{b}$. The value $\theta = 0$ corresponds to parallel vectors $\mathbf{a}$ and $\mathbf{b}$. When $\theta = \pi/2$, the vectors are orthogonal. The vectors point in the opposite directions if $\theta = \pi$. Let θ be the angle between the gradient $\nabla f(\mathbf{r}_0)$ and the unit vector $\hat{\mathbf{u}}$. Then

$$D_{\mathbf{u}} f(\mathbf{r}_0) = \nabla f(\mathbf{r}_0) \cdot \hat{\mathbf{u}} = \|\nabla f(\mathbf{r}_0)\|\|\hat{\mathbf{u}}\| \cos\theta = \|\nabla f(\mathbf{r}_0)\| \cos\theta \tag{13.16}$$

because $\|\hat{\mathbf{u}}\| = 1$ (the unit vector). As the components of the gradient are fixed numbers (the values of the partial derivatives at a particular point $\mathbf{r}_0$), the directional derivative at $\mathbf{r}_0$ varies only if the vector $\hat{\mathbf{u}}$ changes. Thus, the rates of change of f in all directions that have the same angle θ with the gradient are the same. In the two-variable case, only two such directions are possible if $\hat{\mathbf{u}}$ is not parallel to the gradient, while in the three-variable case the rays from $\mathbf{r}_0$ in all such directions form a cone whose axis is along the gradient. It is then concluded that the rate of change of f attains its absolute maximum or minimum when $\cos\theta$ does. Therefore, *the maximal rate is attained in the direction of the gradient ($\theta = 0$) and is equal to the magnitude of the gradient $\|\nabla f(\mathbf{r}_0)\|$, whereas the minimal rate of change $-\|\nabla f(\mathbf{r}_0)\|$*

occurs in the direction of $-\nabla f(\mathbf{r}_0)$, that is, opposite to the gradient $(\theta = \pi)$.

The graph of a function of two variables $z = f(x, y)$ may be viewed as the shape of a hill. Then the gradient at a particular point shows the direction of the *steepest ascent*, while its opposite points in the direction of the *steepest descent*. In Example 13.25, the maximal slope at the point $(1, 2)$ is $\|\nabla f(\mathbf{r}_0)\| = (1/\sqrt{2})\ \|(-3, 2)\| = \sqrt{13/2}$. It occurs in the direction of $(-3/\sqrt{2}, 2/\sqrt{2})$ or $(-3, 2)$ (the multiplication of a vector by a positive constant does not change its direction). If φ is the angle between the positive x axis (or the vector $\hat{\mathbf{e}}_1$) and the gradient, then $\tan\varphi = -2/3$ or $\varphi \approx 146°$. If the hiker goes in this direction, he has to climb up at an angle of $\tan^{-1}(\sqrt{13/2}) \approx 69°$ with the horizon. Also, note the hiker's original direction was $\varphi = 135°$, which makes the angle $11°$ with the direction of the steepest ascent. So the slope in the direction $\varphi = 146° + 11° = 157°$ has the same slope as the hiker's original one. As has been argued, in the two-variable case, there can only be two directions with the same slope.

Next, consider a level curve $f(x, y) = k$ of a function of two variables. There is a vector function $\mathbf{r}(t) = (x(t), y(t))$ that traces out the level curve. This vector function is defined by the condition that $f(x(t), y(t)) = k$ for all values of the parameter t. By the definition of level curves, the function f has a constant value k along its level curve. Therefore, by the chain rule,

$$\frac{d}{dt} f(x(t), y(t)) = 0 \quad \Longrightarrow \quad \frac{df}{dt} = f'_x x'(t) + f'_y y'(t) = \nabla f(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = 0$$

for any value of t. For any particular value $t = t_0$, the point $\mathbf{r}_0 = \mathbf{r}(t_0)$ lies on the level curve, while the derivative $\mathbf{r}'(t_0)$ is a tangent vector to the curve at the point $\mathbf{r}_0$. Thus, *the gradient $\nabla f(\mathbf{r}_0)$ is orthogonal to a tangent vector at the point $\mathbf{r}_0$ to the level curve of f through that point. One can also say the gradient of f is always normal to the level curves of f.*

Recall that a function $f(x, y)$ can be described by a contour map, which is a collection of level curves. If level curves are smooth enough to have tangent vectors everywhere, then one can define a curve through a particular point that is normal to all level curves in some neighborhood of that point. This curve is called the *curve of steepest descent or ascent*. The tangent vector of this curve at any point is parallel to the gradient at that point. The values of the function increase (or decrease) most rapidly along this curve. If a hiker follows the direction of the gradient of the height, he would go along the path of steepest ascent or descent.

Consider a level surface $f(x, y, z) = k$. Let $\mathbf{r}(t) = (x(t), y(t), z(t))$ be a smooth curve on the level surface, that is, $f(\mathbf{r}(t)) = k$ for all values of t. Since the values of f do not change along the curve, $df/dt = 0$. Making use of the chain rule, it is concluded that

$$df/dt = f'_x x' + f'_y y' + f'_z z' = \nabla f(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = 0.$$

There are many curves through a point $\mathbf{r}_0 = \mathbf{r}(t_0)$ that lie in the level surface. The gradient $\nabla f(\mathbf{r}_0)$ is orthogonal to a tangent vector to *any* such curve, that is, to *any* line that is tangent to the level surface at $\mathbf{r}_0$. Thus, *the gradient* $\nabla f(\mathbf{r}_0)$ *is normal to the tangent plane to the level surface through the point* $\mathbf{r}_0$.

All the findings are summarized in the following theorem, which has been proved above.

Theorem 13.15. (Geometrical Properties of the Gradient). *Let f be differentiable on an open set D and let $\mathbf{r}_0 \in D$. Let S be the level surface (or curve) through the point $\mathbf{r}_0$. Then*

1. *The maximal rate of change of f at $\mathbf{r}_0$ occurs in the direction of the gradient $\nabla f(\mathbf{r}_0)$ and is equal to its magnitude $\|\nabla f(\mathbf{r}_0)\|$.*
2. *The minimal rate of change of f at $\mathbf{r}_0$ occurs in the direction opposite to the gradient $-\nabla f(\mathbf{r}_0)$ and equals $-\|\nabla f(\mathbf{r}_0)\|$.*
3. *The gradient $\nabla f(\mathbf{r}_0)$ is normal to S at $\mathbf{r}_0$.*

Example 13.27. *Find an equation of the tangent plane to the ellipsoid $x^2 + 2y^2 + 3z^2 = 11$ at the point $(2, 1, 1)$.*

Solution: The equation of the ellipsoid can be viewed as the level surface $f(x, y, z) = 11$ of the function $f(x, y, z) = x^2+2y^2+3z^2$ through the point $\mathbf{r}_0 = (2, 1, 1)$ because $f(2, 1, 1) = 11$. By the geometrical property of the gradient, the vector $\mathbf{n} = \nabla f(\mathbf{r}_0)$ is normal to the plane in question. Since $\nabla f = (2x, 4y, 6z)$, one has $\mathbf{n} = (4, 4, 6)$. An equation of the plane through the point $(2, 1, 1)$ and normal to $\mathbf{n}$ is $4(x - 2) + 4(y - 1) + 6(z - 1) = 0$ or $2x + 2y + 3z = 9$. □

Theorem 13.15 holds for functions of more than three variables as well. Equation (13.15) was obtained for any number of variables, and the representation of the dot product (13.16) holds in any Euclidean space. Thus, the first two properties of the gradient are valid in any multivariable case. The third property is harder to visualize as the level surface of a function of m variables is an $(m - 1)$-dimensional surface embedded in an m-dimensional Euclidean space. Such surfaces are called *hypersurfaces* to emphasize the fact that they are not two-dimensional surfaces embedded in a three-dimensional Euclidean space. However, if $\mathbf{r}(t)$ is a curve in the level hypersurface $f(\mathbf{r}) = k$, then the multivariable function f has a constant value along any such curve, and, by the chain rule, it immediately follows that $df(\mathbf{r}(t))/dt = \nabla f(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = 0$ for any t. Hence, *the gradient is normal to the level hypersurface in the sense that it is normal to any curve in it.*

Remark. It is interesting to note compact formulas for the linearization $L(\mathbf{r})$ of $f(\mathbf{r})$ at $\mathbf{r}_0$ and the differential df:

$$L(\mathbf{r}) = f(\mathbf{r}_0) + \nabla f(\mathbf{r}_0) \cdot (\mathbf{r} - \mathbf{r}_0)\,, \qquad df = \nabla f \cdot d\mathbf{r}$$

that are valid for any number of variables.

93.3. Study Problems.

Problem 13.8. *Suppose that three level surfaces $f(x, y, z) = 1$, $g(x, y, z) = 2$, and $h(x, y, z) = 3$ of differentiable functions are intersecting along a smooth curve C. Let P be a point on C. Find $\nabla f \cdot (\nabla g \times \nabla h)$ at P.*

Solution: Let $\mathbf{v}$ be a tangent vector to C at the point P (it exists because the curve is smooth). Since C lies in the surface $f(x, y, z) = 1$, the gradient $\nabla f(P)$ is orthogonal to $\mathbf{v}$. Similarly, the gradients $\nabla g(P)$ and $\nabla h(P)$ must be orthogonal to $\mathbf{v}$. Therefore, all the gradients must be in a plane perpendicular to the vector $\mathbf{v}$. The triple product for any three coplanar vectors vanishes, and hence $\nabla f \cdot (\nabla g \times \nabla h) = 0$ at P. □

Problem 13.9. *Consider Newton's second law $m\mathbf{a} = \mathbf{F}$. Suppose that the force is the gradient $\mathbf{F} = -\nabla U$, where $U = U(\mathbf{r})$. Let $\mathbf{r} = \mathbf{r}(t)$ be the trajectory satisfying Newton's law. Prove that the quantity $E = mv^2/2 + U(\mathbf{r})$, where $v = \|\mathbf{r}'(t)\|$ is the speed, is a constant of motion, that is, $dE/dt = 0$. This constant is called the total energy of a particle.*

Solution: First, note that $v^2 = \mathbf{v} \cdot \mathbf{v}$. Hence, $(v^2)' = 2\mathbf{v} \cdot \mathbf{v}' = 2\mathbf{v} \cdot \mathbf{a}$. Using the chain rule, $dU/dt = U'_x x'(t) + U'_y y'(t) + U'_z z'(t) = \mathbf{r}' \cdot \nabla U = \mathbf{v} \cdot \nabla U$. It follows from these two relations that

$$\frac{dE}{dt} = \frac{m}{2}(v^2)' + \frac{dU}{dt} = m\mathbf{v} \cdot \mathbf{a} + \mathbf{v} \cdot \nabla U = \mathbf{v} \cdot (m\mathbf{a} - \mathbf{F}) = 0$$

So the total energy is conserved for the trajectory of the motion. □

94. Maximum and Minimum Values

94.1. Critical Points of Multivariable Functions. The positions of the local maxima and minima of a one-variable function play an important role when analyzing its overall behavior. In Calculus I, it was shown how the derivatives can be used to find local maxima and minima. Here this analysis is extended to multivariable functions.

The following notation will be used. An open ball of radius δ centered at a point $\mathbf{r}_0$ is denoted $B_\delta = \{\mathbf{r} \,|\, \|\mathbf{r}-\mathbf{r}_0\| < \delta\}$; that is, it is a set of points whose distance from $\mathbf{r}_0$ is less than $\delta > 0$. A neighborhood N_δ of a point $\mathbf{r}_0$ in a set D is a set of common points of D and B_δ; that is, $N_\delta = D \cap B_\delta$ contains all points in D whose distance from $\mathbf{r}_0$ is less than δ.

Definition 13.23. (Absolute and Local Maxima or Minima). *A function f on a set D is said to have a local maximum at $\mathbf{r}_0 \in D$ if there is a neighborhood N_δ of $\mathbf{r}_0$ such that $f(\mathbf{r}_0) \geq f(\mathbf{r})$ for all $\mathbf{r} \in N_\delta$. The number $f(\mathbf{r}_0)$ is called a* local maximum value*. If there is a neighborhood N_δ of $\mathbf{r}_0$ such that $f(\mathbf{r}_0) \leq f(\mathbf{r})$ for all $\mathbf{r} \in N_\delta$, then f is said to have a local minimum at $\mathbf{r}_0$ and the number $f(\mathbf{r}_0)$ is called a* local minimum value*. If the inequality $f(\mathbf{r}_0) \geq f(\mathbf{r})$ or $f(\mathbf{r}_0) \leq f(\mathbf{r})$ holds for all points $\mathbf{r}$ in the domain of f, then f has an absolute maximum or absolute minimum at $\mathbf{r}_0$, respectively.*

There is an extension of Fermat's theorem to multivariable functions that is helpful in finding the possible positions of the local maxima and minima, provided the function has first-order partial derivatives.

Theorem 13.16. *If f has a local maximum or minimum at an interior point $\mathbf{r}_0$ of its domain D and the first-order partial derivatives exist at $\mathbf{r}_0$, then they vanish at $\mathbf{r}_0$, $f'_{x_i}(\mathbf{r}_0) = 0$, $i = 1, 2, ..., m$.*

Proof. Consider a line $\mathbf{r}(t) = \mathbf{r}_0 + t\hat{\mathbf{u}}$ through $\mathbf{r}_0$ and parallel to a unit vector $\mathbf{u}$. The function $F(t) = f(\mathbf{r}(t))$ defines the values of f along the line. Therefore, $F(t)$ must have a local maximum or minimum at $t = 0$. The derivative F' exists at $t = 0$ because, by the definition of the directional derivative, $F'(0) = D_{\mathbf{u}} f(\mathbf{r}_0) = \nabla f(\mathbf{r}_0) \cdot \hat{\mathbf{u}}$ and the gradient $\nabla f(\mathbf{r}_0)$ exists by the hypothesis. By Fermat's theorem, $F'(0) = 0$. So the rate of change of f vanishes in any direction and, in particular, along the coordinate axes, that is, $\hat{\mathbf{u}} = \hat{\mathbf{e}}_i$ and $D_{\mathbf{u}} f(\mathbf{r}_0) = f'_{x_i}(\mathbf{r}_0) = 0$. □

The converse of this theorem is not true. This is illustrated by the example of the function $f(x, y) = xy - 2y - x$. It is differentiable everywhere. In particular, the system of equations $f'_x = y - 1 =$ and $f'_y = x - 2 = 0$ has the solution $(x, y) = (2, 1)$. However, the function has neither a local maximum nor a local minimum. Indeed, consider a straight line through $(2, 1)$, $x = 2+u_1 t$, $y = 1+u_2 t$, that is parallel to a unit vector $\hat{\mathbf{u}} = (u_1, u_2)$. Then the values of f along the line are $F(t) = f(x(t), y(t)) = -1 + at^2$, where $a = u_1 u_2$. So $F(t)$ has a minimum at $t = 0$ if $a > 0$ or a maximum if $a < 0$. However, the coefficient a may be either positive or negative, depending on the choice of the line (or the components of $\hat{\mathbf{u}}$). Thus, the value $F(0) = f(2, 1) = -1$ cannot be a local maximum or a local minimum value. The graph of f looks like a saddle in the neighborhood of $(2, 1)$.

Definition 13.24. (Saddle Point).
If the number $f(\mathbf{r}_0)$ is the maximal value f along some lines through $\mathbf{r}_0$ in a small open ball (disk) centered at $\mathbf{r}_0$, whereas $f(\mathbf{r}_0)$ is the minimal value of f along all the other such lines, then $\mathbf{r}_0$ is called a saddle point *of f.*

Remark. Consider all lines through a point $\mathbf{r}_0$ in the domain of a function f. Suppose that f attains a local maximum along every line through $\mathbf{r}_0$; that is, the function $g(t) = f(\mathbf{r}_0+\mathbf{u}t)$ has a local maximum at $t = 0$ for any choice of the vector $\mathbf{u}$. One might tend to conclude that in this case the function f should have a local maximum at $\mathbf{r}_0$. This conclusion is *wrong*! An example is given at the end of this section (see Study Problems 13.9). The remark also applies to the case of a local minimum.

A local maximum or minimum may occur at a point where some of the partial derivatives do not exist. For example, $f(x, y) = |x| + |y|$ is defined everywhere and has an absolute minimum at $(0, 0)$. However, the partial derivatives $f'_x(0, 0)$ and $f'_y(0, 0)$ do not exist.

Finally, a local minimum or maximum may occur at a point of the domain that is not an interior point, and hence the partial derivatives are not defined at that point. For example, the domain of the function $f(x, y) = \sqrt{1 - x^2 - y^2}$ is the disk $x^2 + y^2 \leq 1$. Its boundary points $x^2+y^2 = 1$ are not interior points. But this function attains its absolute minimum on the circle $x^2 + y^2 = 1$.

Definition 13.25. (Critical Points).
An interior point $\mathbf{r}_0$ of the domain of a function f is said to be a critical point of f if either $\nabla f(\mathbf{r}_0) = \mathbf{0}$ or the gradient does not exist at $\mathbf{r}_0$.

Thus, *if f has a local maximum or minimum at $\mathbf{r}_0$, then $\mathbf{r}_0$ is a critical point of f. However, not all critical points correspond to either a local maximum or a local minimum.*

94.2. Second-Derivative Test. Suppose that a function $f(x, y)$ has continuous second derivatives in an open ball centered at $\mathbf{r}_0$. The second derivatives $a = f''_{xx}(\mathbf{r}_0)$, $b = f''_{yy}(\mathbf{r}_0)$, and $c = f''_{xy}(\mathbf{r}_0) = f''_{yx}(\mathbf{r}_0)$ (Clairaut's theorem) can be arranged into a 2×2 symmetric matrix whose diagonal elements are a and b and whose off-diagonal elements c. The quadratic polynomial of a variable λ,

$$P_2(\lambda) = \det \begin{pmatrix} a - \lambda & c \\ c & b - \lambda \end{pmatrix} = (a - \lambda)(b - \lambda) - c^2,$$

is called the *characteristic polynomial* of the matrix of second partial derivatives of f at $\mathbf{r}_0$.

THEOREM 13.17. (Second-Derivative Test).
Let $\mathbf{r}_0$ be a critical point of a function f. Suppose that the second-order partial derivatives of f are continuous in an open ball (disk) B_δ centered at $\mathbf{r}_0$. Let $P_2(\lambda)$ be the characteristic polynomial of the matrix of second derivatives at $\mathbf{r}_0$. Let λ_i, $i = 1, 2$, be the roots of $P_2(\lambda)$. Then

- *If the roots are strictly positive, $\lambda_i > 0$, then f has a local minimum at $\mathbf{r}_0$.*
- *If the roots are strictly negative, $\lambda_i < 0$, then f has a local maximum at $\mathbf{r}_0$.*
- *If the roots do not vanish but have different signs, then $\mathbf{r}_0$ is a saddle point of f.*
- *If at least one of the roots vanishes, then f may have a local maximum, a local minimum, a saddle, or none of the above (the second-derivative test is inconclusive).*

PROOF. Consider the second directional derivative of f at any point $\mathbf{r}$ in a neighborhood N_δ of $\mathbf{r}_0$:

$$D^2_{\mathbf{u}} f = D_{\mathbf{u}}(f'_x u_1 + f'_y u_2) = f''_{xx} u_1^2 + 2 f''_{xy} u_1 u_2 + f''_{yy} u_2^2,$$

which is a quadratic function in components of the vector $\hat{\mathbf{u}}$. It determines the concavity of the curve obtained by the intersection of the graph of f with a plane parallel to both the z axis and $\hat{\mathbf{u}}$ and going through the point $\mathbf{r} = (x, y)$. Hence, if $D^2_{\mathbf{u}} f > 0$ for all $\mathbf{r}$ in N_δ, then the graph is concave downward and f must have a local minimum at $\mathbf{r}_0$. Similarly, f has a local maximum at $\mathbf{r}_0$ if $D^2_{\mathbf{u}} f < 0$ for all $\mathbf{r}$ in N_δ.

For any unit vector, $u_1 = \cos\varphi$ and $u_2 = \sin\varphi$. Making use of the double-angle formulas $\cos^2\varphi = (1+\cos(2\varphi)/2$, $\sin^2\varphi = (1-\cos(2\varphi)/2$, and $2\sin\varphi\cos\varphi = \sin(2\varphi)$, the second directional derivative can be written in the form

$$2D^2_{\mathbf{u}}f = (f''_{xx} + f''_{yy}) + (f''_{xx} - f''_{yy})\cos(2\varphi) + 2f''_{xy}\sin(2\varphi).$$

Put $A^2 = (f''_{xx} - f''_{yy})^2 + 4(f''_{xy})^2$ and define an angle α by $\cos\alpha = (f''_{xx} - f''_{yy})/A$ and $\sin\alpha = -2f''_{xy}/A$. Since $\cos(2\varphi+\alpha) = \cos\alpha\cos(2\varphi) - \sin\alpha\sin(2\varphi)$, one infers that

$$2D^2_{\mathbf{u}}f = (f''_{xx} + f''_{yy}) + A\cos(2\varphi+\alpha) = (f''_{xx} + f''_{yy}) + A\cos\phi,$$

where $\phi = 2\varphi + \alpha$ takes values in $[0, 2\pi]$ for all $\hat{\mathbf{u}}$. Define λ_1 and λ_2 by the relations

$$\lambda_1 + \lambda_2 = f''_{xx} + f''_{yy}\,, \quad \lambda_1\lambda_2 = f''_{xx}f''_{yy} - (f''_{xy})^2.$$

Note that if $\mathbf{r} = \mathbf{r}_0$, then λ_1 and λ_2 are roots of the characteristic polynomial $P_2(\lambda) = 0$ as they satisfy the conditions $\lambda_1 + \lambda_2 = a + b$ and $\lambda_1\lambda_2 = ab - c^2$. By the continuity of the second derivatives, λ_1 and λ_2 are continuous as well. Hence, if the roots of $P_2(\lambda)$ do not vanish, then λ_1 and λ_2 do not vanish in some neighborhood N_δ. Then it follows that $(\lambda_1 - \lambda_2)^2 = (\lambda_1 + \lambda_2)^2 - 4\lambda_1\lambda_2 = A^2$ or $A = |\lambda_1 - \lambda_2|$, and the second derivative becomes

$$\begin{aligned} D^2_{\mathbf{u}}f &= \frac{\lambda_1 + \lambda_2}{2} + \frac{|\lambda_1 - \lambda_2|}{2}\cos\phi \\ &= \begin{cases} \lambda_1\cos^2(\phi/2) + \lambda_2\sin^2(\phi/2),\ \lambda_1 \geq \lambda_2, \\ \lambda_1\sin^2(\phi/2) + \lambda_2\cos^2(\phi/2),\ \lambda_1 < \lambda_2, \end{cases} \end{aligned}$$

where $(1+\cos\phi) = 2\cos^2\phi$ and $(1-\cos\phi) = 2\sin^2\phi$ have been used. Suppose that $\lambda_{1,2} > 0$ at the critical point $\mathbf{r}_0$. Then, by the continuity of the second derivatives, $\lambda_{1,2} > 0$ in some neighborhood N_δ of $\mathbf{r}_0$. Thus, in this case, $D^2_{\mathbf{u}}f > 0$ in N_δ and f has a local minimum at $\mathbf{r}_0$. Similarly, if the roots are strictly negative, then $D^2_{\mathbf{u}}f < 0$ in N_δ and f has a local maximum at $\mathbf{r}_0$. If $\lambda_{1,2} \neq 0$ but have different signs, then $D^2_{\mathbf{u}}f$ changes its sign in N_δ. Since the sign of $D^2_{\mathbf{u}}$ does not change when $\hat{\mathbf{u}} \to -\hat{\mathbf{u}}$, on any straight line through $\mathbf{r}_0$ and parallel to $\hat{\mathbf{u}}$, f has a fixed concavity along each line, which means that f has a saddle point at $\mathbf{r}_0$.

The inconclusiveness of the second-derivative test when at least one of the roots vanishes is easily established by specific examples.

Consider the function $f(x, y) = x^2 + sy^4$, where s is a number. It has a critical point $(0, 0)$ because $f'_x(0, 0) = f'_y(0, 0) = 0$ and $a = f''_{xx}(0, 0) = 2$, $b = f''_{yy}(0, 0) = 0$, and $c = f''_{xy}(0, 0) = 0$. Therefore,

$P_2(\lambda) = -(2-\lambda)\lambda$ has the roots $\lambda_1 = 2$ and $\lambda_2 = 0$. If $s > 0$, then $f(x, y) \leq 0$ for all (x, y) and f has a minimum at $(0, 0)$. If $s < 0$, the function f has a minimum along the line $x = t$, $y = 0$ ($F(t) = t^2$), while it has a maximum along the line $x = 0$, $y = t$ ($F(t) = st^4$, $s < 0$); that is, $(0, 0)$ is a saddle point. The function $f(x, y) = -(x^2 + sy^4)$ has a maximum at $(0, 0)$ if $s > 0$, and if $s < 0$, the critical point $(0, 0)$ is a saddle point. So, if one of the roots vanishes, then f may have a local maximum or a local minimum, or a saddle. The same conclusion is reached when $\lambda_1 = \lambda_2 = 0$ by studying the functions $f(x, y) = \pm(x^4 + sy^4)$ along the similar lines of arguments.

Furthermore, consider the function $f(x, y) = xy^2$. It also has a critical point at the origin, and all its second derivatives vanish at $(0, 0)$, that is, $P_2(\lambda) = \lambda^2$ and $\lambda_1 = \lambda_2 = 0$. The values of f along any line through the origin $x = u_1 t$, $y = u_2 t$ are $F(t) = st^3$, where $s = u_1 u_2^2$. For any $s \neq 0$, $F(t)$ has an *inflection* point at $t = 0$. Therefore, f cannot have a maximum or minimum or saddle at $(0, 0)$ because in any of these situations f should have either a minimum or a maximum along any straight line. A critical point may be a general inflection point if one the of roots does not vanish. An example is provided by $f(x, y) = x^2 + y^3$, which has a critical point $(0, 0)$ and $a = 2$, $b = 0$, and $c = 0$, that is $\lambda_1 = 2$ and $\lambda_2 = 0$. There are lines through the origin along which f has either a minimum or an inflection at $t = 0$. For example, along the line $x = 0$, $y = t$, the function f has an inflection point ($F(t) = t^3$) at $t = 0$, whereas, along the line $x = t$, $y = 0$, it has a minimum ($F(t) = t^2$). The very existence of lines along which f has an inflection precludes us from concluding that f can have a maximum or a minimum or a saddle at $(0, 0)$.

This concludes the proof of the second-derivative test in the case of two-variable functions. □

EXAMPLE 13.28. *Suppose that $\lambda_1 = 0$ and $\lambda_2 < 0$ for $f(x, y)$ at its critical point. Find the directions at which $D^2_{\mathbf{u}} f$ vanishes at the critical point.*

SOLUTION: Put $\hat{\mathbf{u}} = (\cos\varphi, \sin\varphi)$. Then $D^2_{\mathbf{u}} f = \lambda_2 \sin^2(\theta + \alpha/2) = 0$ or $\theta = -\alpha/2$ and $\theta = -\alpha/2 + \pi$, where $\alpha = -\sin^{-1}(2c/A)$, $A = \sqrt{(a-b)^2 + 4c^2} = |\lambda_1 - \lambda_2| = |\lambda_2|$ (see the proof of the second-derivative test). Therefore, the directions are $\hat{\mathbf{u}} = \pm(\cos(\alpha/2), -\sin(\alpha/2))$. □

EXAMPLE 13.29. *Find all critical points of the function $f(x, y) = \frac{1}{3}x^3 + xy^2 - x^2 - y^2$ and determine whether f has a local maximum, minimum, or saddle at them.*

Solution: Critical points. The function is a polynomial, and therefore it has partial derivatives everywhere. So its critical points are solutions of the system of equations

$$\begin{cases} f'_x = x^2 + y^2 - 2x = 0 \\ f'_y = 2xy - 2y \qquad = 0 \end{cases}$$

It is important not to lose solutions when transforming the system of equations $\nabla f(\mathbf{r}) = \mathbf{0}$ *for the critical points.* It follows from the second equation that $y = 0$ or $x = 2$. Therefore, the original system of equations is *equivalent* to two systems of equations:

$$\begin{cases} f'_x = x^2 + y^2 - 2x = 0 \\ x \ = 1 \end{cases} \quad \text{or} \quad \begin{cases} f'_x = x^2 + y^2 - 2x = 0 \\ y \ = 0 \end{cases}.$$

Solutions of the first system are $(1, 1)$ and $(1, -1)$. Solutions of the second system are $(0, 0)$ and $(2, 0)$. Thus, the function has four critical points. *It is advisable to check if all points found do satisfy the original system because, when transforming a system of nonlinear equations, one might get points that do not satisfy the original system or one might simply make an error.*

Second-derivative test. The second derivatives are

$$f''_{xx} = 2x - 2\,, \quad f''_{yy} = 2x - 2\,, \quad f''_{xy} = 2y.$$

For the points $(1, \pm 1)$, $a = b = 0$ and $c = \pm 2$. The characteristic polynomial is $P_2(\lambda) = \lambda^2 - 4$. Its roots $\lambda = \pm 2$ do not vanish and have opposite signs. Therefore, the function has a saddle at the points $(1, \pm 1)$. For the point $(0, 0)$, $a = b = -2$ and $c = 0$. The characteristic polynomial is $P_2(\lambda) = (-2 - \lambda)^2$. It has one root of multiplicity 2, that is, $\lambda_1 = \lambda_2 = -2 < 0$, and f has a local maximum at $(0, 0)$. Finally, for the point $(2, 0)$, $a = b = 2$ and $c = 0$. The characteristic polynomial $P_2(\lambda) = (2 - \lambda)^2$ has one root of multiplicity 2, $\lambda_1 = \lambda_2 = 2 > 0$; that is, the function has a local minimum at $(2, 0)$. □

94.3. Study Problems.

Problem 13.10. *Define* $f(0, 0) = 0$ *and*

$$f(x, y) = x^2 + y^2 - 2x^2y - \frac{4x^6y^2}{(x^4 + y^2)^2}$$

if $(x, y) \neq (0, 0)$*. Show that, for all* (x, y)*, the following inequality holds:* $4x^4y^2 \leq (x^4 + y^2)^2$*. Use it and the squeeze principle to conclude that* f *is continuous. Next, consider a line through* $(0, 0)$ *and parallel to* $\hat{\mathbf{u}} = (\cos\varphi, \sin\varphi)$ *and the values of* f *on it:*

$$F_\varphi(t) = f(t\cos\varphi, t\sin\varphi).$$

Show that $F_\varphi(0) = 0$, $F'_\varphi(0) = 0$, *and* $F''_\varphi(0) = 2$ *for all* $0 \le \varphi \le 2\pi$. *Thus,* f *has a minimum at* $(0, 0)$ *along any straight line through* $(0, 0)$. *Show that nevertheless* f *has no minimum at* $(0, 0)$ *by studying its value along the parabolic curve* $(x, y) = (t, t^2)$.

Solution: One has $0 \ge (a-b)^2 = a^2-2ab+b^2$ and hence $2ab \le a^2+b^2$ for any numbers a and b. Therefore, $4ab = 2ab+2ab \le 2ab+a^2+b^2 = (a+b)^2$. By setting $a = x^4$ and $b = y^2$, the said inequality is established. The continuity of the last term in f at $(0, 0)$ has to be verified. By the found inequality,

$$\frac{4x^6y^2}{(x^4+y^2)^2} \le \frac{4x^6y^2}{4x^4y^2} = x^2 \to 0 \quad \text{as} \quad (x, y) \to (0, 0).$$

Thus, $f(x, y) \to f(0, 0) = 0$ as $(x, y) \to (0, 0)$, and f is continuous everywhere. If $\varphi = \pm\pi/2$, that is, the line coincides with the x axis, $(x, y) = (t, 0)$, one has $F_\varphi(t) = t^2$, from which it follows that $F_\varphi(0) = F'_\varphi(0) = 0$ and $F''_\varphi(0) = 2$. When $\varphi \ne \pm\pi/2$ so that $\sin\varphi \ne 0$, one has

$$F_\varphi(t) = t^2 + at^3 + \frac{bt^4}{(1+ct^2)^2},$$

$$a = -2\cos^2\varphi\sin\varphi, \quad b = -\frac{4\cos^6\varphi}{\sin^2\varphi}, \quad c = \frac{\cos^4\varphi}{\sin^2\varphi}.$$

A straightforward differentiation shows that $F_\varphi(0) = F'_\varphi(0) = 0$ and $F''_\varphi(0) = 2$ as stated, and $F_\varphi(t)$ has an absolute minimum at $t = 0$, or f attains an absolute minimum at $(0, 0)$ along any straight line through $(0, 0)$. Nevertheless, the latter *does not imply that* f *has a minimum at* $(0, 0)$! Indeed, along the parabola $(x, y) = (t, t^2)$, the function f behaves as

$$f(t, t^2) = -t^4,$$

which attains an *absolute maximum* at $t = 0$. Thus, along the parabola, f has a maximum value at the origin and hence cannot have a local minimum there. The problem illustrates the remark given earlier in this section. □

95. Maximum and Minimum Values (Continued)

95.1. Second-Derivative Test for Multivariable Functions. In the proof of the second-derivative test, it has been established that

$$D^f_{\mathbf{u}} = f''_{xx}u_1^2 + 2f''_{xy}u_1u_2 + f''_{yy}u_2^2 = \lambda_1 v_1^2 + \lambda_2 v_2^2, \tag{13.17}$$

where $\hat{\mathbf{u}} = (u_1, u_2) = (\cos\varphi, \sin\varphi)$ and $\hat{\mathbf{v}} = (v_1, v_2) = (\cos\beta, \sin\beta)$ are unit vectors such that $\beta = \phi/2 = \varphi + \alpha/2$. This fact has a remarkable geometrical interpretation. Note that $D^2_{\mathbf{u}} f$ is a quadratic function in components of $\hat{\mathbf{u}}$. When discussing the shapes of quadric surfaces, in particular, quadric cylinders, it has been shown that, by a suitable rotation of the coordinate axes, the "mixed" term $2f''_{xy}u_1u_2$ can be eliminated. When the coordinate system is rotated, the angle φ is simply shifted by the rotation angle. So the vector $\hat{\mathbf{v}}$ is obtained from the vector $\hat{\mathbf{u}}$ by rotating the latter through the angle $\alpha/2$, which is determined by the second-order partial derivatives of f. In doing so, *the quadratic function $D^2_{\mathbf{u}} f$ is brought into the standard form in which the coefficients are determined by the roots of the characteristic polynomial.* This result holds for any number of variables.

First, note that $D_{\mathbf{u}}(D_{\mathbf{u}} f) = u_1 D_{\mathbf{u}} f'_{x_1} + u_2 D_{\mathbf{u}} f'_{x_2} + \cdots + u_m D_{\mathbf{u}} f'_{x_m}$. Put $D_{ij} = f''_{x_i x_j} = D_{ji}$ (by Clairaut's theorem). Then

$$D^2_{\mathbf{u}} f = \sum_{i=1}^{m} \sum_{j=1}^{m} D_{ij} u_i u_j,$$

which is a quadratic function. The numbers D_{ij} can be arranged into a square $m \times m$ matrix. The polynomial of degree m,

$$P_m(\lambda) = \det \begin{pmatrix} D_{11} - \lambda & D_{12} & D_{13} & \cdots & D_{1m} \\ D_{21} & D_{22} - \lambda & D_{23} & \cdots & D_{2m} \\ D_{31} & D_{32} & D_{33} - \lambda & \cdots & D_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D_{m1} & D_{m2} & D_{m3} & \cdots & D_{mm} - \lambda \end{pmatrix},$$

is called the characteristic polynomial of the matrix of second derivatives. *For any symmetric real matrix ($D_{ij} = D_{ji}$), the roots of its characteristic polynomial are proved to be real.* It can further be proved that there exists a rotation of the coordinate system under which $\hat{\mathbf{u}}$ goes into $\hat{\mathbf{v}}$ such that

$$D^2_{\mathbf{u}} f = \sum_{i=1}^{m} \sum_{j=1}^{m} D_{ij} u_i u_j = \lambda_1 v_1^2 + \lambda_2 v_2^2 + \cdots + \lambda_m v_m^2, \tag{13.18}$$

where λ_i are roots of P_m. This equation shows that the second-derivative test formulated in the preceding section holds for any number of variables.

THEOREM 13.18. (Second-Derivative Test for m Variables). *Let $\mathbf{r}_0$ be a critical point of f and suppose that f has continuous second-order partial derivatives D_{ij} in some open ball centered at $\mathbf{r}_0$. Then the* *characteristic polynomial $P_m(\lambda)$ of the matrix $D_{ij}(\mathbf{r}_0)$ has real roots λ_i, $i = 1, 2, ..., m$, and*

- *If all the roots are strictly positive, $\lambda_i > 0$, then f has a local minimum.*
- *If all the roots are strictly negative, $\lambda_i < 0$, then f has a local maximum.*
- *If all the roots do not vanish but have different signs, then f has an m-dimensional saddle point.,*
- *If some of the roots vanish, then f may have a local maximum, a local minimum, a saddle, or none of the above (the test is inconclusive).*

In the two-variable case, the proof uses special properties of the roots of quadratic polynomials. In the general case, the goal is achieved by means of *linear algebra* methods.

Remark. If at least one of the roots of the characteristic polynomials vanishes, the second-derivative test is inconclusive. How can the local behavior of a function be analyzed near its critical point? If the function in question is differentiable sufficiently many times near a critical point, then the Taylor polynomial approximation at the critical point provides a useful technique for answering this question because it is easier to study a polynomial rather than a general function.

Example 13.30. *Investigate the local behavior of the function defined by $f(0, y) = f(x, 0) = 1$ and $f(x, y) = \sin(xy)/(xy)$ if $x \neq 0$ and $y \neq 0$.*

Solution: Since $u = xy$ is small near the origin, $\sin u$ can be approximated by its Taylor polynomial $T_3(u) = u - u^3/6$. Hence, the corresponding Taylor polynomial $T_n^f(x, y)$ for the function f at $(0, 0)$ reads:

$$T_n^f(x,y) = \frac{T_3(u)}{u} = 1 - \frac{u^2}{6} = 1 - \frac{x^2y^2}{6}.$$

So the function attains a local maximum at $(0, 0)$ because $x^2y^2 \geq 0$ for all (x, y). It is worth noting that the first nonconstant polynomial has degree $n = 4$ and therefore $T_2^f = 1$, which means that all the first and second derivatives of f vanish at $(0, 0)$; that is, the characteristic polynomial for the matrix of second derivatives is $P_2(\lambda) = \lambda^2$ has root $\lambda = 0$ of multiplicity 2. The second-derivative test would be inconclusive. □

95.2. Absolute Maximal and Minimal Values. For a function f of one variable, the extreme value theorem says that if f is continuous on a closed interval $[a, b]$, then f has an absolute minimum value and an absolute maximum value. For example, the function $f(x) = x^2$ on $[-1, 2]$ attains an absolute minimum value at $x = 0$ and an absolute maximum value at $x = 2$. Note that f is defined for all x, and therefore its critical points are determined by $f'(x) = 0$. So the absolute minimum value occurs at the critical point $x = 0$ inside the interval, while the absolute maximum value occurs on the boundary of the interval that is not a critical point of f. Thus, to find the absolute maximum and minimum values of a function f in a closed interval in the domain of f, the values of f must be evaluated not only at the critical points but also at the boundaries of the interval.

The situation for multivariable functions is similar.

Definition 13.26. (Closed Set).
A set D in a Euclidean space is said to be closed if it contains all its limit points.

Recall that any neighborhood of a limit point of D contains points of D. If a limit point of D is not an interior point of D, then it lies on a boundary of D. So a closed set contains its boundaries. All points of an open interval (a, b) are its limit points, but, in addition, the boundaries a and b are also its limit points, so when they are added, a closed set $[a, b]$ is obtained. Similarly, the set in the plane $D\{(x, y)|x^2 + y^2 < 1\}$ has limit points on the circle $x^2 + y^2 = 1$ (the boundary of D), which is not in D. By adding these points, a closed set is obtained, $x^2 + y^2 \leq 1$.

Definition 13.27. (Bounded Set).
A set D in a Euclidean space is said to be bounded if it is contained in some ball.

In other words, for any two points in a bounded set, the distance between them cannot exceed some value (the diameter of the ball that contains the set).

Theorem 13.19. (Extreme Value Theorem).
If f is continuous on a closed, bounded set D in a Euclidean space, then f attains an absolute maximum value $f(\mathbf{r}_1)$ and an absolute minimum value $f(\mathbf{r}_2)$ at some points $\mathbf{r}_1 \in D$ and $\mathbf{r}_2 \in D$.

By this theorem, it follows that the points $\mathbf{r}_1$ and $\mathbf{r}_2$ are either critical points of f (because a local maximum or minimum always occurs at a critical point) or lie on the boundary of D. So, to find the absolute minimum and maximum values of a continuous function f on a closed, bounded set D, one has to

(1) Find the values of f at the critical points of f in D.
(2) Find the extreme values of f on the boundary of D.
(3) The largest of the values obtained in Steps 1 and 2 is the absolute maximum value, and the smallest of these values is the absolute minimum value.

EXAMPLE 13.31. *Find the absolute maximum and minimum values of $f(x, y) = x^2 + y^2 + xy$ on the disk $x^2 + y^2 \leq 4$ and the points at which they occur.*

SOLUTION: The function f is a polynomial. It is continuous and differentiable on the whole plane.

Step 1. Critical points of f satisfy the system of equations $f'_x = 2x+y = 0$ and $f'_y = 2y + x = 0$; that is, $(0, 0)$ is the only critical point of f and it happens to be in the disk. The value of f at the critical point is $f(0, 0) = 0$.
Step 2. The boundary of the disk is the circle $x^2 + y^2 = 4$. To find the extreme values of f on it, take the parametric equations of the circle $x(t) = 2\cos t$, $y(t) = 2\sin t$, where $t \in [0, 2\pi]$. One has $F(t) = f(x(t), y(t)) = 4 + 4\cos t \sin t = 4 + 2\sin(2t)$. The function $F(t)$ attains its maximal value 6 on $[0, 2\pi]$ when $\sin(2t) = 1$ or $t = \pi/4$ and $t = \pi/4 + \pi$. These values of t correspond to the points $(\sqrt{2}, \sqrt{2})$ and $(-\sqrt{2}, -\sqrt{2})$. Similarly, $F(t)$ attains its minimal value 2 on $[0, 2\pi]$ when $\sin(2t) = -1$ or $t = 3\pi/4$ and $t = 3\pi/4 + \pi$. These values of t correspond to the points $(-\sqrt{2}, \sqrt{2})$ and $(\sqrt{2}, -\sqrt{2})$.
Step 3. The largest number of 0, 2, and 6 is 6. So the absolute minimum value of f is 6; it occurs at the points $(\sqrt{2}, \sqrt{2})$ and $(-\sqrt{2}, -\sqrt{2})$. The smallest number of 0, 2, and 6 is 0. So the absolute minimum value of f is 0; it occurs at the point $(0, 0)$. □

EXAMPLE 13.32. *Find the absolute maximum and minimum values of $f(x, y, z) = x^2+y^2-z^2+2z$ on the closed set $D = \{(x, y, z) \mid x^2+y^2 \leq z \leq 4\}$.*

SOLUTION: The set D is the solid bounded from below by the paraboloid $z = x^2 + y^2$ and from the top by the plane $z = 4$. It is a bounded set, and f is continuous on it as any polynomial function.

Step 1. Since f is differentiable everywhere, its critical points satisfy the equations $f'_x = 2x = 0$, $f'_y = 2y = 0$, and $f'_z = -2z + 2 = 0$. There is only one critical point $(0, 0, 1)$, and it happens to be in D. The value of f at it is $f(0, 0, 1) = 1$.
Step 2. The boundary consists of two surfaces, the disk $S_1 = \{(x, y, z) \,|\, z = 4,\ x^2 + y^2 \le 4\}$ in the plane $z = 4$ and the portion of the paraboloid $S_2 = \{(x, y, z) \,|\, z = x^2 + y^2,\ x^2 + y^2 \le 4\}$. The values of f on S_1 are $F_1(x, y) = f(x, y, 4)$, where the points (x, y) lie in the disk of radius 2, $x^2 + y^2 \le 4$. The problem now is to find the maximal and minimal values of a two-variable function F_1 on the disk. In principle, at this point, Steps 1, 2, and 3 have to be applied to F_1. These technicalities can be avoided in this particular case by noting that $F_1(x, y) = x^2 + y^2 - 8 = r^2 - 8$, where $r^2 = x^2 + y^2 \le 4$. Therefore, the maximal value of F_1 is reached when $r^2 = 4$, and its minimal value is reached when $r^2 = 0$. So the maximal and minimal values of f on S_1 are -4 and -8. The values of f on S_2 are $F_2(x, y) = f(x, y, x^2 + y^2) = 3r^2 - r^4 = g(r)$, where $r^2 = x^2 + y^2 \le 4$ or $r \in [0, 2]$. The critical points of $g(r)$ satisfy the equation $g'(r) = 6r - 4r^3 = 0$ whose solutions are $r = 0$, $r = \pm\sqrt{3/2}$. Therefore, the maximal value of f on S_2 is $9/4$, which is the largest of $g(0) = 0$, $g(\sqrt{3/2}) = 9/4$, and $g(2) = -4$, and the minimal value is -4 as the smallest of these numbers.
Step 3. The absolute maximum value of f on D is $\max\{1, -8, -4, 9/4\} = 9/4$, and the absolute minimum value of f on D is $\min\{1, -8, -4, 9/4\} = -8$. Both values occur on the boundary of D: $f(0, 0, 4) = -8$ and the absolute maximal value is attained along the circle of intersection of the plane $z = 3/2$ with the paraboloid $z = x^2 + y^2$. □

96. Lagrange Multipliers

Let $f(x, y)$ be the height of a hill as a function of position. A hiker walks along a path $\mathbf{r}(t) = (x(t), y(t))$. What are the local maxima and minima along the path? What are the maximum and minimum heights along the path? These questions are easy to answer if the parametric equations of the path are explicitly known. Indeed, the height along the path is the single-variable function $F(t) = f(\mathbf{r}(t))$ and the problem is reduced to the standard extreme value problem for $F(t)$ on an interval $t \in [a, b]$.

Example 13.33. *The height as a function of position is $f(x, y) = xy$. Find the local maxima and minima of the height along the circular path $x^2 + y^2 = 4$.*

Solution: The parametric equation of the circle can be taken in the form $\mathbf{r}(t) = (2\cos t, 2\sin t)$, where $t \in [0, 2\pi]$. The height a long the path is $F(t) = 4\cos t\ \sin t = 2\sin(2t)$. On the interval $[0, 2\pi]$, the function $\sin(2t)$ attains its absolute maximum value at $t = \pi/4$ and $t = \pi/4 + \pi$ and its absolute minimum value at $t = 3\pi/4$ and $t = 3\pi/4 + \pi$. So, along the path, the function f attains the absolute maximum value 2 at $(\sqrt{2}, \sqrt{2})$ and $(-\sqrt{2}, -\sqrt{2})$ and the absolute minimum value -2 at $(-\sqrt{2}, \sqrt{2})$ and $(\sqrt{2}, -\sqrt{2})$. □

However, in many similar questions, an explicit form of $\mathbf{r}(t)$ is not known or not easy to find. An algebraic condition $g(x, y) = 0$ is a more general way to describe a curve. It simply says that only the points (x, y) that satisfy this condition are permitted in the argument of f; that is, the variables x and y are no longer independent. The condition $g(x, y) = 0$ is called a *constraint*.

Problems of this type occur for functions of more than two variables. For example, let $f(x, y, z)$ be the temperature as a function of position. A reasonable question to ask is: What are the maximum and minimum temperatures on a surface? A surface may be described by imposing one constraint $g(x, y, z) = 0$ on the variables x, y, and z. Nothing precludes us from asking about the maximum and minimum temperatures along a curve defined as an intersection of two surfaces $g_1(x, y, z) = 0$ and $g_2(x, y, z) = 0$. So the variables x, y, and z are now subject to two constraints. In general, what are the extreme values of a multivariable function $f(\mathbf{r})$ whose arguments are subject to several constraints $g_a(\mathbf{r}) = 0$, $a = 1, 2, ..., M$? Naturally, the number of independent constraints should not exceed the number of variables.

Definition 13.28. (Local Maxima and Minima Subject to Constraints). *A function $f(\mathbf{r})$ has a local maximum (or minimum) at $\mathbf{r}_0$ on the set defined by the constraints $g_a(\mathbf{r}) = 0$ if $f(\mathbf{r}) \leq f(\mathbf{r}_0)$ (or $f(\mathbf{r}) \geq f(\mathbf{r}_0)$) for all $\mathbf{r}$ in some neighborhood of $\mathbf{r}_0$ that satisfy the constraints, that is, $g_a(\mathbf{r}) = 0$.*

Note that a function f may not have local maxima or minima in its domain. However, when its arguments become subject to constraints, it may well have local maxima and minima on the set defined by the constraints. In the example considered, $f(x, y) = xy$ has no local maxima or minima, but, when it is restricted on the circle by imposing the constraint $g(x, y) = x^2 + y^2 - 4 = 0$, it happens to have two local minima and maxima.

96.1. Critical Points of a Function Subject to a Constraint. The extreme value problem with constraints amounts to finding the critical points of a function whose arguments are subject to constraints. The example discussed above shows that the equation $\nabla f = \mathbf{0}$ no longer determines the critical points for differentiable functions if its arguments are constrained.

Consider first the case of a single constraint for two variables $\mathbf{r} = (x, y)$. Suppose the function f and the function g that define the constraint are differentiable. Let $\mathbf{r}_0$ be a point at which $f(\mathbf{r})$ has a local maximum or minimum on the set S defined by the constraint $g(\mathbf{r}) = 0$, which is a curve in the two-variable case. Let $\mathbf{r}(t)$ be parametric equations of this curve in a neighborhood of $\mathbf{r}_0$, that is, for some $t = t_0$, $\mathbf{r}(t_0) = \mathbf{r}_0$. Assuming that $\mathbf{r}(t)$ is differentiable, it is concluded that $F'(t_0) = 0$, where $F(t) = f(\mathbf{r}(t))$ are values of f along the curve. The chain rule yields

$$F'(t_0) = f'_x(\mathbf{r}_0)x'(t_0) + f'_y(\mathbf{r}_0)y'(t_0) = \nabla f(\mathbf{r}_0) \cdot \mathbf{r}'(t_0) = 0 \implies \nabla f(\mathbf{r}_0) \perp \mathbf{r}'(t_0)\,.$$

The gradient $\nabla f(\mathbf{r}_0)$ is orthogonal to a tangent vector to the curve at *the point where f has a local maximum or minimum on the curve.* On the other hand, the gradient $\nabla g(\mathbf{r})$ at *any* point is normal to the level curve $g(\mathbf{r}) = 0$, that is, $\nabla g(\mathbf{r}(t)) \perp \mathbf{r}'(t)$ for any t. Therefore, the gradients $\nabla f(\mathbf{r}_0)$ and $\nabla g(\mathbf{r}_0)$ must be parallel at $\mathbf{r}_0$. This geometrical statement can be translated into an algebraic one: there should exist a number λ such that $\nabla f(\mathbf{r}_0) = \lambda \nabla g(\mathbf{r}_0)$. This proves the following theorem.

THEOREM 13.20. (Critical Points Subject to a Constraint). *Suppose that f and g are differentiable at $\mathbf{r}_0$ and f has a local maximum or minimum at $\mathbf{r}_0$ in the set defined by the constraint $g(\mathbf{r}) = 0$. Then there exists a number λ such that*

$$\nabla f(\mathbf{r}_0) = \lambda \nabla g(\mathbf{r}_0).$$

The theorem holds for three-variable functions as well. Indeed, if $\mathbf{r}(t)$ is a curve through $\mathbf{r}_0$ in the level surface $g(x, y, z) = 0$. Then the derivative $F'(t) = (d/dt) f(\mathbf{r}(t)) = f'_x x' + f'_y y' + f'_z z' = \nabla f \cdot \mathbf{r}'$ must vanish at t_0, that is, $F'(t_0) = \nabla f(\mathbf{r}_0) \cdot \mathbf{r}'(t_0) = 0$. Therefore, $\nabla f(\mathbf{r}_0)$ is orthogonal to a tangent vector of *any* curve in the surface S at $\mathbf{r}_0$, and hence $\nabla f(\mathbf{r}_0)$ is normal to the tangent plane to S through $\mathbf{r}_0$. On the other hand, the gradient ∇g is normal to the tangent plane to a level surface of g at any point (see the properties of the gradient). Therefore, at the point $\mathbf{r}_0$, the gradients of f and g must be parallel. A similar line of reasoning proves the theorem for any number of variables.

This theorem provides a powerful method to find the critical points of f subject to a constraint $g = 0$ if f and g are differentiable. It is called the *method of Lagrange multipliers.* To find the critical points of f, the following system of equations must be solved:

$$\nabla f(\mathbf{r}) = \lambda \nabla g(\mathbf{r})\,, \quad g(\mathbf{r}) = 0. \tag{13.19}$$

If $\mathbf{r} = (x, y)$, this is a system of three equations, $f'_x = \lambda g'_x$, $f'_y = \lambda g'_y$, and $g = 0$ for three variables (x, y, λ). For each solution (x_0, y_0, λ_0), the corresponding critical point of f is (x_0, y_0). The numerical value of λ is not relevant; only its existence must be established by solving the system. In the three-variable case, the system contains four equations for four variables (x, y, z, λ). For each solution $(x_0, y_0, z_0, \lambda_0)$, the corresponding critical point of f is (x_0, y_0, z_0).

Example 13.34. *Use the method of Lagrange multipliers to solve the problem in Example 13.33.*

Solution: Put $g(x, y) = x^2 + y^2 - 4$. Then

$$\begin{cases} f'_x = \lambda g'_x \\ f'_y = \lambda g'_y \\ g \;\; = 0 \end{cases} \quad \Longrightarrow \quad \begin{cases} y = 2\lambda x \\ x = 2\lambda y \\ x^2 + y^2 = 4 \end{cases} .$$

The substitution of the first equation into the second one gives $x = 4\lambda^2 x$. This means that either $x = 0$ or $\lambda = \pm 1/2$. If $x = 0$, then $y = 0$ by the first equation, which contradicts the constraint. For $\lambda = 1/2$, $x = y$ and the constraint gives $2x^2 = 4$ or $x = \pm\sqrt{2}$. The critical points corresponding to $\lambda = 1/2$ are $(\sqrt{2}, \sqrt{2})$ and $(-\sqrt{2}, -\sqrt{2})$. If $\lambda = -1/2$, $x = -y$ and the constraint gives $2x^2 = 4$ or $x = \pm\sqrt{2}$. The critical points corresponding to $\lambda = -1/2$ are $(\sqrt{2}, -\sqrt{2})$ and $(-\sqrt{2}, \sqrt{2})$. So $f(\pm\sqrt{2}, \pm\sqrt{2}) = 2$ is the maximal value and $f(\mp\sqrt{2}, \pm\sqrt{2}) = -2$ is the minimal one. □

Example 13.35. *A rectangular box without a lid is to be made from cardboard. Find the dimensions of the box of a given volume V such that the cost of material is minimal.*

Solution: Let the dimensions be x, y, and z, where z is the height. The amount of cardboard needed is determined by the surface area $f(x, y, z) = xy + 2xz + 2yz$. The question is to find the minimal value of f subject to constraint $g(x, y, z) = xyz - V = 0$. The Lagrange multiplier method gives

$$\begin{cases} f'_x = \lambda g'_x \\ f'_y = \lambda g'_y \\ f'_z = \lambda g'_z \\ g \;\; = 0 \end{cases} \Longrightarrow \begin{cases} y + 2z = \lambda yz \\ x + 2z = \lambda xz \\ 2x + 2y = \lambda xy \\ xyz = V \end{cases} \Longrightarrow \begin{cases} xy + 2xz = \lambda V \\ xy + 2zy = \lambda V \\ 2xz + 2yz = \lambda V \\ xyz = V \end{cases},$$

where the last system has been obtained by multiplying the first equation by x, the second one by y, and the third one by z with the subsequent use of the constraint. Combining the first two equations, one infers $2z(y - x) = 0$. Since $z \neq 0$ ($V \neq 0$), one has $y = x$. Combining the first and third equations, one infers $y(x-2z) = 0$ and hence $x = 2z$. The substitution $y = x = 2z$ into the constraint yields $4z^3 = V$. Hence, the optimal dimensions are $x = y = (2V)^{1/3}$ and $z = (2V)^{1/3}/2$. The amount of cardboard minimizing the cost is $3(2V)^{2/3}$ (the value of f at the critical point). From the geometry of the problem, it is clear that f attains its minimum value at the only critical point. □

96.2. The Case of Two or More Constraints. Let f be a function of three variables subject to two constraints $g_1(\mathbf{r}) = 0$ and $g_2(\mathbf{r}) = 0$. Each constraint defines a surface in the domain of f (level surfaces of g_1 and g_2). So the set defined by the constraints is the curve of intersection of the level surfaces $g_1 = 0$ and $g_2 = 0$. Let $\mathbf{r}_0$ be a point of the curve at which f has a local maximum or minimum. Let $\mathbf{v}$ be a tangent vector to the curve at $\mathbf{r}_0$. Since the curve lies in the level surface $g_1 = 0$, by the earlier arguments, $\nabla f(\mathbf{r}_0)\perp\mathbf{v}$ and $\nabla g_1(\mathbf{r}_0)\perp\mathbf{v}$. On the other hand, the curve also lies in the level surface $g_2 = 0$ and hence $\nabla g_2(\mathbf{r}_0)\perp\mathbf{v}$. It follows that the gradients ∇f, ∇g_1, and ∇g_2 become *coplanar* at the point $\mathbf{r}_0$ as they lie in the plane normal to $\mathbf{v}$. Therefore, there exist numbers λ_1 and λ_2 such that

$$\nabla f(\mathbf{r}) = \lambda_1 \nabla g_1(\mathbf{r}) + \lambda_2 \nabla g_2(\mathbf{r}), \qquad g_1(\mathbf{r}) = g_2(\mathbf{r}) = 0$$

when $\mathbf{r} = \mathbf{r}_0$ (see Study Problem 11.6). This is a system of five equations for five variables $(x, y, z, \lambda_1, \lambda_2)$. For any solution $(x_0, y_0, z_0, \lambda_{10}, \lambda_{20})$, the point (x_0, y_0, z_0) is a critical point of f on the set defined by the constraints. In general, the following result can be proved by a similar line of reasoning.

THEOREM 13.21. (Critical Points Subject to Constraints). *Suppose that f and g_a, $a = 1, 2, ..., M$, are functions of m variables, $m > M$, which are differentiable at $\mathbf{r}_0$, and f has a local maximum or minimum at $\mathbf{r}_0$ in the set defined by the constraints $g_a(\mathbf{r}) = 0$. Then* *there exist numbers* λ_a *such that*

$$\nabla f(\mathbf{r}_0) = \lambda_1 \nabla g_1(\mathbf{r}_0) + \lambda_2 \nabla g_2(\mathbf{r}_0) + \cdots + \lambda_M \nabla g_M(\mathbf{r}_0).$$

Let $f(\mathbf{r})$ be a function subject to a constraint $g(\mathbf{r})$. Define the function

$$F(\mathbf{r}, \lambda) = f(\mathbf{r}) - \lambda g(\mathbf{r}),$$

where λ is viewed as an additional independent variable. Then critical points of F are determined by (13.19). Indeed, the condition $\partial F/\partial \lambda = 0$ yields the constraint $g(\mathbf{r}) = 0$, while the differentiation with respect to the variables $\mathbf{r}$ gives $\nabla F = \nabla f - \lambda \nabla g = 0$, which coincides with the first equation in (13.19). Similarly, if there are several constraints, critical points of the function with additional variables λ_a, $a = 1, 2, ..., M$,

$$F(\mathbf{r}, \lambda_1, \lambda_2, ..., \lambda_n) = f(\mathbf{r}) - \lambda_1 g_1(\mathbf{r}) - \lambda_2 g_2(\mathbf{r}) - \cdots - \lambda_M g_M(\mathbf{r}) \tag{13.20}$$

coincide with the critical points of f subject to the constraints $g_a = 0$ as stated in Theorem 13.21. The functions F and f have the same values on the set defined by the constraints $g_a = 0$ because they differ by a linear combination of constraint functions with the coefficients being the *Lagrange multipliers*.

96.3. Finding Local Maxima and Minima. In the simplest case of a two-variable function f subject to a constraint, the nature of critical points (local maximum or minimum) can be solved by geometrical means. Suppose that the level curve $g(x, y) = 0$ is closed. Then, by the extreme value theorem, f attains its maximum and minimum values on it at some of the critical points. Suppose f attains its absolute maximum at a critical point $\mathbf{r}_1$. Then f should have either a local minimum or an inflection at the neighboring critical point $\mathbf{r}_2$ along the curve. Let $\mathbf{r}_3$ be the critical point next to $\mathbf{r}_2$ along the curve. Then f has a local minimum at $\mathbf{r}_2$ if $f(\mathbf{r}_2) < f(\mathbf{r}_3)$ and an inflection if $f(\mathbf{r}_2) > f(\mathbf{r}_3)$. This procedure may be continued until all critical points are exhausted. Compare this pattern of critical points with the behavior of a height along a closed hiking path.

Remark. If the constraints can be solved, then an explicit form of f on the set defined by the constraints can be found, and the standard second-derivative test applies! For instance, in Example 13.35, the constraint can be solved $z = V/(xy)$. The values of the function f on the constraint surface are $F(x, y) = f(x, y, V/(xy)) = xy + 2V(x + y)/(xy)$. The equations $F'_x = 0$ and $F'_y = 0$ determine the critical point $x = y = (2V)^{1/3}$ (and $z = V/(xy) = (2V)^{1/3}/2$). So the second-derivative test can be applied to the function $F(x, y)$ at the critical point $x = y = (2V)^{1/3}$ to show that indeed F has a minimum and hence f has a minimum on the constraint surface.

There is an analog of the second-derivative test for critical points of functions subject to constraints. Its general formulation is not simple. So the discussion is limited to the simplest case of a function of two variables subject to a constraint.

Suppose that $\nabla g(\mathbf{r}_0) \neq \mathbf{0}$. Then g'_x and g'_y cannot simultaneously vanish at the critical point. Without loss of generality, assume that $g'_y \neq 0$ at $\mathbf{r}_0 = (x_0, y_0)$. By the implicit function theorem (Theorem 13.11), there is a neighborhood of $\mathbf{r}_0$ in which the equation $g(x, y) = 0$ has a unique solution $y = h(x)$. The values of f on the level curve $g = 0$ near the critical point are $F(x) = f(x, h(x))$. By the chain rule, one infers that $F' = f'_x + f'_y h'$ and

$$F'' = (d/dx)(f'_x + f'_y h') = f''_{xx} + 2f''_{xy}h' + f''_{yy}(h')^2 + f'_y h''. \tag{13.21}$$

So, in order to find $F''(x_0)$, one has to calculate $h'(x_0)$ and $h''(x_0)$. This task is accomplished by the implicit differentiation. By the definition of $h(x)$, $G(x) = g(x, h(x)) = 0$ for all x in an open interval containing x_0. Therefore, $G'(x) = 0$, which defines h' because $G' = g'_x + g'_y h' = 0$ and $h' = -g'_x/g'_y$. Similarly, $G''(x) = 0$ yields

$$G'' = g''_{xx} + 2g''_{xy}h' + g''_{yy}(h')^2 + g'_y h'' = 0, \tag{13.22}$$

which can be solved for h'', where $h' = -g'_x/g'_y$. The substitution of $h'(x_0)$, $h''(x_0)$, and all the values of all the partial derivatives of f at the critical point (x_0, y_0) into (13.21) gives the value $F''(x_0)$. If $F''(x_0) > 0$ (or $F''(x_0) < 0$), then f has a local minimum (or maximum) at (x_0, y_0) along the curve $g = 0$. Note also that $F'(x_0) = 0$ as required owing to the conditions $f'_x = \lambda g'_x$ and $f'_y = \lambda g'_y$ satisfied at the critical point.

If $g'_y(\mathbf{r}_0) = 0$, then $g'_x(\mathbf{r}_0) \neq 0$, and there is a function $x = h(y)$ that solves the equation $g(x, y) = 0$. So, by swapping x and y in the above arguments, the same conclusion is proved to hold.

EXAMPLE 13.36. *Show that the point $\mathbf{r}_0 = (0, 0)$ is a critical point of the function $f(x, y) = x^2y + y + x$ subject to the constraint $e^{xy} = x + y + 1$ and determine whether f has a local minimum or maximum at it.*

SOLUTION:
Critical point. Put $g(x, y) = e^{xy} - x - y - 1$. Then $g(0, 0) = 0$; that is, the point $(0, 0)$ satisfies the constraint. The first partial derivatives of f and g are $f'_x = 2xy + 1$, $f'_y = x^2 + 1$, $g'_x = ye^{xy} - 1$, and $g'_y = xe^{xy} - 1$.

Therefore, both equations $f'_x(0,0) = \lambda g'_x(0,0)$ and $f'_y(0,0) = \lambda g'_y(0,0)$ are satisfied at $\lambda = -1$. Thus, the point $(0,0)$ is a critical point of f subject to the constraint $g = 0$.

Second-derivative test. Since $g'_y(0,0) = -1 \neq 0$, there is a function $y = h(x)$ near $x = 0$ such that $G(x) = g(x, h(x)) = 0$. By the implicit differentiation,

$$h'(0) = -g'_x(0,0)/g'_y(0,0) = -1\,.$$

The second partial derivatives of g are

$$g''_{xx} = y^2 e^{xy}\,, \quad g''_{yy} = x^2 e^{xy}\,, \quad g''_{xy} = e^{xy} + xye^{xy}\,.$$

The derivative $h''(0)$ is found from (13.22), where $g''_{xx}(0,0) = g''_{yy}(0,0) = 0$, $g''_{xy}(0,0) = 1$, $h'(0) = -1$, and $g'_y(0,0) = -1$:

$$h''(0) = -[g''_{xx}(0,0) + 2g''_{xy}(0,0)h'(0) + g''_{yy}(0,0)(h'(0))^2]/g'_y(0,0) = -2.$$

The second partial derivatives of f are

$$f''_{xx} = 2y\,, \quad f''_{yy} = 0\,, \quad f''_{xy} = 2x.$$

The substitution of $f''_{xx}(0,0) = f''_{yy}(0,0) = f''_{xy}(0,0) = 0$, $h'(0) = -1$, $f'_y(0,0) = 1$, and $h''(0) = -2$ into (13.21) gives $F''(0) = -2 < 0$. Therefore, f attains a local maximum at $(0,0)$ along the curve $g = 0$. Note also that $F'(0) = f'_x(0,0) + f'_y(0,0)h'(0) = 1 - 1 = 0$ as required. □

The implicit differentiation and the implicit function theorem can be used to establish the second-derivative test for the multivariable case with constraints (see another example in Study Problem 13.11).

96.4. Study Problems.

Problem 13.11. *Let f be a twice continuously differentiable function of $\mathbf{r} = (x, y, z)$ subject to a constraint $g(\mathbf{r}) = 0$. Assume that g is a twice continuously differentiable function. Use the implicit differentiation to establish the second-derivative test for critical points of f on the surface $g = 0$.*

SOLUTION: Suppose that $\nabla g(\mathbf{r}_0) \neq \mathbf{0}$ at a critical point $\mathbf{r}_0$. Without loss of generality, one can assume that $g'_z(\mathbf{r}_0) \neq 0$. By the implicit function theorem, there exists a function $z = h(x, y)$ such that $G(x, y) = g(x, y, h(x, y)) = 0$ in some neighborhood of the critical point. Then the equations $G'_x(x, y) = 0$ and $G'_y(x, y) = 0$ determine the first partial derivatives of h:

$$g'_x + g'_z h'_x = 0 \implies h'_x = -g'_x/g'_z\,; \quad g'_y + g'_z h'_y = 0 \implies h'_y = -g'_y/g'_z\,.$$

The second partial derivatives of h are found from the equations

$$G''_{xx} = 0 \Longrightarrow g''_{xx} + 2g''_{xz}h'_x + g''_{zz}(h'_x)^2 + g'_z h''_{xx} = 0,$$
$$G''_{yy} = 0 \Longrightarrow g''_{yy} + 2g''_{yz}h'_y + g''_{zz}(h'_y)^2 + g'_z h''_{yy} = 0,$$
$$G''_{xy} = 0 \Longrightarrow g''_{xy} + g''_{xz}h'_x + g''_{yz}h'_y + g''_{zz}h'_x h'_y + g'_z h''_{xy} = 0.$$

The values of the function $f(x, y, z)$ of the level surface $g(x, y, z) = 0$ near the critical points are $F(x, y) = f(x, y, h(x, y))$. To apply the second-derivative test to the function F, its second partial derivatives have to be computed at the critical point. The implicit differentiation gives

$$F''_{xx} = (f'_x + f'_z h'_x)'_x = f''_{xx} + 2f''_{xz}h'_x + f''_{zz}(h'_x)^2 + f'_z h''_{xx},$$
$$F''_{yy} = (f'_y + f'_z h'_y)'_y = f''_{yy} + 2f''_{yz}h'_y + f''_{zz}(h'_y)^2 + f'_z h''_{yy},$$
$$F''_{xy} = (f'_x + f'_z h'_x)'_y = f''_{xy} + f''_{xz}h'_x + f''_{yz}h'_y + f''_{zz}h'_x h'_y + f'_z h''_{xy},$$

where the partial derivatives of h are determined by the partial derivatives of the constraint function g as specified. If (x_0, y_0, z_0) is the critical point found by the Lagrange multiplier method, then $a = F''_{xx}(x_0, y_0)$, $b = F''_{yy}(x_0, y_0)$, and $c = F''_{xy}(x_0, y_0)$ in the second-derivative test for the two-variable function F. □

CHAPTER 14

Multiple Integrals

97. Double Integrals

97.1. The Volume Problem. Suppose one needs to determine the volume of a hill whose height $f(\mathbf{r})$ as a function of position $\mathbf{r} = (x, y)$ is known. For example, the hill must be leveled to construct a highway. Its volume is required to estimate the number of truck loads needed to move the soil away. The following procedure can be used to estimate the volume. The base D of the hill is first partitioned into small pieces D_p of area ΔA_p, where $p = 1, 2, ..., N$ enumerates the pieces; that is, the union of all the pieces D_p is the region D. The partition elements should be small enough so that the height $f(\mathbf{r})$ has no significant variation when $\mathbf{r}$ is in D_p. The volume of the portion of the hill above each partition element D_p is approximately $\Delta V_p \approx f(\mathbf{r}_p)\, \Delta A_p$, where $\mathbf{r}_p$ is a point in D_p. The approximation becomes better for smaller D_p. The volume of the hill can therefore be estimated as

$$V \approx \sum_{p=1}^{N} f(\mathbf{r}_p)\, \Delta A_p.$$

For practical purposes, the values $f(\mathbf{r}_p)$ can be found, for example, from a detailed contour map of f.

The approximation is expected to become better and better as the size of the partition elements gets smaller (naturally, their number N has to increase). If R_p is the smallest radius of a disk that contains D_p, then put $R_N = \max_p R_p$, which determines the size of the largest partition element. When a larger number N of partition elements is taken to improve the accuracy of the approximation, one has to reduce R_N at the same time to make variations of f within each partition element smaller. Note that the reduction of the maximal area $\max_p \Delta A_p$ versus the maximal size R_N may not be good enough to improve the accuracy of the estimate. If D_p looks like a narrow strip, its area is small, but the variation of the height f along the strip may be significant and the accuracy of the approximation $\Delta V_p \approx f(\mathbf{r}_p)\, \Delta A_p$ is poor. One can therefore expect that the exact value of the volume is obtained in the limit

$$V = \lim_{\substack{N\to\infty \\ (R_N\to 0)}} \sum_{p=1}^{N} f(\mathbf{r}_p)\,\Delta A_p\,. \tag{14.1}$$

The volume V may be viewed as the volume of a solid bounded from above by the surface $z = f(x, y)$, which is the graph of f, and by the portion D of the xy plane. Naturally, it is not expected to depend on the way the region D is partitioned, neither should it depend on the choice of sample points $\mathbf{r}_p$ in each partition element.

The limit (14.1) resembles the limit of a Riemann sum for a single-variable function $f(x)$ on an interval $[a, b]$ used to determine the area under the graph of f. Indeed, if x_k, $k = 0, 1, ..., N$, $x_0 = a < x_1 < \cdots < x_{N-1} < x_N = b$ is the partition of $[a, b]$, then ΔA_p is the analog of $\Delta x_k = x_k - x_{k-1}$, $k = 1, 2, ..., N$, the number R_N is the analog of $\Delta_N = \max_k \Delta x_k$, and the values $f(\mathbf{r}_p)$ are analogous to $f(x_k^*)$, where $x_k^* \in [x_{k-1}, x_k]$. The area under graph is then

$$A = \lim_{\substack{N\to\infty \\ (\Delta_N\to 0)}} \sum_{k=1}^{N} f(x_k^*)\,\Delta x_k = \int_a^b f(x)\,dx\,.$$

So, the limit (14.1) seems to define an integral over a two-dimensional region D (i.e., with respect to both variables x and y used to label points in D). This observation leads to the concept of a *double integral*. However, the qualitative construction used to analyze the volume problem still lacks the level of rigor used to define the single-variable integration. For example, how does one choose the "shape" of the partition elements D_p, or how does one calculate their areas? These kinds of questions were not even present in the single-variable case and have to be addressed.

97.2. The Double Integral. Let D be a closed, bounded region. The boundaries of D are assumed to be piecewise-smooth curves. Let $f(\mathbf{r})$ be a *bounded* function on D, that is, $m \le f(\mathbf{r}) \le M$ for some numbers M and m and all $\mathbf{r} \in D$. The numbers m and M are called *lower and upper bounds* of f on D. Evidently, upper and lower bounds are not unique because any number smaller than m is also a lower bound, and, similarly, any number greater than M is an upper bound. However, the smallest upper bound and the largest lower bound are unique.

DEFINITION 14.1. (Supremum and Infimum).
Let f be bounded on D. The smallest upper bound of f on D is called

the supremum *of f on D and denoted by* $\sup_D f$*. The largest lower bound of f on D is called the* infimum *of f on D and denoted by* $\inf_D f$*.*

As a bounded region, D can always be embedded in a rectangle $R_D = \{(x,y)\,|\, x \in [a,b],\ y \in [c,d]\}$ (i.e., D is a subset of R_D). The function f is then *extended* to the rectangle R_D by setting its values to 0 for all points outside D, that is, $f(\mathbf{r}) = 0$ if $\mathbf{r} \in R_D$ and $\mathbf{r} \notin D$. Consider a partition x_k, $k = 0, 1, ..., N_1$, of the interval $[a,b]$, where $x_k = a + k\,\Delta x$, $\Delta x = (b-a)/N_1$, and a partition y_j, $j = 0, 1, ..., N_2$, of the interval $[c,d]$, where $y_j = c + j\,\Delta y$ and $\Delta y = (d-c)/N_2$. These partitions induce a partition of the rectangle R_D by rectangles $R_{kj} = \{(x,y)\,|\, x \in [x_{k-1}, x_k],\ y \in [y_{j-1}, y_j]\}$, where $k = 1, 2, ..., N_1$ and $j = 1, 2, ..., N_2$. The area of each partition rectangle R_{kj} is $\Delta A = \Delta x\,\Delta y$. This partition is called a *rectangular partition* of R_D. For every partition rectangle R_{kj}, there are numbers $M_{jk} = \sup f(\mathbf{r})$ and $m_{jk} = \inf f(\mathbf{r})$, the supremum and infimum of f on R_{kj}.

Definition 14.2. (Upper and Lower sums). *Let f be a bounded function on a closed bounded region D. Let R_D be a rectangle that contains D and let the function f be defined to have zero value for all points of R_D that do not belong to D. Given a rectangular partition R_{kj} of R_D, let $M_{jk} = \sup f$ and $m_{jk} = \inf f$ be the supremum and infimum of f on R_{jk}. The sums*

$$U(f, N_1, N_2) = \sum_{j=1}^{N_1}\sum_{k=1}^{N_2} M_{kj}\,\Delta A\,, \quad L(f, N_1, N_2) = \sum_{j=1}^{N_1}\sum_{k=1}^{N_2} m_{kj}\,\Delta A$$

are called the upper and lower sums*.*

The sequences of upper and lower sums have the following important property.

Corollary 14.1. (Property of Upper and Lower Sums). *The sequence of upper sums is decreasing, while the sequence of lower sums is increasing, that is, $U(f, N_1, N_2) \geq U(f, N_1', N_2')$ and $L(f, N_1, N_2) \leq L(f, N_1', N_2')$ if $N_1' \geq N_1$ and $N_2' \geq N_2$.*

Proof. Consider any partition rectangle R (the indices jk are omitted) of area ΔA. Put $M = \sup_R f$ and $m = \inf_R f$. When the partition is refined, R becomes a union of several rectangles R_p of area ΔA_p, $p = 1, 2, ..., q$, so that $\Delta A = \sum_p \Delta A_p$. Put $M_p = \sup_{R_p} f$ and $m_p = \inf_{R_p} f$. Since R_p is contained in R, one has $M_p \leq M$ and $m_p \geq m$. Therefore, if the partition is refined, the term $M\,\Delta A$ in the upper sum is replaced by $\sum_p M_p\,\Delta A_p \le M\sum_p \Delta A_p = M\,\Delta A$, and the term $m\,\Delta A$ in the lower sum is replaced by $\sum_p m_p\,\Delta A_p \ge m\sum_p \Delta A_p = m\,\Delta A$; that is, the upper sum either decreases or does not change, while the lower sum either increases or does not change. □

Continuing the analogy with the volume problem, the upper and lower sums represent upper and lower estimates of the volume. They should become closer and closer to the volume as the partition becomes finer and finer. This leads to the following natural definition of the double integral.

DEFINITION 14.3. (Double Integral).
If the limits of the upper and lower sums exist as $N_{1,2} \to \infty$ (or $(\Delta x, \Delta y) \to (0,0)$) and coincide, then f is said to be Riemann integrable on D, and the limit of the upper and lower sums

$$\iint_D f(x,y)\,dA = \lim_{N_{1,2}\to\infty} U(f, N_1, N_2) = \lim_{N_{1,2}\to\infty} L(f, N_1, N_2)$$

is called the double integral *of f over the region D.*

It should be emphasized that the double integral is defined as the two-variable limit $(\Delta x, \Delta y) \to (0,0)$. The upper and lower sums are functions of Δx and Δy because $N_1 = (b-a)/\Delta x$ and $N_2 = (d-c)/\Delta y$. The existence of the limit and its value must be established accordingly.

Let us discuss this definition from the point of view of the volume problem. First, note that a specific partition of D by rectangles has been used. In this way, the area ΔA_p of the partition element has been given a precise meaning as the area of a rectangle. Later, it will be shown that if the double integral exists in the sense of the above definition, then it exists if the rectangular partition is replaced by any partition of D by elements D_p of an arbitrary shape subject to certain conditions that allow for a precise evaluation of their area. Second, the volume (14.1) is indeed given by the double integral of f, and its value is *independent of the choice of sample points* $\mathbf{r}_p$. This is an extremely useful property that allows one to approximate the double integral with any desired accuracy by evaluating a suitable *Riemann sum*.

DEFINITION 14.4. (Riemann Sum).
Let f be a function on D that is contained in a rectangle R_D. Let f be defined by zero values outside of D in R_D. Let $\mathbf{r}^*_{jk}$ be a point in a partition rectangle R_{jk}, where R_{jk} form a partition of R_D. The sum

$$R(f, N_1, N_2) = \sum_{j=1}^{N_1}\sum_{k=1}^{N_2} f(\mathbf{r}^*_{jk})\,\Delta A$$

is called a Riemann sum.

THEOREM 14.1. (Convergence of Riemann Sums).
If a function f is integrable on D, then its Riemann sums for any choice of sample points $\mathbf{r}^*_{jk}$ converge to the double integral:

$$\lim_{N_{1,2}\to\infty} R(f, N_1, N_2) = \iint_D f\, dA.$$

PROOF. For any partition rectangle R_{jk} and any sample point $\mathbf{r}^*_{jk}$ in it, $m_{jk} \le f(\mathbf{r}^*_{jk}) \le M_{jk}$. It follows from this inequality that $L(f, N_1, N_2) \le R(f, N_1, N_2) \le U(f, N_1, N_2)$. Since f is integrable, the limits of the upper and lower sums exist and coincide. The conclusion of the theorem follows from the squeeze principle for limits. □

Approximation of Double Integrals. If f is integrable, its double integral can be approximated by a suitable Riemann sum. A commonly used choice of sample points is to take $\mathbf{r}^*_{jk}$ to be the intersection of the diagonals of partition rectangles R_{jk}, that is, $\mathbf{r}^*_{jk} = (\bar{x}_j, \bar{y}_k)$, where $\bar{x}_j$ and $\bar{y}_k$ are the midpoints of the intervals $[x_{j-1}, x_j]$ and $[y_{k-1}, y_k]$, respectively. This rule is called the *midpoint rule*. The accuracy of the midpoint rule approximation can be assessed by finding the upper and lower sums; their difference gives the upper bound on the absolute error of the approximation. Alternatively, if the integral is to be evaluated up to some significant decimals, the partition in the Riemann sum has to be refined until its value does not change in the significant digits. The integrability of f guarantees the convergence of Riemann sums and the independence of the limit from the choice of sample points.

97.3. Continuity and Integrability.

An Example of a Nonintegrable Function. Not every bounded function is integrable. Suppose f is defined on the square $x \in [0, 1]$ and $y \in [0, 1]$ so that $f(x, y) = 1$ if both x and y are rational, $f(x, y) = 2$ if both x and y are irrational, and $f(x, y) = 0$ otherwise. This function is not integrable. Recall that any interval $[a, b]$ contains both rational and irrational numbers. Therefore, any partition rectangle R_{jk} contains points whose coordinates are both rational, or both irrational, pairs of rational and irrational numbers. Hence, $M_{jk} = 2$ and $m_{jk} = 0$. The lower sum vanishes for any partition and therefore its limit is 0, whereas the upper sum is $2\sum_{jk} \Delta A = 2A = 2$ for any partition, where A is the area of the square. The limits of the upper and lower sums do not coincide, $2 \ne 0$, and the double integral of f does not exist. The Riemann sum for this function can converge to any number between 2 and 0, depending on the choice of sample points. For example, if the sample points have rational coordinates, then the Riemann sum equals 1. If the sample points have irrational coordinates, then the Riemann sum equals 2. If the sample points are such that one coordinate is rational while the other is irrational, then the Riemann sum vanishes.

The following theorem describes a class of integrable functions that is sufficient in many practical applications.

Theorem 14.2. (Integrability of Continuous Functions). *Let D be a closed, bounded region whose boundaries are piecewise-smooth curves. If a function f is continuous on D, then it is integrable on D.*

Note that the converse is not true; that is, the class of integrable functions is wider than the class of all continuous functions. This is a rather natural conclusion in view of the analogy between the double integral and the volume. The volume of a solid below a graph $z = f(x, y) \geq 0$ of a continuous function on D should exist. On the other hand, let $f(x, y)$ be defined on $D = \{(x, y)|x \in [0, 2],\ y \in [0, 1]\}$ so that $f(x, y) = m$ if $x \leq 1$ and $f(x, y) = M$ if $x > 1$. The function is piecewise constant and has a jump discontinuity along the line $x = 1$ in D. The volume below the graph $z = f(x, y)$ and above D is easy to find; it is the sum of volumes of two rectangular boxes with the same base area $A_1 = A_2 = 1$ and different heights M and m, $V = MA_1 + mA_2 = M + m$. The double integral of f exists and also equals $M + m$. Indeed, for any rectangular partition, the numbers M_{jk} and m_{jk} differs only for partition rectangles intersected by the discontinuity line $x = 1$, that is, $M_{jk} - m_{jk} = M - m$ for all such rectangles. Therefore, the difference between the upper and lower sums is $l\,\Delta x(M - m)$, where $l = 1$ is the length of the discontinuity curve. In the limit $\Delta x \to 0$, the difference vanishes. As noted earlier, the upper and lower sums are the upper and lower estimates of the volume and should therefore converge to it as their limits coincide. Using a similar line of arguments, one can prove the following.

Corollary 14.2. *If f is bounded on D and discontinuous only on a finite number of smooth curves, then it is integrable on D.*

98. Properties of the Double Integral

The properties of the double integral are similar to those of an ordinary integral and can be established directly from the definition.

Linearity. Let f and g be functions integrable on D and let c be a number. Then

$$\iint_D (f+g)\,dA = \iint_D f\,dA + \iint_D g\,dA\,,$$

$$\iint_D cf\,dA = c\iint_D f\,dA\,.$$

Area. The function χ is called the *characteristic function* of the region D if $\chi(\mathbf{r}) = 1$ if $\mathbf{r} \in D$ and $\chi(\mathbf{r}) = 0$ otherwise. Since χ is constant on D, it is also continuous on D and hence integrable. If follows that

$$\iint_D \chi\,dA = \iint_D dA = A(D), \tag{14.2}$$

where $A(D)$ is called the *area* of D. The region D can always be covered by the union of adjacent rectangles of area $\Delta A = \Delta x\,\Delta y$. In the limit $(\Delta x, \Delta y) \to (0,0)$, the total area of these rectangles converges to the area of D.

Additivity. Suppose that D is the union of D_1 and D_2 such that the area of their intersection is 0; that is, D_1 and D_2 may only have common points at their boundaries or no common points at all. If f is integrable on D, then

$$\iint_D f\,dA = \iint_{D_1} f\,dA + \iint_{D_2} f\,dA\,.$$

This property is the most difficult to prove directly from the definition. However, it appears rather natural when making the analogy of the double integral and the volume. If the region D is cut into two pieces D_1 and D_2, then the solid above D is also cut into two solids, one above D_1 and the other above D_2. Naturally, the volume is additive.

Suppose that f is nonnegative on D_1 and nonpositive on D_2. The double integral over D_1 is the volume of the solid above D_1 and below the graph of f. Since $-f \geq 0$ on D_2, the double integral over D_2 is the *negative* volume of the solid *below* D_2 and *above* the graph of f. When f becomes negative, its graph goes below the plane $z = 0$ (the xy plane). So, in general, the double integral may vanish or take negative values, depending on which volume (above or below the xy plane is larger). This property is analogous to the familiar relation between the ordinary integral and the area under the graph.

Positivity. If $f(\mathbf{r}) \geq 0$ for all $\mathbf{r} \in D$, then

$$\iint_D f\,dA \geq 0,$$

and, as a consequence of the linearity,

$$\iint_D f\,dA \geq \iint_D g\,dA$$

if $f(\mathbf{r}) \geq g(\mathbf{r})$ for all $\mathbf{r} \in D$.

Upper and Lower Bounds. Let $m = \inf_D f$ and $M = \sup_D f$. Then $m \leq f(\mathbf{r}) \leq M$ for all $\mathbf{r} \in D$. From the positivity of the double integral, it follows that

$$mA(D) \leq \iint_D f\,dA \leq MA(D)\,.$$

This inequality is easy to visualize. If f is positive, then the double integral is the volume of the solid below the graph of f. The solid lies in the cylinder with the cross section D. The graph of f lies between the planes $z = m$ and $z = M$. Therefore, the volume of the cylinder of height m cannot exceed the volume of the solid, whereas the latter cannot exceed the volume of the cylinder of height M.

Theorem 14.3. (Integral Mean Value Theorem).
If f is continuous on D, then there exists a point $\mathbf{r}_0 \in D$ such that

$$\iint_D f\,dA = f(\mathbf{r}_0)A(D)\,.$$

Proof. Let h be a number. Put $g(h) = \iint_D (f-h)\,dA = \iint_D f\,dA - hA(D)$. From the upper and lower bounds for the double integral, it follows that $g(M) \leq 0$ and $g(m) \geq 0$. Since $g(h)$ is linear in h, there exists $h = h_0 \in [m, M]$ such that $g(h_0) = 0$. On the other hand, a continuous function on a closed, bounded region D takes its maximal and minimal values as well as all the values between them. Therefore, for any $m \leq h_0 \leq M$, there is $\mathbf{r}_0 \in D$ such that $f(\mathbf{r}_0) = h_0$. □

A geometrical interpretation of the integral mean value theorem is rather simple. Imagine that the solid below the graph of f is made of clay. The shape of a piece of clay may be deformed while the volume is preserved under deformation. The nonflat top of the solid can be deformed so that it becomes flat, turning the solid into a cylinder of height h_0, which, by volume preservation, should be between the smallest and the largest heights of the original solid. The integral mean value theorem merely states the existence of such an *average* height at which the volume of the cylinder coincides with the volume of the solid with a nonflat top. The continuity of the function is sufficient (but not necessary) to establish that there is a point at which the average height coincides with the value of the function.

Integrability of the Absolute Value. Suppose that f is integrable on a bounded, closed region D. Then its absolute value $|f|$ is also integrable and

$$\left|\iint_D f\,dA\right| \le \iint_D |f|\,dA.$$

A proof of the integrability of $|f|$ is rather technical. Once the integrability of $|f|$ is established, the inequality is a simple consequence of $|a+b| \le |a| + |b|$ applied to a Riemann sum of f. Making the analogy between the double integral and the volume, suppose that $f \ge 0$ on D_1 and $f \le 0$ on D_2, where $D_{1,2}$ are two portions of D. If V_1 and V_2 stand for the volumes of the solids bounded by the graph of f and D_1 and D_2, respectively, then the double integral of f over D is $V_1 - V_2$, while the double integral of $|f|$ is $V_1 + V_2$. Naturally, $|V_1 - V_2| \le V_1 + V_2$ for positive $V_{1,2}$.

Independence of Partition. It has been argued that the volume of a solid under the graph of f and above a region D can be computed by (14.1) in which the Riemann sum is defined for an *arbitrary* (nonrectangular) partition of D. Can the double integral of f over D be computed in the same way? The answer is affirmative. However, the proof goes beyond the scope of this course. If f is continuous, then the assertion is not so difficult to establish. Let D be partitioned by piecewise smooth curves into partition elements D_p, $p = 1, 2, ..., N$, so that the union of D_p is D and $A(D) = \sum_{p=1}^{N} \Delta A_p$, where ΔA_p is the area of D_p defined by (14.2). If R_p is the smallest radius of a disk that contains D_p, put $R_N = \max R_p$; that is, R_p characterizes the size of the partition element D_p and R_N is the size of the largest partition element. Recall that the largest partition element does not necessarily have the largest area. The partition is said to be refined if $R_N < R_{N'}$ for $N < N'$; that is, the size of the largest partition element decreases. Then, if f is continuous on D, there are points $\mathbf{r}_p \in D_p$ such that

$$\iint_D f\,dA = \sum_{p=1}^{N} \iint_{D_p} f\,dA = \sum_{p=1}^{N} f(\mathbf{r}_p)\,\Delta A_p\,.$$

The first equality follows from the additivity of the double integral, and the second one holds by the integral mean value theorem. Consider the Riemann sum

$$R(f,N) = \sum_{p=1}^{N} f(\mathbf{r}_p^*)\,\Delta A_p,$$

where $\mathbf{r}_p^* \in D_p$ are sample points. If $\mathbf{r}_p^* \ne \mathbf{r}_p$, then the Riemann sum does not coincide with the double integral. However, its limit as $N \to$

∞ equals the double integral. Indeed, put $c_p = |f(\mathbf{r}_p^*) - f(\mathbf{r}_p)|$ and $c_N = \max c_p$, $p = 1, 2, ..., N$. Since f is continuous, $c_N \to 0$ as $N \to \infty$ because any partition element D_p is contained in a disk of radius $R_p \leq R_N \to 0$ as $N \to \infty$. Therefore, the deviation of the Riemann sum from the double integral converges to 0:

$$\begin{aligned}\Big|\iint f\,dA - R(f,N)\Big| &= \Big|\sum_{p=1}^{N}(f(\mathbf{r}_p) - f(\mathbf{r}_p^*))\,\Delta A_p\Big| \\ &\leq \sum_{p=1}^{N} |f(\mathbf{r}_p) - f(\mathbf{r}_p^*)|\Delta A_p \\ &= \sum_{p=1}^{N} c_p\,\Delta A_p \leq c_N \sum_{p=1}^{N} \Delta A_p = c_N A(D) \to 0\end{aligned}$$

as $N \to \infty$. So the double integral can be approximated by Riemann sums for arbitrary partitions subject to the conditions specified above, that is,

$$\iint_D f\,dA = \lim_{\substack{N\to\infty \\ (R_N \to 0)}} \sum_{p=1}^{N} f(\mathbf{r}_p^*)\,\Delta A_p \tag{14.3}$$

for any choice of sample points $\mathbf{r}_p^*$. Note that the region D is no longer required to be embedded in a rectangle and f does not have to be extended outside of D. This property is useful for evaluating double integrals by means of *change of variables* discussed later in this chapter. It is also useful to simplify calculations of Riemann sums.

Example 14.1. *Find the double integral of $f(x, y) = x^2 + y^2$ over the disk D $x^2 + y^2 \leq 1$ using the partition of D by concentric circles and rays from the origin.*

Solution: Consider circles $x^2 + y^2 = r_p^2$, where $r_p = p\,\Delta r$, $\Delta r = 1/N$, and $p = 0, 1, 2, ..., N$. If θ is the polar angle in the plane, then points with a fixed value of θ form a ray from the origin. Let the disk D be partitioned by circles of radii r_p and rays $\theta = \theta_k = k\,\Delta\theta$, $\Delta\theta = 2\pi/n$, $k = 1, 2, ..., n$. Each partition element lies in the sector of angle $\Delta\theta$ and is bounded by two circles whose radii differ by Δr. The area of a sector of radius r_p is $r_p^2\,\Delta\theta/2$. Therefore, the area of a partition element between circles of radii r_p and r_{p+1} is $\Delta A_p = r_{p+1}^2\,\Delta\theta/2 - r_p^2\,\Delta\theta/2 = (r_{p+1}^2 - r_p^2)\,\Delta/2\theta = (r_{p+1} + r_p)\,\Delta r\,\Delta\theta/2$. In the Riemann sum, use the midpoint rule; that is, the sample points are intersections of the circles of radius $\bar{r}_p = (r_{p+1} + r_p)/2$ and the rays with angles $\bar{\theta}_k = (\theta_{k+1} + \theta_k)/2$.

The values of f at the sample points are $f(\mathbf{r}_p^*) = \bar{r}_p^2$, the area elements are $\Delta A_p = \bar{r}_p\,\Delta r\,\Delta\theta$, and the corresponding Riemann sum reads

$$R(f, N, n) = \sum_{k=1}^{n}\sum_{p=1}^{N} \bar{r}_p^3\ \Delta r\,\Delta\theta = 2\pi \sum_{p=1}^{N} \bar{r}_p^3\ \Delta r$$

because $\sum_{k=1}^{n} \Delta\theta = 2\pi$, the total range of θ in the disk D. The sum over p is the Riemann sum for the single-variable function $g(r) = r^3$ on the interval $r \in [0, 1]$. In the limit $N \to \infty$, this sum converges to the integral of g over the interval $[0, 1]$, that is,

$$\iint_D (x^2 + y^2)\,dA = 2\pi \lim_{N\to\infty} \sum_{p=1}^{N} \bar{r}_p^3\ \Delta r = 2\pi \int_0^1 r^3\,dr = \pi/2\,.$$

So, by choosing the partition according to the shape of D, the double Riemann sum has been reduced to a Riemann sum for a single-variable function. □

The numerical value of the double integral in this example is the volume of the solid that lies between the paraboloid $z = x^2 + y^2$ and the disk D of unit radius. It can also be represented as the volume of the cylinder with height $h = 1/2$, $V = hA(D) = \pi h = \pi/2$. This observation illustrates the integral mean value theorem. The function f takes the value $h = 1/2$ on the circle $x^2 + y^2 = 1/2$ of radius $1/\sqrt{2}$ in D.

99. Iterated Integrals

Here a practical method for evaluating double integrals will be developed. To simplify the technicalities, the derivation of the method is given for continuous functions. However, the method is also valid for bounded functions that are discontinuous on a finite number of smooth curves, which is sufficient for many practical applications.

99.1. Rectangular Domains. Let a function f be continuous on D. Suppose first that D is a rectangle $x \in [a, b]$ and $y \in [c, d]$. Let R_{jk} be a rectangular partition of D as defined earlier. For any choice of sample points $r_{jk}^* = (x_j^*, y_k^*)$, where $x_j^* \in [x_{j-1}, x_j]$ and $y_k^* \in [y_{k-1}, y_k]$, the Riemann sum converges to the double integral of f over D. Since the double limit of the Riemann sum (as $(\Delta x, \Delta y) \to (0, 0)$) exists, it should not depend on the order in which the limits $N_1 \to \infty$ (or $\Delta x \to 0$) and

$N_2 \to \infty$ (or $\Delta y \to 0$) are computed. This is the key observation for what follows. Suppose the limit $\Delta y \to 0$ is to be evaluated first:

$$\begin{aligned}\iint_D f\,dA &= \lim_{N_{1,2}\to\infty} R(f, N_1, N_2)\\ &= \lim_{N_1\to\infty} \sum_{j=1}^{N_1} \left(\lim_{N_2\to\infty} \sum_{k=1}^{N_2} f(x_j^*, y_k^*)\,\Delta y \right) \Delta x\,.\end{aligned}$$

The expression in parenthese is nothing but the Riemann sum for the single-variable function $g_j(y) = f(x_j^*, y)$ on the interval $y \in [c, d]$. So, if the functions $g_j(y)$ are integrable on $[c, d]$, then the limit of their Riemann sums is the integral of g_j over the interval. If f is continuous on D, then it must also be continuous along the lines $x = x_j^*$ in D; that is, $g_j(y) = f(x_j^*, y)$ is continuous and hence integrable on $[c, d]$. Thus,

$$\lim_{N_2\to\infty} \sum_{k=1}^{N_2} f(x_j^*, y_k^*)\,\Delta y = \int_c^d f(x_j^*, y)\,dy\,. \tag{14.4}$$

Define a function $A(x)$ by

$$A(x) = \int_c^d f(x, y)\,dy\,. \tag{14.5}$$

The value of A at x is given by the integral of f with respect to y; the integration with respect to y is carried out as if x were a fixed number. For example, put $f(x, y) = x^2y + e^{xy}$ and $[c, d] = [0, 1]$. Then an antiderivative $F(x, y)$ of $f(x, y)$ with respect to y is $F(x, y) = x^2y^2/2 + e^{xy}/x$, which means that $F_y'(x, y) = f(x, y)$. Therefore,

$$A(x) = \int_0^1 (x^2y + e^{xy})\,dy = x^2y^2/2 + e^{xy}/x\Big|_0^1 = x^2/2 + e^x/x - 1/x\,.$$

A geometrical interpretation of $A(x)$ is simple. If $f \geq 0$, then $A(x_j^*)$ is the area of the cross section of the solid below the graph $z = f(x, y)$ by the plane $x = x_j^*$, and $A(x_j^*)\,\Delta x$ is the volume of the slice of the solid of width Δx.

The second sum in the Riemann sum for the double integral in the Riemann sum of $A(x)$ on the interval $[a, b]$:

$$\begin{aligned}\iint_D f\,dA &= \lim_{N_1\to\infty} \sum_{j=1}^{N_1} A(x_j^*)\,\Delta x = \int_a^b A(x)\,dx\\ &= \int_a^b \left(\int_c^d f(x, y)\,dy \right) dx,\end{aligned}$$

where the integral exists by the continuity of A. The integral on the right side of this equality is called the *iterated integral*. In what follows, the parenthese in the iterated integral will be omitted. The order in which the integrals are evaluated is specified by the order of the differentials in it; for example, $dy\,dx$ means that the integration with respect to y is to be carried out first. In a similar fashion, by computing the limit $\Delta x \to 0$ first, the double integral can be expressed as an iterated integral in which the integration is carried out with respect to x and then with respect to y. So the following result has been established.

THEOREM 14.4. (Fubini's Theorem).
If f is continuous on the rectangle $D = \{(x, y)\,|\,x \in [a, b],\ y \in [c, d]\}$, then

$$\iint_D f(x,y)\,dA = \int_c^d \int_a^b f(x,y)\,dx\,dy = \int_a^b \int_c^d f(x,y)\,dy\,dx.$$

More generally, this is true for any bounded f on D that is discontinuous on a finite number of smooth curves.

Think of a loaf of bread with a rectangular base and with a top having the shape of the graph $z = f(x, y)$. It can be sliced along either of the two directions parallel to adjacent sides of its base. Fubini's theorem says that the volume of the loaf is the sum of the volumes of the slices and is independent of how the slicing is done.

EXAMPLE 14.2. *Find the volume of the solid bounded from above by the portion of the paraboloid $z = 4 - x^2 - 2y^2$ and from below by the portion of the paraboloid $z = -4 + x^2 + 2y^2$, where $x \in [0, 1]$ and $y \in [0, 1]$.*

SOLUTION: If the height of the solid at any $(x, y) \in D$ is $h(x, y) = z_{\text{top}}(x, y) - z_{\text{bot}}(x, y)$, where the graphs $z = z_{\text{top}}(x, y)$ and $z = z_{\text{bot}}(x, y)$ are the top and bottom boundaries of the solid, then the volume is

$$\begin{aligned} V &= \iint_D h(x,y)\,dA = \iint_D [z_{\text{top}}(x,y) - z_{\text{bot}}(x,y)]\,dA \\ &= \iint_D (8 - 2x^2 - 4y^2)\,dA = \int_0^1 \int_0^1 (8 - 2x^2 - 4y^2)\,dy\,dx \\ &= \int_0^1 [(8 - 2x^2)y - 4y^3/3]\Big|_0^1 dx = \int_0^1 (8 - 2x^2 - 4/3)\,dx = 6. \end{aligned}$$

□

COROLLARY 14.3. (Factorization of Iterated Integrals).
Let D be a rectangle $\{(x,y)\,|\,x\in[a,b],\ y\in[c,d]\}$. Suppose $f(x,y)=g(x)h(y)$, where the functions g and h are integrable on $[a,b]$ and $[c,d]$, respectively. Then

$$\iint_D f(x,y)\,dA=\int_a^b g(x)\,dx\int_c^d h(y)\,dy\,.$$

This simple consequence of Fubini's theorem is quite useful.

EXAMPLE 14.3. *Evaluate the double integral of $f(x,y)=\sin(x+y)$ over the rectangle $x\in[0,\pi]$ and $y\in[-\pi/2,\pi/2]$.*

SOLUTION: One has $\sin(x+y)=\sin x\cos y+\cos x\sin y$. The integral of $\sin y$ over $[-\pi/2,\pi/2]$ vanishes by symmetry. So, by the factorization property of the iterated integral, only the first term contributes to the double integral:

$$\iint_D \sin(x+y)\,dA=\int_0^\pi \sin x\,dx\int_{-\pi/2}^{\pi/2}\cos y\,dy=4\,.$$

□

The following example illustrates the use of the additivity of a double integral.

EXAMPLE 14.4. *Evaluate the double integral of $f(x,y)=15x^4y^2$ over the region D, which is the rectangle $[-2,2]\times[-2,2]$ with the rectangular hole $[-1,1]\times[-1,1]$.*

SOLUTION: Let $D_1=[-2,2]\times[-2,2]$ and let $D_2=[-1,1]\times[-1,1]$. The rectangle D_1 is the union of D and D_2 such that their intersection has no area. Hence,

$$\begin{aligned}\iint_{D_2} f\,dA=\iint_D f\,dA+\iint_{D_1} f\,dA\quad&\Rightarrow\quad\iint_D f\,dA\\&=\iint_{D_2} f\,dA-\iint_{D_1} f\,dA.\end{aligned}$$

By evaluating the double integrals over $D_{1,2}$,

$$\begin{aligned}\iint_{D_1} 15x^4y^2\,dA&=15\int_{-2}^2 x^4\,dx\int_{-2}^2 y^2\,dy=2^{10},\\ \iint_{D_2} 15x^4y^2\,dA&=15\int_{-1}^1 x^4\,dx\int_{-1}^1 y^2\,dy=4.\end{aligned}$$

the double integral over D is obtained, $1024-4=1020$. □

99.2. Study Problems.

Problem 14.1. *Suppose a function f has continuous second derivatives on the rectangle $R = [0,1] \times [0,1]$. Find $\iint_R f''_{xy}\,dA$ if $f(0,0) = 1$, $f(0,1) = 2$, $f(1,0) = 3$, and $f(1,1) = 5$.*

Solution: By Fubini's theorem,

$$\begin{aligned}
\iint_R f''_{xy}\,dA =& \int_0^1 \int_0^1 \frac{\partial}{\partial x} f'_y(x,y)\,dx\,dy = \int_0^1 f'_y(x,y)\Big|_0^1 dy \\
=& \int_0^1 [f'_y(1,y) - f'_y(0,y)]\,dy = \int_0^1 \frac{d}{dy}[f(1,y) - f(0,y)]\,dy \\
=& [f(1,y) - f(0,y)]\Big|_0^1 \\
=& \; [f(1,1) - f(0,1)] - [f(1,0) - f(0,0)] = 1.
\end{aligned}$$

By Clairaut's theorem $f''_{xy} = f''_{yx}$ and the value of the integral is independent of the order of integration. □

100. Double Integrals Over General Regions

The concept of the iterated integral can be extended to general regions subject to the following conditions.

100.1. Simple Regions.

Definition 14.5. (Simple and Convex Regions).
A region D is said to be simple *in the direction* $\mathbf{u}$ *if any line parallel to the vector* $\mathbf{u}$ *intersects D along at most one straight line segment. A region D is called* convex *if it is simple in any direction.*

Suppose D is simple in the direction of the y axis. It will be referred to as *y simple* or *vertically simple*. Since D is bounded, there is an interval $[a,b]$ such that vertical lines $x = x_0$ intersect D if $x_0 \in [a,b]$. In other words, the region D lies within the vertical strip $a \le x \le b$. Take a vertical line $x = x_0 \in [a,b]$ and consider all points of D that also belong to the line, that is, pairs $(x_0, y) \in D$, where the first coordinate is fixed. Since the line intersects D along a segment, the variable y ranges over an interval. The endpoints of this interval depend on the line or the value of x_0; that is, for every $x_0 \in [a,b]$, $y_{\rm bot} \le y \le y_{\rm top}$, where the numbers $y_{\rm bot}$ and $y_{\rm top}$ depend on x_0. In other words, vertically simple regions admit the following algebraic description.

Algebraic Description of Vertically Simple Regions. If D is vertically simple, then it lies in the vertical strip $a \le x \le b$ and is bounded from below by the graph $y = y_{\rm bot}(x)$ and from above by the graph $y = y_{\rm top}(x)$:

$$D = \{(x, y)\,|\, y_{\rm bot}(x) \le y \le y_{\rm top}(x)\,, \quad x \in [a, b]\}. \tag{14.6}$$

The numbers a and b are, respectively, the smallest and the largest values of the x coordinate of points of D. For example, the half-disk $x^2 + y^2 \le 1$, $y \ge 0$, is a vertically simple region. The x coordinate of any point in the disk lies in the interval $[a, b] = [-1, 1]$. For every x in this interval, the y coordinate lies in the interval $0 \le y \le \sqrt{1 - x^2}$; that is, in the vertical direction, the top boundary of the disk is the graph $y = \sqrt{1 - x^2} = y_{\rm top}(x)$ and the bottom boundary is the graph $y = 0 = y_{\rm bot}(x)$.

Suppose D is simple in the direction of the x axis. It will be referred to as *x simple* or *horizontally simple*. Since D is bounded, there is an interval $[c, d]$ such that horizontal lines $y = y_0$ intersect D if $y_0 \in [c, d]$. In other words, the region D lies within the horizontal strip $c \le y \le d$. Take a horizontal line $y = y_0 \in [c, d]$ and consider all points of D that also belong to the line, that is, pairs $(x, y_0) \in D$, where the second coordinate is fixed. Since the line intersects D along a segment, the variable x ranges over an interval. The endpoints of this interval depend on the line or the value of y_0; that is, for every $y_0 \in [c, d]$, $x_{\rm bot} \le x \le x_{\rm top}$, where the numbers $x_{\rm bot}$ and $x_{\rm top}$ depend on y_0. In other words, horizontally simple regions admit the following algebraic description.

Algebraic Description of Horizontally Simple Regions. If D is horizontally simple, then it lies in a horizontal strip $c \le y \le b$ and is bounded from below by the graph $x = x_{\rm bot}(y)$ and from above by the graph $x = x_{\rm top}(y)$:

$$D = \{(x, y)\,|\, x_{\rm bot}(y) \le x \le x_{\rm top}(y)\,, \quad y \in [c, d]\}. \tag{14.7}$$

The numbers c and d are, respectively, the smallest and the largest values of the y coordinate of points of D. The terms “below” and “above” are now defined relative to the line of sight in the direction of the x axis. For example, the half-disk $x^2 + y^2 \le 1$, $y \ge 0$, is also a horizontally simple region. The y coordinate of any point in the disk lies in the interval $[c, d] = [0, 1]$. For every y in this interval, the x coordinate lies in the interval $-\sqrt{1 - y^2} \le x \le \sqrt{1 - y^2}$; that is, in the horizontal direction, the top boundary of the disk is the graph $x = \sqrt{1 - y^2} = x_{\rm top}(y)$ and the bottom boundary is the graph $x = -\sqrt{1 - y^2} = x_{\rm bot}(y)$.

100.2. Iterated Integrals for Simple Regions. Suppose D is vertically simple. Then it should have an algebraic description according to (14.6). For the embedding rectangle R_D, one can take $[a,b]\times[c,d]$, where $c \le y_{\rm bot}(x) \le y_{\rm top}(x) \le d$ for all $x \in [a,b]$. The function f is continuous in D and defined by zero values outside D; that is, $f(x,y)=0$ if $c \le y < y_{\rm bot}(x)$ and $y_{\rm top}(x) < y \le d$, where $x \in [a,b]$. Consider a Riemann sum for a rectangular partition of R_D with sample points (x_j^*, y_k^*) just like in the case of rectangular domains discussed earlier. Since f is integrable, the double integral exists, and the double limit of the Riemann sum should not depend on the order in which the limits $\Delta x \to 0$ and $\Delta y \to 0$ are taken. For a vertically simple D, the limit $\Delta y \to 0$ is taken first. Similarly to (14.4), one infers that

$$\lim_{N_2\to\infty}\sum_{k=1}^{N_2} f(x_j^*, y_k^*)\,\Delta y = \int_c^d f(x_j^*, y)\,dy = \int_{y_{\rm bot}(x_j^*)}^{y_{\rm top}(x_j^*)} f(x_j^*, y)\,dy$$

because the function f vanishes outside the interval $y_{\rm bot}(x) \le y \le y_{\rm top}(x)$ for any $x \in [a,b]$. The area of the slice of the solid below the graph $z = f(x,y)$ is also given by (14.5):

$$A(x) = \int_c^d f(x,y)\,dy = \int_{y_{\rm bot}(x)}^{y_{\rm top}(x)} f(x,y)\,dy\,.$$

Note that the last equality is only possible for a vertically simple base D of the solid. If D were not vertically simple, then such a slice would not have been a single slice but rather a few disjoint slices, depending on how many disjoint intervals are in the intersection of a vertical line with D. In this case, the integration with respect to y would have yielded a sum of integrals over all such intervals. The reason the integration with respect to y is to be carried out first only for vertically simple regions is exactly to avoid the necessity to integrate over a union of disjoint intervals. Finally, the value of the double integral is given by the integral of $A(x)$ over the interval $[a,b]$.

Iterated Integral for Vertically Simple regions. Let D be a vertically simple region; that is, it admits the algebraic description (14.6). The double integral of f over D is then given by the iterated integral

$$\text{(14.8)} \qquad \iint_D f(x,y)\,dA = \int_a^b \int_{y_{\rm bot}(x)}^{y_{\rm top}(x)} f(x,y)\,dy\,dx.$$

Iterated Integral for Horizontally Simple Regions. Naturally, for horizontally simple regions, the integration with respect to x should be carried out first; that is, in the Riemann sum, the limit $\Delta x \to 0$ is taken first. Let D be a horizontally simple region; that is, it admits the algebraic description (14.7). The double integral of f over D is then given by the iterated integral

$$\iint_D f(x,y)\,dA = \int_c^d \int_{x_{\rm bot}(y)}^{x_{\rm top}(y)} f(x,y)\,dx\,dy. \tag{14.9}$$

Iterated Integrals for Nonsimple Regions. If the integration region D is not simple, how can one evaluate the double integral? Any nonsimple region can be cut into simple regions D_p, $p = 1, 2, ..., n$. The double integral over simple regions can then be evaluated. The double integral over D is then the sum of the double integrals over D_p by the additivity property (see Example 14.4).

EXAMPLE 14.5. *Evaluate the double integral of $f(x,y) = 6yx^2$ over the region D bounded by the line $y = 1$ and the parabola $y = x^2$.*

SOLUTION: The region D is both horizontally and vertically simple. It is therefore possible to use either (14.8) or (14.9). To find an algebraic description of D as a vertically simple region, one has to first specify the maximal range of the x coordinate in D. It is determined by the intersection of the line $y = 1$ and the parabola $y = x^2$, that is, $1 = x^2$, and hence $x \in [a,b] = [-1,1]$ for all points of D. For any $x \in [-1,1]$, the y coordinate of points of D attains the smallest value on the parabola (i.e., $y_{\rm bot}(x) = x^2$), and the largest value on the line (i.e., $y_{\rm top}(x) = 1$). One has

$$\iint_D 6yx^2\,dA = 6\int_{-1}^1 x^2 \int_{x^2}^1 y\,dy\,dx = 3\int_{-1}^1 x^2(1-x^4)\,dx = 8/7\,.$$

It is also instructive to obtain this result using the reverse order of integration. To find an algebraic description of D as a horizontally simple region, one has to first specify the maximal range of the y coordinate in D. The smallest value of y is 0 and the largest value is 1; that is, $y \in [c,d] = [0,1]$ for all points of D. For any fixed $y \in [0,1]$, the x coordinate of points of D attains the smallest and largest values when $y = x^2$ or $x = \pm\sqrt{y}$, that is, $x_{\rm bot}(y) = -\sqrt{y}$ and $x_{\rm top}(y) = \sqrt{y}$. One has

$$\begin{aligned}\iint_D 6yx^2\,dA &= 6\int_0^1 y\int_{-\sqrt{y}}^{\sqrt{y}} x^2\,dx\,dy = 2\int_0^1 y(2y^{3/2})\,dy\\ &= 4\int_0^1 y^{5/2}\,dy = 8/7\,.\end{aligned}$$

□

100.3. Reversing the Order of Integration. By reversing the order of integration, a simplification of technicalities involved in evaluating double integrals can be achieved, but not always, though.

EXAMPLE 14.6. *Evaluate the double integral of $f(x, y) = 2x$ over the region D bounded by the line $x = 2y+2$ and the parabola $x = y^2-1$.*

SOLUTION: The region D is both vertically and horizontally simple. However, the iterated integral based on the algebraic description of D as a vertically simple region is more involved. Indeed, the largest value of the x coordinate in D occurs at one of the points of intersection of the line and the parabola, $2y + 2 = y^2 - 1$ or $(y - 1)^2 = 4$, and hence, $y = -1$, 3. The largest value of x in D is $x = 3^2 - 1 = 8$. The smallest value of x occurs at the point of intersection of the parabola with the x axis, $x = -1$. So $[a, b] = [-1, 8]$. However, the algebraic expression for the top boundary $y_{\text{top}}(x)$ is not the same for all $x \in [-1, 8]$. For any fixed $x \in [-1, 0]$, the range of the y coordinate is determined by the parabola, $-\sqrt{x+1} \leq y \leq \sqrt{x+1}$, while for any fixed $x \in [0, 8]$, the top and bottom boundaries of the range of y are determined by the line and parabola, respectively, $-\sqrt{x+1} \leq y \leq (x - 2)/2$. This dictates the necessity to split the region D into two regions D_1 and D_2 such that $x \in [-1, 0]$ for all points in D_1 and $x \in [0, 8]$ for all points in D_2. The corresponding iterated integral reads

$$\begin{aligned}\iint\limits_{D} 2x\,dA &= \iint\limits_{D_1} 2x\,dA + \iint\limits_{D_2} 2x\,dA \\ &= 2\int\limits_{-1}^{0} x \int\limits_{-\sqrt{x+1}}^{\sqrt{x+1}} dy\,dx + 2\int\limits_{0}^{8} x \int\limits_{-\sqrt{x+1}}^{x/2+1} dy\,dx.\end{aligned}$$

On the other hand, if the iterated integral corresponding to the algebraic description of D as a horizontally simple region is used, the technicalities are greatly simplified. The smallest and largest values of y in D occur at the points of intersection of the line and the parabola found above, $y = -1, 3$, that is, $[c, d] = [-1, 3]$. For any fixed $y \in [-1, 3]$, the x coordinate ranges from its value on the parabola to its value on the line, $x_{\text{bot}}(y) = y^2 - 1 \leq x \leq 2y + 2 = x_{\text{bot}}(y)$. The corresponding iterated integral reads

$$\iint_D 2x\,dA = 2\int_{-1}^{3}\int_{y^2-1}^{2y+2} x\,dx\,dy = \int_{-1}^{3}(-y^4+6y^2+8y+3)\,dy = 256/5,$$

which is simpler to evaluate than the previous one. □

Sometimes the iterated integration cannot even be carried out in one order, but it can still be done in the other order.

EXAMPLE 14.7. *Evaluate the double integral of* $f(x, y) = \sin(y^2)$ *over the region* D*, which is the rectangle bounded by* $x = 0$*,* $y = x$*, and* $y = \sqrt{\pi}$*.*

SOLUTION: Suppose the iterated integral for vertically simple regions is used. The range of the x coordinate is $x \in [0, \sqrt{\pi}] = [a, b]$, and, for every fixed $x \in [0, \sqrt{\pi}]$, the range of the y coordinate is $y_{\rm bot}(x) = x \le y \le \sqrt{\pi} = y_{\rm top}(x)$ in D. The iterated integral reads

$$\iint_D \sin(y^2)\, dA = \int_0^{\sqrt{\pi}} \int_x^{\sqrt{\pi}} \sin(y^2)\, dy\, dx\,.$$

However, the antiderivative of $\sin(y^2)$ cannot be expressed in elementary functions! Let us reverse the order of integration. The maximal range of the y coordinate in D is $[0, \sqrt{\pi}] = [c, d]$. For every fixed $y \in [0, \sqrt{\pi}]$, the range of the x coordinate is $x_{\rm bot}(y) = 0 \le x \le y = x_{\rm top}(y)$ in D. Therefore, the iterated integral reads

$$\begin{aligned}\iint\limits_D \sin(y^2)\, dA &= \int\limits_0^{\sqrt{\pi}} \sin(y^2) \int\limits_0^{y} dx\, dy \\ &= \int\limits_0^{\sqrt{\pi}} \sin(y^2) y\, dy = -\frac{1}{2}\cos(y^2)\Big|_0^{\sqrt{\pi}} = 1\,.\end{aligned}$$

□

100.4. The Use of Symmetry. The symmetry property has been established in single-variable integration:

$$f(-x) = -f(x) \quad \Rightarrow \quad \int_{-a}^{a} f(x)\, dx = 0,$$

which has proved to be quite useful. For example, the integral of $\sin(x^{2011})$ over any symmetric interval $[-a, a]$ vanishes because $\sin(x^{2011})$ is an antisymmetric function. A similar property can be established for double integrals. Consider a transformation that maps each point (x, y) of the plane to another point (x_s, y_s). A region D is said to be *symmetric* under a transformation $(x, y) \to (x_s, y_s)$ if the image D^s of D coincides with D (i.e., $D^s = D$). For example, let D be bounded by an ellipse $x^2/a^2 + y^2/b^2 = 1$. Then D is symmetric under reflections about the x axis, the y axis, or their combination, that is,

$(x,y) \to (x_s,y_s) = (-x,y)$, $(x,y) \to (x_s,y_s) = (x,-y)$, or $(x,y) \to (x_s,y_s) = (-x,-y)$. A transformation of the plane $(x,y) \to (x_s,y_s)$ is said to be *area preserving* if the image D^s of any region D under this transformation has the same area, that is, $A(D) = A(D^s)$. For example, translations, rotations, reflections about lines, and their combinations are area-preserving transformations.

Theorem 14.5. (Symmetry Property).
Let a region D be symmetric under an area-preserving transformation $(x,y) \to (x_s,y_s)$ such that $f(x_s,y_s) = -f(x,y)$. Then the integral of f over D vanishes:

$$\iint_D f(x,y)\,dA = 0\,.$$

A general proof is postponed until the change of variables in double integrals is discussed. Here the simplest case of a reflection about a line is considered. If D is symmetric under this reflection, then the line cuts D into two equal-area regions D_1 and D_2 so that $D_1^s = D_2$ and $D_2^s = D_1$. The double integral is independent of the choice of partition (see (14.3)). Consider a partition of D_1 by elements D_{1p}, $p = 1, 2, ..., N$. By symmetry, the images D_{1p}^s of the partition elements D_{1p} form a partition of D_2 such that $\Delta A_p = A(D_{1p}) = A(D_{1p}^s)$ by area preservation. Choose elements D_{1p} and D_{1p}^s to partition the region D. Now recall that the double integral is also independent of the choice of sample points. Suppose (x_p,y_p) are sample points in D_{1p}. Choose sample points in D_{1p}^s to be the images (x_{ps},x_{ps}) of (x_p,y_p) under the reflection. With these choices of the partition of D and sample points, the Riemann sum (14.3) vanishes:

$$\iint_D f\,dA = \lim_{N\to\infty}\sum_{p=1}^{N}\Big(f(x_p,y_p)\,\Delta A_p + f(x_{ps},y_{ps})\,\Delta A_p\Big) = 0\,,$$

where the two terms in the sum correspond to partitions of D_1 and D_2 in D; by the hypothesis, the function f is antisymmetric under the reflection and therefore $f(x_{ps},y_{ps}) = -f(x_p,y_p)$ for all p. From a geometrical point of view, the portion of the solid bounded by the graph $z = f(x,y)$ that lies above the xy plane has exactly the same shape as that below the xy plane, and therefore their volumes contribute with opposite signs to the double integral and cancel each other.

Example 14.8. *Evaluate the double integral of* $\sin[(x-y)^3]$ *over the portion D of the disk $x^2+y^2 \le 1$ that lies in the first quadrant ($x,y \ge 0$).*

Solution: The region D is symmetric under the reflection about the line $y = x$, that is, $(x, y) \to (x_s, y_s) = (y, x)$, whereas the function is anti-symmetric, $f(x_s, y_s) = f(y, x) = \sin[(y - x)^3] = \sin[-(x - y)^3] = -\sin[(x - y)^3] = -f(x, y)$. By the symmetry property, the double integral vanishes. □

Example 14.9. *Evaluate the double integral of* $f(x, y) = x^2y^3$ *over the region* D*, which is obtained from the elliptic region* $x^2/4 + y^2/9 \leq 1$ *by removing the square* $[0, 1] \times [0, 1]$.

Solution: Let D_1 and D_2 be the elliptic and square regions, respectively. The elliptic region D_1 is large enough to include the square D_2. Therefore, the additivity of the double integral can be used (compare Example 14.4) to transform the double integral over a non-simple region D into two double integrals over simple regions:

$$\begin{aligned}\iint_D x^2y^3\,dA &= \iint_{D_1} x^2y^3\,dA - \iint_{D_2} x^2y^3\,dA \\ &= -\iint_{D_2} x^2y^3\,dA = -\int_0^1 x^2\,dx \int_0^1 y^3\,dy = -1/12\,;\end{aligned}$$

the integral over D_1 vanishes because the elliptic region D_1 is symmetric under the reflection $(x, y) \to (x_s, y_s) = (x, -y)$, whereas the integrand is anti-symmetric, $f(x, -y) = x^2(-y)^3 = -x^2y^3 = -f(x, y)$. □

101. Double Integrals in Polar Coordinates

The polar coordinates are defined by the following relations:

$$x = r\cos\theta\,, \quad y = r\sin\theta\,, \quad \text{or} \quad r = \sqrt{x^2 + y^2}\,, \quad \theta = \tan^{-1}(y/x)\,,$$

where r is the distance from the origin to the point (x, y) and θ is the angle between the positive x axis and the ray from the origin through the point (x, y) counted counterclockwise. The value of $\tan^{-1}$ must be taken according to the geometrical definition of θ. If (x, y) lies in the first quadrant, then the value of $\tan^{-1}$ must be in the interval $[0, \pi/2)$ and $\tan^{-1}(\infty) = \pi/2$ and similarly for the other quadrants. These equations define a *one-to-one* correspondence between all points (x, y) of the plane and points of the strip $(r, \theta) \in (0, \infty) \times [0, 2\pi)$. Alternatively, one can also set the range of θ to be the interval $[-\pi, \pi)$. The ordered pair (r, θ) can be viewed as a point of an auxiliary plane or *polar plane*. In what follows, the r axis in this plane is set to be vertical, and the θ axis is set to be horizontal. For any region D, there is an image D' of D in the polar plane defined by the transformation of an ordered pair $(x, y) \in D$ to the ordered pair $(r, \theta) \in D'$. Note that the boundaries of D' are mapped onto the boundaries of D by

$x = r\cos\theta$ and $y = r\sin\theta$. For example, let D be the portion of the disk $x^2 + y^2 \le 1$ in the first quadrant. Then the shape of D' can be found from the images of boundaries of D in the polar plane:

$$\begin{aligned} \text{boundaries of } D &\leftrightarrow \text{boundaries of } D' \\ x^2 + y^2 = 1 &\leftrightarrow r = 1 \\ y = 0,\ x \ge 0 &\leftrightarrow \theta = 0 \\ x = 0,\ y \ge 0 &\leftrightarrow \theta = \pi/2 \end{aligned}$$

Since $r \ge 0$, the region D' is the rectangle $(r, \theta) \in [0, 1] \times [0, \pi/2] = D'$.

Let D' be the image of D on the polar plane and let R'_D be a rectangle containing D' so that the image of R'_D contains D. As before, a function f on D is extended outside D by setting its values to 0. Consider a rectangular partition of D' such that each partition rectangle D'_{jk} is bounded by the coordinate lines $r = r_j$, $r = r_{j+1} = r_j + \Delta r$, $\theta = \theta_k$, and $\theta = \theta_{k+1} = \theta_k + \Delta\theta$. Each partition rectangle has the area $\Delta A' = \Delta r\,\Delta\theta$. The image of the coordinate line $r = r_k$ in the xy plane is the circle of radius r_k centered at the origin. The image of the coordinate line $\theta = \theta_k$ on the xy plane is the ray from the origin that makes the angle θ_k with the positive x axis counted counterclockwise. The rays and circles are called *coordinate curves* of the polar coordinate system, that is, the curves along which either the coordinate r or the coordinate θ remains constant (circles and rays, respectively). A rectangular partition of D' induces a partition of D by the coordinate curves. Each partition element D_{jk} is the image of the rectangle D'_{jk} and is bounded by two circles and two rays.

Let $f(x, y)$ be an integrable function on D. The double integral of f over D can be computed as the limit of the Riemann sum. According to (14.3), the limit does not depend on either the choice of partition or the sample points. Let ΔA_{jk} be the area of D_{jk}. The area of the sector of the disk of radius r_j that has the angle $\Delta\theta$ is $r_j^2\Delta\theta/2$. Therefore,

$$\Delta A_{jk} = \frac{1}{2}(r_{j+1}^2 - r_j^2)\,\Delta\theta = \frac{1}{2}(r_{j+1} + r_j)\,\Delta r\,\Delta\theta = \frac{1}{2}(r_{j+1} + r_j)\,\Delta A'\,.$$

In (14.3), put $\Delta A_p = \Delta A_{jk}$, $\mathbf{r}_p \in D_{jk}$ being the image of a sample point $(r_j^*, \theta_k^*) \in D'_{jk}$ so that $f(\mathbf{r}_p) = f(r_j^* \cos\theta_k^*, r_j^* \sin\theta_k^*)$. The limit in (14.3) is understood as the double limit $(\Delta r, \Delta\theta) \to (0, 0)$. Owing to the independence of the limit of the choice of sample points, put $r_j^* = (r_{j+1} + r_j)/2$ (the midpoint rule). With this choice, $(r_{j+1} + r_j)\,\Delta r/2 = r_j^*\,\Delta r$. By taking the limit of the Riemann sum (14.3)

$$\lim_{\substack{N\to\infty \\ (R_N\to 0)}} \sum_{p=1}^{N} f(\mathbf{r}_p^*)\,\Delta A_p = \lim_{\substack{N_{1,2}\to\infty \\ (\Delta r,\Delta\theta)\to(0,0)}} \sum_{j=1}^{N_1}\sum_{k=1}^{N_2} f(r_j^* \cos\theta_k^*, r_j^* \sin\theta_k^*) r_j^*\,\Delta A',$$

one obtains the double integral of the function $f(r\cos\theta, r\sin\theta)J(r)$ over the region D' (the image of D), where $J(r) = r$ is called the *Jacobian* of the polar coordinates. The Jacobian defines the area element transformation

$$dA = J\,dA' = r\,dA'.$$

Definition 14.6. (Double Integral in Polar Coordinates). *Let D' be the image of D in the polar plane spanned by ordered pairs (r, θ) of polar coordinates. The double integral of f over D in polar coordinates is*

$$\iint\limits_D f(x, y)\,dA = \iint\limits_{D'} f(r\cos\theta, r\sin\theta)\,J(r)\,dA'\,, \quad J(r) = r\,.$$

A similarity between the double integral in rectangular and polar coordinates is that they both use partitions by corresponding coordinate curves. Note that horizontal and vertical lines are coordinate curves of the rectangular coordinates. So the very term "a double integral in polar coordinates" refers to a specific partitioning D in the Riemann sum, namely, by *coordinate curves of polar coordinates* (by circles and rays). *The double integral over D' can be evaluated by the standard means, that is, by converting it to a suitable iterated integral with respect to r and θ.*

Example 14.10. *Use polar coordinates to evaluate the double integral of $f(x, y) = xy^2\sqrt{x^2 + y^2}$ over D, which is the portion of the disk $x^2 + y^2 \leq 1$ that lies in the first quadrant.*

Solution: First, the image D' of D has to be found. Using the boundary transformation, as explained at the beginning of this section, D' is the rectangle $r \in [0, 1]$ and $\theta \in [0, \pi/2]$. Second, the function has to be written in polar coordinates, $f(r\cos\theta, r\sin\theta) = r^4\cos\theta\sin^2\theta$. Third, the double integral of this function, *multiplied by the Jacobian* r, has to be evaluated over D'. As D' is a rectangle, by Fubini's theorem, the order of integration in the iterated integral is irrelevant:

$$\iint_D f\,dA = \int_0^{\pi/2} \sin^2\theta\cos\theta\,d\theta \int_0^1 r^5 dr = \frac{1}{3}\sin^3\theta\Big|_0^{\pi/2} \cdot \frac{1}{6}\,r^6\Big|_0^1 = \frac{1}{18}.$$

□

This example shows that the technicalities involved in evaluating the double integral have been substantially simplified by passing to polar coordinates. The simplification is twofold. First, the domain of integration has been simplified; the new domain is a rectangle, which is much simpler to handle in the iterated integral than a portion of a disk. Second, the evaluation of ordinary integrals with respect to r and θ appears to be simpler than the integration of f with respect to either x or y needed in the iterated integral. However, these simplifications cannot always be achieved by converting the double integral to polar coordinates. The region D and the integrand f should have some particular properties that guarantee the observed simplifications and thereby justify the use of polar coordinates. Here are some guiding principles to decide whether the conversion of a double integral to polar coordinates could be helpful:

- The domain D is bounded by circles, lines through the origin, and polar graphs.
- The function $f(x, y)$ depends on either the combination $x^2 + y^2 = r^2$ or $y/x = \tan\theta$.

Indeed, if D is bounded only by circles centered at the origin and lines through the origin, then the image D' is a rectangle because the boundaries of D are *coordinate curves* of polar coordinates. If the boundaries of D contain circles not centered at the origin or, generally, polar graphs, that is, curves defined by the relations $r = g(\theta)$, then an algebraic description of the boundaries of D' is simpler than that of the boundaries of D. If $f(x, y) = h(u)$, where $u = x^2 + y^2 = r^2$ or $u = y/x = \tan\theta$, then in the iterated integral one of the integrations, either with respect to θ or r, becomes trivial.

EXAMPLE 14.11. *Evaluate the double integral of $f(x, y) = xy$ over the region D that lies in the first quadrant and is bounded by the circles $x^2 + y^2 = 4$ and $x^2 + y^2 = 2x$.*

SOLUTION: First, the image D' of D must be found. Using the principle that the boundaries of D are mapped onto the boundaries of D', one finds the equations of the boundaries of D' by converting the equations for the boundaries of D into polar coordinates. The boundary of the region D consists of three curves:

$$\begin{aligned} x^2 + y^2 &= 4 \to r = 2\,, \\ x^2 + y^2 &= 2x \to r = 2\cos\theta\,, \\ x = 0,\ y &\geq 0 \to \theta = \pi/2. \end{aligned}$$

So, in the polar plane, the region D' is bounded by the horizontal line $r = 2$, the graph $r = 2\cos\theta$, and the vertical line $\theta = \pi/2$. In particular, it is convenient to use an algebraic description of D' as a vertically simple region; that is, $(r, \theta) \in D'$ if $r_{\text{bot}}(\theta) = 2\cos\theta \leq r \leq 2 = r_{\text{top}}(\theta)$ and $\theta \in [0, \pi/2] = [a, b]$ (because $r_{\text{top}}(0) = r_{\text{bot}}(0)$). Second, the function is written in polar coordinates, $f(r\cos\theta, r\sin\theta) =$

$r^2 \sin\theta \cos\theta$. Multiplying it by the Jacobian $J = r$, the integrand is obtained. One has

$$\begin{aligned}\iint_D xy\,dA &= \iint_{D'} r^3 \sin\theta\cos\theta\,dA' = \int_a^b \sin\theta\cos\theta \int_{r_{\text{bot}}(\theta)}^{r_{\text{top}}(\theta)} r^3\,dr\,d\theta \\ &= \int_0^{\pi/2} \sin\theta\cos\theta \int_{2\cos\theta}^{2} r^3\,dr\,d\theta \\ &= 4\int_0^{\pi/2} (1-\cos\theta)^4 \cos\theta\sin\theta\,d\theta \\ &= 4\int_0^1 (1-u)^4 u\,du = 4\int_0^1 v^4(1-v)\,dv = \frac{4}{15},\end{aligned}$$

where two changes of variables have been used to simplify the calculations, $u = \cos\theta$ and $v = 1 - u$. □

Example 14.12. *Find the area of the region D that is bounded by two spirals $r = \theta$ and $r = 2\theta$, where $\theta \in [0, 2\pi]$, and the positive x axis.*

Before solving the problem, let us make a few comments about the shape of D. The boundaries $r = \theta$ and $r = 2\theta$ are examples of polar graphs, $r = g(\theta)$, where $g(\theta) = \theta$ and $g(\theta) = 2\theta$. They can be visualized by means of a simple geometrical procedure. Take a ray corresponding to a fixed value of the polar angle θ. On this ray, mark the point that is a distance $r = g(\theta)$ from the origin. All such points obtained for all values of θ form a curve, called the *polar graph*. When $g(\theta) = \theta$, the distance $r = \theta$ increases as the ray rotates about the origin, and the polar graph is a spiral winding about the origin. The region D lies between two spirals; it is not simple in any direction. Any smooth curve in the xy plane can always be defined by an equation $h(x, y) = 0$. In this case, by converting the polar graph equation into the rectangular coordinates, one has $\sqrt{x^2 + y^2} = \tan^{-1}(y/x)$ or $y = x\tan(\sqrt{x^2 + y^2})$. There is no way to find an analytic solution of this equation to express y as a function of x or vice versa. Therefore, had one tried to evaluate the double integral in the rectangular coordinates, one would have faced an *unsolvable* problem of finding the equations for the boundaries of D in the form $y = y_{\text{top}}(x)$ and $y = y_{\text{bot}}(x)$!

Solution: The region D is bounded by three curves, two spirals (polar graphs), and the line $y = 0$, $x \ge 0$. Their images on the polar plane are the lines $r = \theta$, $r = 2\theta$, and the vertical line $\theta = 2\pi$. They form the boundaries of D'. An algebraic description of D' as a vertically simple region is convenient to use, $(r,\theta)\in D'$ if $r_{\rm bot}(\theta)=\theta\le r\le 2\theta=r_{\rm top}(\theta)$ and $\theta\in[0,2\pi]=[a,b]$. Hence,

$$A(D)=\iint_D dA=\iint_{D'} r\,dA'=\int_0^{2\pi}\int_\theta^{2\theta} r\,dr\,d\theta=\frac{3}{2}\int_0^{2\pi}\theta^2\,d\theta=4\pi^3.$$

□

Example 14.13. *Find the volume of the portion of the solid bounded by the cone $z=2\sqrt{x^2+y^2}$ and the paraboloid $z=2-x^2-y^2$ that lies in the first octant.*

Solution: The intersection of the cone and paraboloid is a circle of unit radius. Indeed, put $r=\sqrt{x^2+y^2}$. Then the points of intersection satisfy the condition $2r=2-r^2$ or $r=1$. So the projection D of the solid onto the xy plane is the portion of the disk $r\le 1$ in the first quadrant. For any $(x,y)\in D$, the height is $h=2-r^2-2r$ (i.e., independent of the polar angle θ). The image D' of D in the polar plane is the rectangle $(r,\theta)\in[0,1]\times[0,\pi/2]$. The volume is

$$\begin{aligned}\iint_D h(x,y)\,dA &= \iint_{D'}(2-r^2-2r)r\,dA'\\ &=\int_0^{\pi/2}d\theta\int_0^1(2r-r^3-2r^2)\,dr=\frac{\pi}{24}.\end{aligned}$$

□

102. Change of Variables in Double Integrals

With an example of polar coordinates, it is quite clear that a smart choice of integration variables can significantly simplify the technicalities involved when evaluating double integrals. The simplification is twofold: simplifying the shape of the integration region (a rectangular shape is most desirable) and finding antiderivatives when calculating the iterated integral. It is therefore of interest to develop a technique for a general change of variable in the double integral so that one would be able to *design* new variables specific to the double integral in question in which the sought-for simplification is achieved.

102.1. Change of Variables. Let the functions $x(u,v)$ and $y(u,v)$ be defined on an open region D'. Then, for every pair $(u,v)\in D'$, one can find a pair (x,y), where $x=x(u,v)$ and $y=y(u,v)$. All such pairs form a region in the xy plane that is denoted D. In other words, the functions $x(u,v)$ and $y(u,v)$ define a *transformation* of a region D' in the uv plane onto a region D in the xy plane. If no two points in D' have the same image point in D, then the transformation is called *one-to-one*. For a one-to-one transformation, one can define the inverse transformation, that is, the functions $u(x,y)$ and $v(x,y)$ that assign a pair $(u,v) \in D'$ to a pair $(x,y) \in D$, where $u = u(x,y)$ and $v = v(x,y)$. Owing to this one-to-one correspondence between rectangular coordinates (x,y) and pairs (u,v), one can describe points in a plane by *new coordinates* (u,v). For example, if polar coordinates are introduced by the relations $x = x(r,\theta) = r\cos\theta$ and $y = y(r,\theta) = r\sin\theta$ for any open set D' of pairs (r,θ) that lie within the half-strip $[0,\infty)\times[0,2\pi)$, then there is a one-to-one correspondence between the pairs $(x,y) \in D$ and $(r,\theta) \in D'$. In particular, the inverse functions are $r(x,y) = \sqrt{x^2+y^2}$ and $\theta(x,y) = \tan^{-1}(y/x)$.

Definition 14.7. (Change of Variables in a Plane).
A one-to-one transformation of an open region D' defined by $x = x(u,v)$ and $y = y(u,v)$ is called a change of variables *if the functions $x(u,v)$ and $y(u,v)$ have continuous first-order partial derivatives on D'.*

The pairs (u,v) are often called *curvilinear coordinates*. Recall that a point of a plane can be described as an intersection point of two coordinate lines of a rectangular coordinate system $x = x_p$ and $y = y_p$. The point $(x_p,y_p) \in D$ is a unique image of a point $(u_p,v_p) \in D'$. Consider the inverse transformation $u = u(x,y)$ and $v = v(x,y)$. Since $u(x_p,y_p) = u_p$ and $v(x_p,y_p) = v_p$, the point $(x_p,y_p) \in D$ can be viewed as the point of intersection of two curves $u(x,y) = u_p$ and $v(x,y) = v_p$. The curves $u(x,y) = u_p$ and $v(x,y) = v_p$ are called *coordinate curves* of the new coordinates u and v; that is, the coordinate u has a fixed value along its coordinate curve $u(x,y) = u_p$, and, similarly, the coordinate v has a fixed value along its coordinate curve $v(x,y) = v_p$. The coordinate curves are images of the straight lines $u = u_p$ and $v = v_p$ in D' under the inverse transformation. If the coordinate curves are not straight lines (as in a rectangular coordinate system), then such coordinates are naturally curvilinear. For example, the coordinate curves of polar coordinates are concentric circles (a fixed value of r) and rays from the origin (a fixed value of θ), and every point in a plane can be viewed as an intersection of one such ray and one such circle.

102.2. Change of Variables in a Double Integral. Consider a double integral of a function $f(x,y)$ over a region D. Let $x = x(u,v)$ and $y = y(u,v)$ define a transformation of a region D' to D, where D' is bounded by piecewise-smooth curves in the uv plane. Suppose that the transformation is a change of variables on an open region that includes D'. Then there is an inverse transformation, that is, a transformation of D to D', which is defined by the functions $u = u(x, y)$ and $v = v(x, y)$. According to (14.3), the double integral of f over D is the limit of a Riemann sum. The limit depends neither on a partition of D by area elements nor on sample points in the partition elements. Following the analogy with polar coordinates, consider a partition of D by coordinate curves $u(x, y) = u_i$, $i = 1, 2, ..., N_1$, and $v(x, y) = v_j$, $j = 1, 2, ..., N_2$, such that $u_{i+1} - u_i = \Delta u$ and $v_{j+1} - v_j = \Delta v$. This partition of D is induced by a rectangular partition of D' by horizontal lines $v = v_j$ and vertical lines $u = u_i$ in the uv plane. Each partition element D'_{ij} of D' has the area $\Delta A' = \Delta u\, \Delta v$. Its image is a partition element D_{ij} of D. If $(u_i^*, v_j^*) \in D'_{ij}$ is a sample point, then the corresponding sample point in D_{ij} is $\mathbf{r}_{ij}^* = (x(u_i^*, v_j^*),\, y(u_i^*, v_j^*))$, and (14.3) becomes

$$\iint_D f\, dA = \lim_{N_1, N_2 \to \infty} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} f(\mathbf{r}_{ij}^*)\, \Delta A_{ij},$$

where ΔA_{ij} is the area of the partition element D_{ij}. The limit N_1, $N_2 \to \infty$ is understood in the sense of a double limit $(\Delta u, \Delta v) \to (0, 0)$. As before, the values of $f(x(u, v), y(u, v))$ outside D' are set to 0 when calculating the value of f in a partition rectangle that intersects the boundary of D'.

As in the case of polar coordinates, the aim is to convert this limit into a double integral of $f(x(u, v), y(u, v))$ over the region D'. This can be accomplished by finding a relation between ΔA_{ij} and $\Delta A'$, that is, the rule of the area element transformation under a change of variables. To simplify the notation, let the index p label partition elements D_{ij}, that is, $D_p = D_{ij}$, $(u_p, v_p) = (u_i, v_j)$, etc. Now take a point (u_p, v_p) and fix the numbers Δu, Δv. Consider a rectangle D'_p in the uv plane bounded by the lines $u = u_p$, $u = u_p + \Delta u$, $v = v_p$, and $v = v_p + \Delta v$. Let A' be the vertex (u_p, v_p), B' be $(u_p + \Delta u, v_p)$, and C' be $(u_p, v_p + \Delta v)$. The image D_p of D'_p in the xy plane is a region bounded by the coordinate curves of the variables u and v. The images A and B of the points A' and B' lie on the coordinate curve $v = v_p$, while A and C (the image of C') are on the coordinate curve $u = u_p$. The numbers Δu and Δv can be viewed as infinitesimal variations of u and v or their differentials. So, when calculating the area ΔA of D_p, it is sufficient to keep only terms *linear* in Δv and Δu; their higher powers are to be neglected (by definition of the differential), that is,

$$\Delta A = J\, \Delta u\, \Delta v = J\, \Delta A',$$

where the coefficient J is to be found. Recall that $J = r$ for polar coordinates.

Since Δu and Δv are infinitesimally small, the area of D_p can be approximated by the area of a parallelogram with adjacent sides $\vec{AB} = \mathbf{b}$ and $\vec{AC} = \mathbf{c}$. The coordinates of A are $(x(u_p, v_p), y(u_p, v_p))$, while the coordinates of B are $(x(u_p + \Delta u, v_p), y(u_p + \Delta u, v_p))$ because they are images of A' and B', respectively, under the inverse transformation $x = x(u, v)$ and $y = y(u, v)$. Therefore,

$$\begin{aligned}\mathbf{b} &= \Big(x(u_p + \Delta u, v_p) - x(u_p, v_p),\ y(u_p + \Delta u, v_p) - y(u_p, v_p),\ 0\Big)\\ &= \Big(x'_u(u_p, v_p)\,\Delta u,\ y'_u(u_p, v_p)\,\Delta u,\ 0\Big)\\ &= \Delta u\Big(x'_u(u_p, v_p),\ y'_u(u_p, v_p),\ 0\Big),\end{aligned}$$

where $x(u_p + \Delta u, v_p) = x(u_p, v_p) + x'_u(u_p, v_p)\,\Delta u$ has been linearized, that is, higher powers of Δu are neglected, and similarly for $y(u_p + \Delta u, v_p)$; the third component of $\mathbf{b}$ is set to 0 as the vector is planar. An analogous calculation for the components of $\mathbf{c}$ yields

$$\mathbf{c} = \Delta v\Big(x'_v(u_p, v_p),\ y'_v(u_p, v_p),\ 0\Big).$$

The area of the parallelogram reads

$$\Delta A = \|\mathbf{b} \times \mathbf{c}\| = \left|\det\begin{pmatrix} x'_u & x'_v \\ y'_u & y'_v \end{pmatrix}\right| \Delta u\,\Delta v = J\,\Delta u\,\Delta v\,. \tag{14.10}$$

Note that the vectors $\mathbf{b}$ and $\mathbf{c}$ are in the xy plane. Therefore, their cross product has only one nonzero component (the z component) given by the determinant. The absolute value of the determinant is needed because the z component of the cross product may be negative, $\|(0, 0, z)\| = \sqrt{z^2} = |z|$.

Definition 14.8. (Jacobian of a Transformation).
The Jacobian of a transformation defined by $x = x(u, v)$ and $y = y(u, v)$ is

$$\frac{\partial(x, y)}{\partial(u, v)} = \det\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = x'_u y'_v - x'_v y'_u\,.$$

The Jacobian coincides with the determinant in (14.10). In this definition, a convenient notation has been introduced. The matrix whose determinant is evaluated has the *first* row composed of the partial derivatives of the *first* variable the numerator with respect to all variables in the denominator, and similarly for the second row. This rule is easy to remember.

Furthermore, the coefficient J in (14.10) is the *absolute value* of the Jacobian. The Jacobian of a change of variables in the double integral should not vanish on D' because $\Delta A \neq 0$. Since the partial derivatives of x and y with respect to u and v are continuous, J is continuous, too. Therefore, for any partition element D_{ij}, the difference $(\Delta A_{ij} - J(u_i^*, v_j^*)\Delta A')/\Delta A'$ vanishes in the limit $(\Delta u, \Delta v) \to (0, 0)$. So, in this limit, one can put $\Delta A_{ij} = J(u_i^*, v_j^*)\,\Delta u\,\Delta v$ in the Riemann sum. The limit of the Riemann sum defines the double integral of the function $f(x(u,v), y(u,v))J(u,v)$ over the region D'. The foregoing arguments suggest that the following theorem is true (a full proof is given in advanced calculus courses).

Theorem 14.6. (Change of Variables in a Double Integral). *Suppose a transformation $x = x(u,v)$, $y = y(u,v)$ has continuous first-order partial derivatives and maps a region D' bounded by piecewise-smooth curves onto a region D. Suppose that this transformation is one-to-one and has a nonvanishing Jacobian, except perhaps on the boundary of D'. Then*

$$\iint_D f(x,y)\,dA = \iint_{D'} f(x(u,v), y(u,v))J(u,v)\,dA'\,,$$
$$J(u,v) = \left|\frac{\partial(x,y)}{\partial(u,v)}\right|\,.$$

Note that in the case of polar coordinates, the boundary of D' may contain the line $r = 0$ on which the Jacobian $J = r$ vanishes. This entire line collapses into a single point, the origin $(x,y) = (0,0)$ in the xy plane, upon the transformation $x = r\cos\theta$ and $y = r\sin\theta$; that is, this transformation is not one-to-one on this line. A full proof of the theorem requires an analysis of such subtleties in a general change of variables, which was excluded in the above derivation by assuming that the transformation is a *genuine* change of variables on a region that contains D'.

The change of variables in a double integral entails the following steps:

1. Finding the image D' of D under the inverse transformation $u = u(x,y)$, $v = v(x,y)$. A useful rule to remember here is:

   $$\text{boundaries of } D \longleftrightarrow \text{boundaries of } D';$$

   that is, if equations of boundaries of D are given, then equations of the corresponding boundaries of D' can be obtained by expressing the former in the new variables by the substitution $x = x(u,v)$ and $y = y(u,v)$.

2. Transformation of the function to new variables
$$f(x,y) = f(x(u,v), y(u,v)).$$
3. Calculation of the Jacobian that defines the area element transformation:
$$dA = \left|\frac{\partial(x,y)}{\partial(u,v)}\right| du\,dv = J\,dA', \qquad J = \left|\frac{\partial(x,y)}{\partial(u,v)}\right| .$$
4. Evaluation of the double integral of fJ over D' by converting it to a suitable iterated integral. The choice of new variables should be motivated by simplifying the shape D' (a rectangular shape is the most desirable).

When calculating the Jacobian, the following statement, given without a proof, might be useful.

Corollary 14.4. *If $u = u(x,y)$ and $v = v(x,y)$ is the inverse of the transformation $x = x(u,v)$ and $y = y(u,v)$, then its Jacobian is*

$$\frac{\partial(x,y)}{\partial(u,v)} = \frac{1}{\frac{\partial(u,v)}{\partial(x,y)}} = \frac{1}{\det\begin{pmatrix} u'_x & u'_y \\ v'_x & v'_y \end{pmatrix}}. \tag{14.11}$$

Equation (14.11) defines the Jacobian as a function of (x,y). Sometimes it is technically simpler to express the product $f(x,y)J(x,y)$ in the new variables rather than doing so for f and J separately. This is illustrated by the following example.

Example 14.14. *Use a suitable change of variables to evaluate the double integral of $f(x,y) = xy^3$ over the region D that lies in the first quadrant and is bounded by the lines $y = x$ and $y = 3x$ and by the hyperbolas $yx = 1$ and $yx = 2$.*

Solution: The line equations can be written in the form $y/x = 1$ and $y/x = 3$ because $y, x > 0$ in D. Note that the equations of boundaries of D depend on just two particular combinations y/x and yx that take constant values on the boundaries of D. So, if the new variables defined by the relations $u = u(x,y) = y/x$ and $v = v(x,y)$, then the image region D' in the uv plane is a rectangle $u \in [1,3]$ and $v \in [1,2]$. Indeed, the boundaries $y/x = 1$ and $y/x = 3$ are mapped onto the vertical lines $u = 1$ and $u = 3$, while the hyperbolas $yx = 1$ and $yx = 2$ are mapped onto the horizontal lines $v = 1$ and $v = 2$. Let us put aside for a moment the problem of expressing x and y as functions of new variables, which is needed to express f and J as functions of u and v, and find first the Jacobian as a function of x and y by means of (14.11):

$$J = \left|\det\begin{pmatrix} u'_x & u'_y \\ v'_x & v'_y \end{pmatrix}\right|^{-1} = \left|\det\begin{pmatrix} -y/x^2 & 1/x \\ y & x \end{pmatrix}\right|^{-1} = |-2y/x|^{-1} = \frac{x}{2y}\,.$$

Note that x and y are strictly positive in D. The integrand becomes $fJ = x^2y^2/2 = v^2/2$. So finding the functions $x = x(u, x)$ and $y = y(u, v)$ happens to be unnecessary in this example! Hence,

$$\iint_D xy^3\,dA = \frac{1}{2}\iint_{D'} v^2\,dA' = \frac{1}{2}\int_1^3 du\int_1^2 v^2\,dv = 7/3.$$

The reader is advised to evaluate the double integral in the original rectangular coordinates to compare the amount of work needed with this solution. □

The following example illustrates how a change of variables can be used to simplify the integrand of a double integral.

Example 14.15. *Evaluate the double integral of the function* $f(x, y) = \cos[(y - x)/(y + x)]$ *over the trapezoidal region with vertices* $(1, 0)$, $(2, 0)$, $(0, 1)$, *and* $(0, 2)$.

Solution: An iterated integral in the rectangular coordinates would contain the integral of the cosine function of a rational argument (either with respect to x or y), which is difficult to evaluate. So a change of variables should be used to simplify the argument of the cosine function. The region D is bounded by the lines $x + y = 1$, $x + y = 2$, $x = 0$, and $y = 0$. Put $u = x + y$ and $v = y - x$ so that the function in the new variables becomes $f = \cos(v/u)$. The lines $x + y = 1$ and $x + y = 2$ are mapped onto the vertical lines $u = 1$ and $u = 2$. Since $y = (u + v)/2$ and $x = (u - v)/2$, the line $x = 0$ is mapped onto the line $v = u$, while the line $y = 0$ is mapped onto the line $v = -u$. Thus, the region $D' = \{(u, v)| - u \le v \le u,\ u \in [1, 2]\}$. The Jacobian of the change of variables is $J = 1/2$. Hence,

$$\begin{aligned}\iint_D \cos\Big(\frac{y-x}{y+x}\Big)dA &= \frac{1}{2}\iint_{D'}\cos\Big(\frac{v}{u}\Big)dA' = \frac{1}{2}\int_1^2\int_{-u}^{u}\cos\Big(\frac{v}{u}\Big)dv\,du \\ &= \frac{1}{2}\int_1^2 u\sin\Big(\frac{v}{u}\Big)\Big|_{-u}^{u}\,du \\ &= \sin(1)\int_1^2 udu = 3\sin(1)/2.\end{aligned}$$

□

Example 14.16. (Area of an Ellipse).
Find the area of the region D bounded by the ellipse $x^2/a^2 + y^2/b^2 = 1$.

Solution: Under the change of variables $u = x/a$, $v = y/b$, the ellipse is transformed into the circle $u^2 + v^2 = 1$ of unit radius. Since the Jacobian of the transformation is $J = ab$,

$$A(D) = \iint_D dA = \iint_{D'} J\, dA' = ab \iint_{D'} dA' = abA(D') = \pi ab\,.$$

□

When $a = b$, the ellipse becomes a circle of radius $R = a = b$, and the area of the ellipse becomes the area of the disk, $A = \pi R^2$.

102.3. Symmetries and a Change of Variables. As noted earlier, the symmetry properties of double integrals are quite helpful for their evaluation. A transformation $x = x(u, v)$, $y = y(u, v)$ that maps D' onto D is said to be area preserving if the absolute value of its Jacobian is 1, that is, $dA = dA'$, from which it immediately follows that the areas of D and D' coincide, $A(D) = A(D')$. For example, rotations, translations, and reflections are area-preserving transformations for obvious geometrical reasons. The following theorem holds.

Theorem 14.7. *Suppose that an area-preserving transformation $x = x(u, v)$, $y = y(u, v)$ maps a region D onto itself. Suppose that a function f is skew-symmetric under this transformation, that is, $f(x(u, v), y(u, v)) = -f(u, v)$. Then the double integral of f over D vanishes.*

Proof. Since $D' = D$ and $dA = dA'$ (i.e., $dx\,dy = du\,dv$), the change of variables yields

$$\begin{aligned} I = \iint_D f(x, y)\, dA &= \iint_D f(x(u, v), y(u, v))\, dA' \\ &= -\iint_D f(u, v)\, dA' = -I, \end{aligned}$$

that is, $I = -I$, or $I = 0$. □

103. Triple Integrals

Suppose a solid region E is filled with an inhomogeneous material. The latter means that, if a small volume ΔV of the material is taken at two distinct points of E, then the masses of these two pieces are different, despite the equality of their volumes. The inhomogeneity of the material can be characterized by the *mass density* as a function of position. Let $\Delta m(\mathbf{r})$ be the mass of a small piece of material of volume ΔV cut out around a point $\mathbf{r}$. Then the mass density is defined by

$$\sigma(\mathbf{r}) = \lim_{\Delta V \to 0} \frac{\Delta m(\mathbf{r})}{\Delta V}.$$

The limit is understood in the following sense. If R is the radius of the smallest ball that contains the region of volume ΔV, then the limit means that $R \to 0$ (i.e., roughly speaking, all the dimensions of the piece decrease simultaneously in the limit). The mass density is measured in units of mass per unit volume. For example, the value $\sigma(\mathbf{r}) = 5$ g/cm^3 means that a piece of material of volume 1 cm^3 cut out around the point $\mathbf{r}$ has a mass of 5 gr.

Suppose that the mass density of the material in a region E is known. The question is: What is the total mass of the material in E? A practical answer to this question is to partition the region E so that each partition element E_p, $p = 1, 2, ..., N$, has a mass Δm_p. The total mass is $M = \sum_p \Delta m_p$. If a partition element E_p has a volume ΔV_p, then $\Delta m_p \approx \sigma(\mathbf{r}_p)\, \Delta V_p$ for some $\mathbf{r}_p \in E_p$. If R_p is the radius of the smallest ball that contains E_p, put $R_N = \max R_p$. Then, by increasing the number N of partition elements so that $R_p \le R_N \to 0$ as $N \to \infty$, the approximation $\Delta m_p \approx \sigma(\mathbf{r}_p)\, \Delta V_p$ becomes more and more accurate by the definition of the mass density because $\Delta V_p \to 0$ for all p. So the total mass is

$$M = \lim_{\substack{N\to\infty \\ (R_N \to 0)}} \sum_{p=1}^{N} \sigma(\mathbf{r}_p)\, \Delta V_p\,, \tag{14.12}$$

which is to be compared with (14.1). In contrast to (14.1), the summation over the partition should include a triple sum, one sum per each direction in space. This gives an intuitive idea of a triple integral. Its abstract mathematical construction follows exactly the footsteps of the double-integral construction.

103.1. Definition of a Triple Integral.

Rectangular Partition. A region E in space is assumed to be bounded; that is, it is contained in a ball of some (finite) radius. The boundaries of E are assumed to be piecewise-smooth surfaces. A smooth surface can be viewed as a level surface of a differentiable function of three variables. The region E is then embedded in a rectangle $R_E = [a, b] \times [c, d] \times [s, q]$, that is, $x \in [a, b]$, $y \in [c, d]$, and $z \in [s, q]$. If $f(\mathbf{r})$ is a bounded function on E, then it is extended to R_E by setting its values to 0 outside E. The rectangle R_E is partitioned by the coordinate planes $x = x_i = a + i\, \Delta x$, $i = 0, 1, ..., N_1$, where $\Delta x = (b-a)/N_1$;

$y = y_j = c + j\,\Delta y$, $i = 0, 1, ..., N_2$, where $\Delta y = (d - c)/N_2$; and $z = z_i = s + k\,\Delta z$, $k = 0, 1, ..., N_3$, where $\Delta z = (q - s)/N_3$. The volume of each partition element is a rectangle R_{ijk} of volume $\Delta V = \Delta x\,\Delta y\,\Delta z$. The total number of rectangles is $N = N_1 N_2 N_3$.

Upper and Lower Sums. By analogy with Definition 14.2, the lower and upper sums are defined. Put $M_{ijk} = \sup f(\mathbf{r})$, and $m_{ijk} = \inf f(\mathbf{r})$ where the supremum and infimum are taken over the partition rectangle R_{ijk}. Then the upper and lower sums are

$$U(f, \mathbf{N}) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sum_{k=1}^{N_3} M_{ijk}\,\Delta V\,, \quad L(f, \mathbf{N}) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sum_{k=1}^{N_3} m_{ijk}\,\Delta V,$$

where $\mathbf{N} = (N_1, N_2, N_3)$.

Definition 14.9. (Triple Integral).
If the limits of the upper and lower sums exist as $N_{1,2,3} \to \infty$ (or $(\Delta x, \Delta y, \Delta z) \to (0, 0, 0)$) and coincide, then f is said to be Riemann integrable on E, and the limit of the upper and lower sums

$$\iiint_E f(x, y, z)\,dV = \lim_{\mathbf{N}\to\infty} U(f, \mathbf{N}) = \lim_{\mathbf{N}\to\infty} L(f, \mathbf{N})$$

is called the triple integral *of f over the region E.*

The limit is understood as a three-variable limit $(\Delta x, \Delta y, \Delta z) \to (0, 0, 0)$.

103.2. Properties of Triple Integrals. The properties of triple integrals are the same as those of the double integral discussed in Section 98; that is, the linearity, additivity, positivity, integrability of the absolute value $|f|$, and upper and lower bounds holds for triple integrals.

Continuity and Integrability. The relation between continuity and integrability is pretty much the same as in the case of double integrals.

Theorem 14.8. (Integrability of Continuous Functions).
Let E be a closed, bounded spatial region whose boundaries are piecewise-smooth surfaces. If a function f is continuous on E, then it is integrable on E. Furthermore, if f has bounded discontinuities only on a finite number of smooth surfaces in E, then it is also integrable on E.

In particular, a constant function is integrable, and the volume of a region E is given by the triple integral

$$V(E) = \iiint_E dV.$$

The Integral Mean-Value Theorem. The integral mean value theorem (Theorem 14.3) is extended to triple integrals, where the double integral is replaced by the triple integral and $A(D)$ by the volume $V(E)$. Its proof follows the same lines as in the case of double integrals.

Riemann Sums. If a function f is integrable, then its triple integral is the limit of a Riemann sum, and its value is independent of the partition of E and a choice of sample points in the partition elements, that is, (14.12) holds:

$$\iiint_E f(\mathbf{r})\,dV = \lim_{\substack{N\to\infty \\ (R_N\to 0)}} \sum_{p=1}^{N} f(\mathbf{r}_p)\,\Delta V_p\,. \tag{14.13}$$

This equation can be used for approximations of triple integrals, when evaluating the latter numerically just like in the case of double integrals.

Symmetry. If a transformation in space preserves the volume of any region, then it is called *volume preserving*. Obviously, rotations, reflections, and translations in space are volume-preserving transformations. Suppose that, under a volume-preserving transformation, a region E is mapped onto itself; that is, E is *symmetric* relative to this transformation. If $\mathbf{r}_s \in E$ is the image of $\mathbf{r} \in E$ under this transformation and the integrand is skew-symmetric, $f(\mathbf{r}_s) = -f(\mathbf{r})$, then the triple integral of f over E vanishes.

Example 14.17. *Evaluate the triple integral of* $f(x,y,z) = x^2 \sin(y^4 z) + 2$ *over a ball centered at the origin of radius* R.

Solution: Put $g(x,y,z) = x^2\sin(y^4z)$ so that $f = g + h$, where $h = 2$ is a constant function. By the linearity property, the triple integral of f is the sum of triple integrals of g and h over the ball. The ball is symmetric relative to the reflection transformation $(x,y,z) \to (x,y,-z)$, whereas the function g is skew-symmetric, $g(x,y,-z) = -g(x,y,z)$. Therefore, its triple integral vanishes, and

$$\begin{aligned} \iiint_E f\,dV &= \iiint_E g\,dV + \iiint_E h\,dV \\ &= 0 + 2\iiint_E dV = 2V(E) = 8\pi R^3/3. \end{aligned}$$

□

One can think of the numerical value of a triple integral of f over E as the total amount of a quantity distributed in the region E with the density f (the amount of the quantity per unit volume). For example, f can be viewed as the density of electric charge distributed in a dielectric occupying a region E. The total electric charge stored in the region E is then given by triple integral of the density over E. The electric charge can be positive and negative. So, if the total positive charge in E is exactly the same as the negative charge, the triple integral vanishes.

103.3. Iterated Triple Integrals. Similar to a double integral, a triple integral can be converted to a triple iterated integral, which can then be evaluated by means of ordinary single-variable integration.

Definition 14.10. (Simple Region).
A spatial region E is said to be simple in the direction of a vector $\mathbf{v}$ if any straight line parallel to $\mathbf{v}$ intersects E along at most one straight line segment.

A triple integral can be converted to an iterated integral if E is simple in a particular direction. If there is no such direction, then E should be split into a union of simple regions with the consequent use of the additivity property of triple integrals. Suppose that $\mathbf{v} = \mathbf{e}_3$; that is, E is simple along the z axis. Then the region E admits the following description:

$$E = \{(x, y, z)|z_{\rm bot}(x, y) \le z \le z_{\rm top}(x, y),\ (x, y) \in D_{xy}\}\,.$$

Indeed, consider all lines parallel to the z axis that intersect E. These lines also intersect the xy plane. The region D_{xy} in the xy plane is the set of all such points of intersection. One might think of D_{xy} as a shadow made by the solid E when it is illuminated by rays of light parallel to the z axis. Take any line through $(x, y) \in D_{xy}$ parallel to the z axis. By the simplicity of E, any such line intersects E along a single segment. If $z_{\rm bot}$ and $z_{\rm top}$ are the minimal and maximal values of the z coordinate along the intersection segment, then, for any $(x, y, z) \in E$, $z_{\rm bot} \le z \le z_{\rm top}$ and any $(x, y) \in D_{xy}$. Naturally, the values $z_{\rm bot}$ and $z_{\rm top}$ may depend on $(x, y) \in D_{xy}$. Thus, the region E is bounded from the top by the graph $z = z_{\rm top}(x, y)$ and from the bottom by the graph $z = z_{\rm bot}(x, y)$. If E is simple along the y or x axis, then E admits similar descriptions:

$$E = \{(x, y, z)|\, y_{\rm bot}(x, z) \le y \le y_{\rm top}(x, z),\ (x, z) \in D_{xz}\}\,, \tag{14.14}$$

$$E = \{(x, y, z)|\, x_{\rm bot}(y, z) \le x \le x_{\rm top}(y, z),\ (y, z) \in D_{yz}\}\,, \tag{14.15}$$

where D_{xz} and D_{yz} are projections of E into the xz and yz planes, respectively; they are defined analogously to D_{xy}.

According to (14.13), the limit of the Riemann sum is independent of partitioning E and choosing sample points. Let D_p, $p = 1, 2, ..., N$, partition the region D_{xy}. Consider a portion E_p of E that is projected on the partition element D_p; E_p is a column with D_p its cross section by a horizontal plane. Since E is bounded, there are numbers s and q such that $s \leq z_{\text{bot}}(x,y) \leq z_{\text{top}}(x,y) \leq q$ for all $(x,y) \in D_{xy}$; that is, E always lies between two horizontal planes $z = s$ and $z = q$. Consider slicing the solid E by equispaced horizontal planes $z = s + k\,\Delta z$, $k = 0, 1, ..., N_3$, $\Delta z = (q - s)/N_3$. Then each column E_p is partitioned by these planes into small regions E_{pk}. The union of all E_{pk} forms a partition of E, which will be used in the Riemann sum (14.13). The volume of E_{pk} is $\Delta V_{pk} = \Delta z\, \Delta A_p$, where ΔA_p is the area of D_p. Assuming, as usual, that f is defined by zero values outside E, sample points may be selected so that, if $(x_p, y_p, 0) \in D_p$, then $(x_p, y_p, z_k^*) \in E_{pk}$, that is, $z_{k-1} \leq z_k^* \leq z_k$ for $k = 1, 2, ..., N_3$. The three-variable limit (14.13) exists and hence can be taken in any particular order. Take first the limit $N_3 \to \infty$ or $\Delta z \to 0$. The double limit of the sum over the partition of D_{xy} is understood as before; that is, as $N \to 0$, the radii R_p of smallest disks containing D_p go to 0 uniformly, $R_p \leq R_N \to 0$. Therefore,

$$\begin{aligned}\iiint_E f\,dV &= \lim_{\substack{N\to\infty \\ (R_N\to 0)}} \sum_{p=1}^{N} \left(\lim_{N_3\to\infty} \sum_{k=1}^{N_3} f(x_p, y_p, z_k^*)\,\Delta z \right) \Delta A_p \\ &= \lim_{\substack{N\to\infty \\ (R_N\to 0)}} \sum_{p=1}^{N} \left(\int_{z_{\text{bot}}(x_p,y_p)}^{z_{\text{top}}(x_p,y_p)} f(x_p, y_p, z)\,dz \right) \Delta A_p\end{aligned}$$

because, for every $(x_p, y_p) \in D_{xy}$, the function f vanishes outside the interval $z \in [z_{\text{bot}}(x_p, y_p), z_{\text{top}}(x_p, y_p)]$. The integration of f with respect to z over the interval $[z_{\text{bot}}(x,y), z_{\text{top}}(x,y)]$ defines a function $F(x,y)$ whose values $F(x_p, y_p)$ at sample points in the partition elements D_p appear in the parenthese. A comparison of the resulting expression with (14.3) leads to the conclusion that, after taking the second limit, one obtains the double integral of $F(x,y)$ over D_{xy}.

THEOREM 14.9. (Iterated Triple Integral).
Let f be integrable on a solid region E. Suppose that E is simple in the z direction so that it is bounded by the graphs $z = z_{\text{bot}}(x,y)$ and $z = z_{\text{top}}(x,y)$ for $(x,y) \in D_{xy}$. Then

$$\begin{aligned}\iiint_E f(x,y,z)\,dV &= \iint_{D_{xy}} \int_{z_{\text{bot}}(x,y)}^{z_{\text{top}}(x,y)} f(x,y,z)\,dz\,dA \\ &= \iint_{D_{xy}} F(x,y)\,dA.\end{aligned}$$

103.3.1. Evaluation of Triple Integrals. In practical terms, an evaluation of a triple integral over a region E is carried out by the following steps:

Step 1. Determine the direction along which E is simple. If no such direction exists, split E into a union of simple regions and use the additivity property. For definitiveness, suppose that E happens to be z simple.
Step 2. Find the projection D_{xy} of E into the xy plane.
Step 3. Find the bottom and top boundaries of E as the graphs of some functions $z = z_{\text{bot}}(x, y)$ and $z = z_{\text{top}}(x, y)$.
Step 4. Evaluate the integral of f with respect to z to obtain $F(x, y)$.
Step 5. Evaluate the double integral of $F(x, y)$ over D_{xy} by converting it to a suitable iterated integral.

Similar iterated integrals can be written when E is simple in the y or x direction. According to (14.14) (or (14.15)), the first integration is carried out with respect to y (or x), and the double integral is evaluated over D_{xz} (or D_{yz}). If E is simple in any direction, then any of the iterated integrals can be used. In particular, just like in the case of double integrals, the choice of an iterated integral for a simple region E should be motivated by the simplicity of an algebraic description of the top and bottom boundaries or by the simplicity of the integrations involved. Technical difficulties may strongly depend on the order in which the iterated integral is evaluated.

Fubini's theorem can be extended to triple integrals.

Theorem 14.10. (Fubini's Theorem).
Let f be integrable on a rectangular region $E = [a, b] \times [c, d] \times [s, q]$. Then

$$\iiint_E f\,dV = \int_a^b \int_c^d \int_s^q f(x, y, z)\,dz\,dy\,dx$$

and the iterated integral can be evaluated in any order.

Here $D_{xy} = [a, b] \times [c, d]$, and the top and bottom boundaries are the planes $z = q$ and $z = s$. Alternatively, one can take $D_{yz} = [c, d] \times [s, q]$, $x_{\text{bot}}(y, z) = a$, and $x_{\text{top}}(y, z) = b$ to obtain an iterated integral in a different order (where the x integration is carried out first).

Example 14.18. *Evaluate the triple integral of $f(x, y, z) = xy^2z^3$ over the rectangle $E = [0, 2] \times [1, 2] \times [0, 3]$.*

Solution: By Fubini's theorem,

$$\iiint_E xy^2z^3\,dV = \int_0^2 x\,dx \int_1^2 y^2\,dy \int_0^3 z^3\,dz = 2 \cdot (7/3) \cdot 9 = 52.$$

This example shows that the factorization property also holds for triple integrals (see Corollary 14.3). □

Example 14.19. *Evaluate the triple integral of $f(x,y,z) = (x^2 + y^2)z$ over the portion of the solid bounded by the cone $z = \sqrt{x^2+y^2}$ and paraboloid $z = 2 - x^2 - y^2$ in the first octant.*

Solution: Following the step-by-step procedure outlined above, the integration region is z simple. The top boundary is the graph of $z_{\text{top}}(x,y) = 2 - x^2 - y^2$, and the graph of $z_{\text{bot}}(x,y) = \sqrt{x^2+y^2}$ is the bottom boundary. To determine the region D_{xy}, note that it has to be bounded by the projection of the curve of the intersection of the cone and paraboloid onto the xy plane. The intersection curve is defined by $z_{\text{bot}} = z_{\text{top}}$ or $r = 2-r^2$, where $r = \sqrt{x^2+y^2}$, and hence $r = 1$, which is the circle of unit radius. Since E is in the first octant, D_{xy} is the quarter of the disk of unit radius in the first quadrant. One has

$$\begin{aligned}\iiint_E (x^2+y^2)z\,dV &= \iint_{D_{xy}} (x^2+y^2)\int_{\sqrt{x^2+y^2}}^{2-x^2-y^2} z\,dz\,dA\\ &= \frac{1}{2}\iint_{D_{xy}} (x^2+y^2)[(2-x^2-y^2)^2 - (x^2+y^2)]\,dA\\ &= \frac{1}{2}\int_0^{\pi/2} d\theta \int_0^1 r^2[(2-r^2)^2 - r^2]r\,dr\\ &= \frac{\pi}{8}\int_0^1 u[(2-u)^2 - u]\,du = \frac{7\pi}{96},\end{aligned}$$

where the double integral has be transformed into polar coordinates because D_{xy} becomes the rectangle $D'_{xy} = [0,1] \times [0,\pi/2]$ in the polar plane. The integration with respect to r is carried out by the change of variable $u = r^2$. □

Example 14.20. *Evaluate the triple integral of $f(x,y,z) = \sqrt{y^2+z^2}$ over the region E bounded by the paraboloid $x = y^2 + z^2$ and the plane $x = 4$.*

Solution: It is convenient to choose an iterated integral for E described as an x simple region (see (14.15)). There are two reasons for doing so. First, the integrand f is independent of x, and hence the first integration with respect to x is trivial. Second, the boundaries of E are already given in the form required by (14.15), that is, $x_{\text{bot}}(y,z) = y^2 + z^2$ and $x_{\text{top}}(y,z) = 4$. The region D_{yz} is determined by the curve of intersection of the boundaries of E, $x_{\text{top}} = x_{\text{bot}}$ or $y^2 + z^2 = 4$. Therefore, D_{yz} is the disk or radius 2. One has

$$\iiint_E \sqrt{y^2+z^2}dV = \iint_{D_{yz}} \sqrt{y^2+z^2}\int_{y^2+z^2}^{4} dx\,dA$$

$$= \iint_{D_{yz}} \sqrt{y^2+z^2}\ [4-(y^2+z^2)]\,dA$$
$$= \int_0^{2\pi} d\theta \int_0^2 r[4-r^2]r\,dr = \frac{128\pi}{15},$$

where the double integral over D_{yz} has been converted to polar coordinates in the yz plane. □

103.4. Study Problems.

Problem 14.2. *Evaluate the triple integral of $f(x,y,z)=z$ over the region E bounded by the cylinder $y^2+z^2=1$ and the planes $z=0$, $y=1$, and $y=x$ in the first octant.*

Solution: The region is z simple and bounded by the xy plane from the bottom (i.e., $z_{\text{bot}}(x,y)=0$), and by the cylinder from the top (i.e., $z_{\text{top}}(x,y)=\sqrt{1-y^2}$) (by taking the positive solution of $y^2+z^2=1$). The region D_{xy} is bounded by the lines of intersection of the planes $x=0$ and $y=x$ and the cylinder $y^2+z^2=1$ with the xy plane (or the plane $z=0$). The cylinder intersects this plane along the line $y=1$ in the first quadrant. Thus, D_{xy} is the triangle bounded by the lines $x=0$, $y=0$, and $y=x$. One has

$$\iiint_E z\,dV = \iint_{D_{xy}} \int_0^{\sqrt{1-y^2}} z\,dz\,dA$$
$$= \frac{1}{2}\iint_{D_{xy}} (1-y^2)\,dA = \frac{1}{2}\int_0^1 (1-y^2)\int_0^y dx\,dy = \frac{1}{8},$$

where the double integral has been evaluated by using the description of D_{xy} as a horizontally simple region, $x_{\text{bot}}=0\le x\le y=x_{\text{top}}$ for all $y\in[0,1]=[c,d]$. □

Problem 14.3. *Evaluate the triple integral of the function $f(x,y,z)=xy^2z^3$ over the region E that is a ball of radius 3 centered at the origin with a cubic cavity $[0,1]\times[0,1]\times[0,1]$.*

Solution: The region E is not simple in any direction. The additivity property must be used. Let E_1 be the ball and let E_2 be the cavity. By the additivity property,

$$\iiint_E xy^2z^3\,dV = \iiint_{E_1} xy^2z^3\,dV - \iiint_{E_2} xy^2z^3\,dV$$
$$= 0 - \int_0^1 x\,dx \int_0^1 y^2\,dy \int_0^1 z^3\,dz = -\frac{1}{12}.$$

The triple integral over E_1 vanishes by the symmetry argument (the ball is symmetric under the reflection $(-x, y, z) \to (-x, y, z)$ whereas $f(-x, y, z) = -f(x, y, z)$). The second integral is evaluated by Fubini's theorem. □

104. Triple Integrals in Cylindrical and Spherical Coordinates

A change of variables has been proved to be quite useful in simplifying the technicalities involved in evaluating double integrals. An essential advantage is a simplification of the integration region. The concept of changing variables can be extended to triple integrals.

104.1. Cylindrical Coordinates. One of the simplest examples of curvilinear coordinates in space is cylindrical coordinates. They are defined by

$$x = r\cos\theta\,, \quad y = r\cos\theta\,, \quad z = z\,. \tag{14.16}$$

In any plane parallel to the xy plane, the points are labeled by polar coordinates, while the z coordinate is not transformed. Equations (14.16) define a transformation of an ordered triple of numbers (r, θ, z) to another ordered triple (x, y, z). A set of triples (r, θ, z) can be viewed as a set of points E' in a Euclidean space in which the coordinate axes are spanned by r, θ, and z. Then, under the transformation (14.16), the region E' is mapped onto an *image* region E. From the study of polar coordinates, the transformation (14.16) is one-to-one if $r \in (0, \infty)$, $\theta \in [0, 2\pi)$, and $z \in (-\infty, \infty)$. The inverse transformation is given by

$$r = \sqrt{x^2 + y^2}\,, \quad \theta = \tan^{-1}(y/x)\,, \quad z = z\,,$$

where the value of $\tan^{-1}$ is taken according to the quadrant in which the pair (x, y) belong (see the discussion of polar coordinates). It maps any region E in the Euclidean space spanned by (x, y, z) onto the image region E'. To find the shape of E' as well as its algebraic description, the same strategy as in the two-variable case should be used:

$$\text{boundaries of } E \quad \longleftrightarrow \quad \text{boundaries of } E'$$

under the transformation (14.16) and its inverse. A particularly important question is how to investigate the shape of *coordinate surfaces* of cylindrical coordinates, that is, surfaces on which each of the cylindrical coordinates has a constant value. If E is bounded by coordinate surfaces only, then its image E' is a rectangle, which is the simplest, most desirable, shape when evaluating a multiple integral.

The coordinate surfaces of r are cylinders, $r = \sqrt{x^2 + y^2} = r_0$ or $x^2 + y^2 = r_0^2$. In the xy plane, the equation $\theta = \theta_0$ defines a ray from the origin at the angle θ_0 to the positive x axis counted counterclockwise. Since θ depends only on x and y, the coordinate surface of θ is the half-plane bounded by the z axis that makes an angle θ_0 with the xz plane (it is swept by the ray when the latter is moved parallel up and down along the z axis). Since the z coordinate is not changed, neither changes its coordinate surfaces; they are planes parallel to the xy plane. So the coordinate surfaces of cylindrical coordinates are

$$\begin{aligned} &r = r_0 \leftrightarrow x^2 + y^2 = r_0^2 &&\text{(cylinder)},\\ &\theta = \theta_0 \leftrightarrow y\cos\theta_0 = x\sin\theta_0 &&\text{(half-plane)},\\ &z = z_0 \leftrightarrow z = z_0 &&\text{(plane)}. \end{aligned}$$

A point in space corresponding to an ordered triple (r_0, θ_0, z_0) is an intersection point of a cylinder, half-plane bounded by the cylinder axis, and a plane perpendicular to the cylinder axis.

Example 14.21. *Find the image E' of the solid region E that is bounded by the paraboloid $z = x^2 + y^2$ and the planes $z = 4$, $y = x$, and $y = 0$ in the first octant.*

Solution: In cylindrical coordinates, the equations of boundaries become, respectively, $z = r^2$, $z = 4$, $\theta = \pi/4$, and $\theta = 0$. Since E lies below the plane $z = 4$ and above the paraboloid $z = r^2$, the range of r is determined by their intersection $4 = r^2$ or $r = 2$ as $r \geq 0$. Thus,

$$E' = \Big\{(r, \theta, z) \,|\, r^2 \leq z \leq 4\,, (r, \theta) \in [0, 2] \times [0, \pi/4]\Big\}.$$

□

104.1.1. Triple Integrals in Cylindrical Coordinates. To change variables in a triple integral to cylindrical coordinates, one has to consider a partition of the integration region E by *coordinate surfaces*, that is, by cylinders, half-planes, and horizontal planes, which corresponds to a rectangular partition of E' (the image of E under the transformation from rectangular to cylindrical coordinates). Then the limit of the corresponding Riemann sum (14.13) has to be evaluated. In the case of cylindrical coordinates, this task can be accomplished by simpler means.

Suppose E is z simple so that by Theorem 14.9 the triple integral can be written as an iterated integral consisting of a double integral over D_{xy} and an ordinary integral with respect to z. The transformation (14.16) merely defines polar coordinates in the region D_{xy}. So, if D'_{xy} is the image of D_{xy} in the polar plane spanned by pairs (r, θ), then, by converting the double integral to polar coordinate,s one infers that

$$\iiint_E f(x,y,z)\,dV = \iint_{D'_{xy}} \int_{z_{\rm bot}(r,\theta)}^{z_{\rm top}(r,\theta)} f(r\cos\theta, r\sin\theta, z) r\,dz\,dA'$$

$$= \iiint_{E'} f(r\cos\theta, r\sin\theta, z) r\,dV', \tag{14.17}$$

where the region E' is the image of E under the transformation from rectangular to cylindrical coordinates,

$$E' = \{(r,\theta,z)\,|\,z_{\rm bot}(r,\theta) \le z \le z_{\rm top}(r,\theta)\,,\ (r,\theta) \in D'_{xy}\},$$

and $z = z_{\rm bot}(r,\theta)$, $z = z_{\rm top}(r,\theta)$ are equations of the bottom and top boundaries of E written in polar coordinates by substituting (14.16) into the equations for boundaries written in rectangular coordinates. Note that $dV' = dz\,dr\,d\theta = dz\,dA'$ is the volume of an infinitesimal rectangle in the space spanned by the triples (r,θ,z). Its image in the space spanned by (x,y,z) lies between two cylinders whose radii differ by dr, between two half-planes with the angle $d\theta$ between them, and between two horizontal planes separated by the distance dz. So its volume is the product of the area dA of the base and the height dz, $dV = dz\,dA = r\,dz\,dA'$ according to the area transformation law for polar coordinates, $dA = rdA'$. So the volume transformation law for cylindrical coordinates reads

$$dV = J\,dV'\,, \quad J = r,$$

where $J = r$ is the Jacobian of transformation to cylindrical coordinates.

Cylindrical coordinates are advantageous when the boundaries of E contain cylinders, half-planes, horizontal planes, or any surfaces with *axial symmetry.* A set in space is said to be *axially symmetric* if there is an axis such that any rotation about it maps the set onto itself. For example, circular cones, circular paraboloids, and spheres are axially symmetric. Note also that the axis of cylindrical coordinates may be chosen to be the x or y axis, which would correspond to polar coordinates in the yz or xz plane.

EXAMPLE 14.22. *Evaluate the triple integral of $f(x,y,z) = x^2z$ over the region E bounded by the cylinder $x^2+y^2 = 1$, the paraboloid $z = x^2+y^2$, and the plane $z = 0$.*

SOLUTION:The solid E is axially symmetric because it is bounded from the bottom by the plane $z = 0$, by the circular paraboloid from the top, and the side boundary is the cylinder. Hence, D_{xy} is a disk of unit radius, and D'_{xy} is a rectangle, $(r, \theta) \in [0, 1] \times [0, 2\pi]$. The top and bottom boundaries are $z = z_{\rm top}(r, \theta) = r^2$ and $z = z_{\rm bot}(r, \theta) = 0$. Hence,

$$\begin{aligned} \iiint_E x^2 z\, dV &= \int_0^{2\pi}\int_0^1\int_0^{r^2} r^2\cos^2\theta\, z\, r d\, z d\, r d\theta \\ &= \frac{1}{2}\int_0^{2\pi}\cos^2\theta\, d\theta \int_0^1 r^7 dr = \frac{\pi}{16}, \end{aligned}$$

where the double-angle formula, $\cos^2\theta = (1+\cos(2\theta))/2$, has been used to evaluate the integral. □

104.2. Spherical Coordinates. Spherical coordinates are introduced by the following geometrical procedure. Let (x, y, z) be a point in space. Consider a ray from the origin through this point. Any such ray lies in the half-plane corresponding to a fixed value of the polar angle θ. Therefore, the ray is uniquely determined by the polar angle θ and the angle ϕ between the ray and the positive z axis. If ρ is the distance from the origin to the point (x, y, z), then the ordered triple of numbers (ρ, ϕ, θ) defines uniquely any point in space. The triples (ρ, ϕ, θ) are called *spherical coordinates* in space.

To find the transformation law from spherical to rectangular coordinates, consider the plane that contains the z axis and the ray from the origin through (x, y, z) and the rectangle with vertices $(0, 0, 0)$, $(0, 0, z)$, $(x, y, 0)$, and (x, y, z) in this plane. The diagonal of this rectangle has length ρ (the distance between $(0, 0, 0)$ and (x, y, z)). Therefore, its vertical side has length $z = \rho\cos\phi$ because the angle between this side and the diagonal is ϕ. Its horizontal side has length $\rho\sin\phi$. On the other hand, it is also the distance between $(0, 0, 0)$ and $(x, y, 0)$, that is, $r = \rho\sin\phi$, where $r = \sqrt{x^2 + y^2}$. Since $x = r\cos\theta$ and $y = r\sin\theta$, it is concluded that

$$x = \rho\sin\phi\cos\theta\,, \quad y = \rho\sin\phi\sin\theta\,, \quad z = \rho\cos\phi\,. \tag{14.18}$$

The inverse transformation follows from the geometrical interpretation of the spherical coordinates:

$$\rho = \sqrt{x^2 + y^2 + z^2}\,, \quad \cot\phi = \frac{z}{r} = \frac{z}{\sqrt{x^2 + y^2}}\,, \quad \tan\theta = \frac{y}{x}. \tag{14.19}$$

If (x, y, z) span the entire space, the maximal range of the variable ρ is the half-axis $\rho \in [0, \infty)$. The variable θ ranges over the interval $[0, 2\pi)$ as it coincides with the polar angle. To determine the range of the *azimuthal* angle ϕ, note that an angle between the positive z axis and any ray from the origin must be in the interval $[0, \pi]$. If $\phi = 0$, the ray coincides with the positive z axis. If $\phi = \pi$, the ray is the negative z axis. Any ray with $\phi = \pi/2$ lies in the xy plane.

104.2.1. Coordinate Surfaces of Spherical Coordinates. All points that have the same value of $\rho = \rho_0$ form a sphere of radius ρ_0 centered at the origin because they are at the same distance ρ_0 from the origin. Naturally, the coordinate surfaces of θ are the half-planes described earlier when discussing cylindrical coordinates. Consider a ray from the origin that has the angle $\phi = \phi_0$ with the positive z axis. By rotating this ray about the z axis, all rays with the fixed value of ϕ are obtained. Therefore, the coordinate surface $\phi = \phi_0$ is a circular cone whose axis is the z axis. For small values of ϕ, the cone is a narrow cone about the positive z axis. The cone becomes wider as ϕ increases so that it coincides with the xy plane when $\phi = \pi/2$. For $\phi > \pi/2$, the cone lies below the xy plane, and it eventually collapses into the negative z axis as soon as ϕ reaches the value π. The algebraic equations of the coordinate surfaces follow from (14.19):

$$\begin{aligned}
\rho = \rho_0 &\leftrightarrow x^2 + y^2 + z^2 = \rho_0^2 && \text{(sphere)},\\
\phi = \phi_0 &\leftrightarrow z = \cot(\phi_0)\sqrt{x^2 + y^2} && \text{(cone)},\\
\theta = \theta_0 &\leftrightarrow y\cos\theta_0 = x\sin\theta_0 && \text{(half-plane)}.
\end{aligned}$$

So any point in space can be viewed as the point of intersection of three coordinate surfaces: the sphere, cone, and half-plane. Under the transformation (14.19), any region E is mapped onto a region E' in the space spanned by the ordered triples (ρ, ϕ, θ). If E is bounded by spheres, cones, and half-planes only, then its image E' is a rectangle. Thus, a change of variables in a triple integral to spherical coordinates is advantageous when E is bounded by spheres, cones, and half-planes.

Example 14.23. *Let E be the portion of the solid bounded by the sphere $x^2 + y^2 + z^2 = 4$ and the cone $z^2 = 3(x^2 + y^2)$ that lies in the first octant. Find its image E' under transformation to spherical coordinates.*

Solution: The region E has four boundaries: the sphere, the cone $z = \sqrt{3}\sqrt{x^2 + y^2}$, the xz plane ($x \geq 0$), and the yz plane ($y \geq 0$). These boundaries are mapped onto, respectively, $\rho = 2$, $\cot\phi = \sqrt{3}$ or $\phi = \pi/3$, $\theta = 0$, and $\theta = \pi/2$. So E' is the rectangle $[0, 2] \times [0, \pi/3] \times [0, \pi/2]$. The region E is intersected by all spheres with radii $0 \leq \rho \leq 2$, all cones with angles $0 \leq \phi \leq \pi/3$, and all half-planes with angles $0 \leq \theta \leq \pi/2$. □

104.2.2. Volume Transformation Law. Let E' be the image of a region E under the transformation to spherical coordinates (14.19). Consider a rectangular partition of E' by equispaced planes $\rho = \rho_i$, $\phi = \phi_j$, and $\theta = \theta_k$ such that $\rho_{i+1} - \rho_i = \Delta\rho$, $\phi_{j+1} - \phi_j = \Delta\phi$, and $\theta_{k+1} - \theta_k = \Delta\theta$, where $\Delta\rho$, $\Delta\phi$, and $\Delta\theta$ are small numbers that can be regarded as differentials (or infinitesimal variations) of the spherical coordinates. Each partition rectangle has volume $\Delta V' = \Delta\rho\,\Delta\phi\,\Delta\theta$. The rectangular partition of E' induces a partition of E by spheres, cones, and half-planes. Each partition element is bounded by two spheres whose radii differ by $\Delta\rho$, by two cones whose angles differ by $\Delta\phi$, and by two half-planes the angle between which is $\Delta\theta$. The volume of any such partition element can be written as

$$\Delta V = J\,\Delta V'$$

because only terms linear in the variations $\Delta\rho = d\rho$, $\Delta\phi = d\phi$, and $\Delta\theta = d\theta$ have to be retained. The value of J depends on a partition element (e.g., partition elements closer to the origin should have smaller volumes by the geometry of the partition). The function J is the *Jacobian* for spherical coordinates.

By means of (14.18), an integrable function $f(x, y, z)$ can be written in spherical coordinates. According to (14.13), in the three-variable limit $(\Delta\rho, \Delta\phi, \Delta\theta) \to (0, 0, 0)$, the Riemann sum for f for the partition constructed converges to a triple integral of fJ expressed in the variables (ρ, ϕ, θ) over the region E' and thereby defines the triple integral of f over E in spherical coordinates.

To find J, consider the image of the rectangle $\rho \in [\rho_0, \rho_0 + \Delta\rho]$, $\phi \in [\phi_0, \phi_0 + \Delta\phi]$, $\theta \in [\theta_0, \theta_0 + \Delta\theta]$ under the transformation (14.18). Since it lies between two spheres of radii ρ_0 and $\rho_0 + \Delta\rho$, its volume can be written as $\Delta V = \Delta\rho\,\Delta A$, where ΔA is the area of the portion of the sphere of radius ρ_0 that lies between two cones and two half-planes. Any half-plane $\theta = \theta_0$ intersects the sphere $\rho = \rho_0$ along a half-circle of radius ρ_0. The arc length of the portion of this circle that lies between the two cones $\phi = \phi_0$ and $\phi = \phi_0 + \Delta\phi$ is therefore $\Delta a = \rho_0\,\Delta\phi$. The cone $\phi = \phi_0$ intersects the sphere $\rho = \rho_0$ along a circle of radius $r_0 = \rho_0 \sin\phi_0$ (see the text above (14.18)). Hence, the arc length of the portion of this circle of intersection that lies between the half-planes $\theta = \theta_0$ and $\theta = \theta_0 + \Delta\theta$ is $\Delta b = r_0\,\Delta\theta = \rho_0 \sin\phi\,\Delta\theta$. The area ΔA can be approximated by the area of a rectangle with adjacent sides Δa and Δb. Since only terms linear in $\Delta\phi$ and $\Delta\theta$ are to be retained, one can write $\Delta A = \Delta a\,\Delta b = \rho_0^2 \sin\phi_0\,\Delta\phi\,\Delta\theta$. Thus, the volume transformation law reads

$$dV = J\,dV'\,, \quad J = \rho^2 \sin\phi\,.$$

By the continuity of the Jacobian, the difference of the values of J at any two sample points in a partition rectangle in E' vanishes in the limit $(\Delta\rho, \Delta\phi, \Delta\theta) \to (0, 0, 0)$; that is, the value of the Jacobian in $\Delta V = J\,\Delta V'$ can be taken at any point within the partition element when evaluating the limit. Therefore, for any choice of sample points, the limit of the Riemann sum (14.13) for the constructed partition is

$$\iiint_E f(x, y, z)\, dV = \iiint_{E'} f\Big(\rho\sin\phi\cos\theta, \rho\sin\phi\sin\theta, \rho\cos\phi\Big)\rho^2\sin\phi\, dV'.$$

This relation defines the triple integral of f over E in spherical coordinates. The triple integral over E' has to be evaluated by converting it to a suitable iterated integral.

Example 14.24. *Find the volume of the solid E bounded by the sphere $x^2 + y^2 + z^2 = 2z$ and the cone $z = \sqrt{x^2 + y^2}$.*

Solution: By completing the squares, the equation $x^2 + y^2 + z^2 = 2z$ is written in the standard form $x^2 + y^2 + (z-1)^2 = 1$, which describes a sphere of unit radius centered at $(0, 0, 1)$. So E is bounded from the top by this sphere, while the bottom boundary of E is the cone, and E has no other boundaries. In spherical coordinates, the top boundary becomes $\rho^2 = 2\rho\cos\phi$ or $\rho = 2\cos\phi$. The bottom boundary is $\phi = \pi/4$. The boundaries of E impose no restriction on θ, which can therefore be taken over its full range. Hence, the image E' admits the following algebraic description:

$$E' = \Big\{(\rho, \phi, \theta)\,|\, 0 \le \rho \le 2\cos\phi\,,\ (\phi, \theta) \in [0, \pi/4] \times [0, 2\pi]\Big\}.$$

Since the range of ρ depends on the other variables, the integration with respect to it must be carried out first when converting the triple integral over E' into an iterated integral (E' is ρ simple, and the projection of E' onto the $\phi\theta$ plane is the rectangle $[0, \pi/4] \times [0, 2\pi]$). The order in which the integration with respect to θ and ϕ is carried out is irrelevant because the angular variables range over a rectangle. One has

$$\begin{aligned} V(E) &= \iiint_E dV = \iiint_{E'} \rho^2\sin\phi\, dV' \\ &= \int_0^{2\pi}\int_0^{\pi/4} \sin\phi \int_0^{2\cos\phi} \rho^2\, d\rho\, d\phi\, d\theta \\ &= \frac{8}{3}\int_0^{2\pi} d\theta \int_0^{\pi/4} \cos^3\phi\sin\phi\, d\phi = \frac{16\pi}{3}\int_{1/\sqrt{2}}^1 u^3\, du = \pi, \end{aligned}$$

where the change of variables $u = \cos\phi$ has been carried out in the last integral. □

105. Change of Variables in Triple Integrals

Consider the transformation of an open region E' in space into an open region E defined by $x = x(u, v, w)$, $y = y(u, v, z)$, and $z = z(u, v, w)$; that is, for every point $(u, v, w) \in E'$, these functions define an image point $(x, y, z) \in E$. If no two points in E' have the same image point, the transformation is *one-to-one*, and there is a *one-to-one correspondence* between points of E and E'. The inverse transformation exists and is defined by the functions $u = u(x, y, z)$, $v = v(x, y, z)$, and $w = w(x, y, z)$. A point $(x_0, y_0, z_0) = \mathbf{r}_0$ is the intersection point of three coordinate planes $x = x_0$, $y = y_0$, and $z = z_0$. Alternatively, it can also be viewed as the point of intersection of three *coordinate surfaces*, $u(x, y, z) = u_0$, $v(x, y, z) = v_0$, and $w(x, y, z) = w_0$, where the image of (u_0, v_0, w_0) under the one-to-one transformation is $\mathbf{r}_0$.

Definition 14.11. (Jacobian of a Mapping).
Suppose that a one-to-one mapping of an open set E' onto E has continuous first-order partial derivatives. The quantity

$$\frac{\partial(x, y, z)}{\partial(u, v, w)} = \det \begin{pmatrix} x'_u & y'_u & z'_u \\ x'_v & y'_v & z'_v \\ x'_w & y'_w & z'_w \end{pmatrix}$$

is called the Jacobian of the mapping.

Definition 14.12. (Change of Variables).
A continuously differentiable one-to-one mapping of an open set E' onto E is called a change of variables *(or a change of coordinates) if the Jacobian of the mapping does not vanish in E'.*

As in the case of double integrals, a change of variables in space can be used to simplify the evaluation of triple integrals. For example, if there is a change of variables whose coordinate surfaces form a boundary of the integration region E, then the new integration region E' is a rectangle, and the limits in the corresponding iterated integral are greatly simplified in accordance with Fubini's theorem.

105.1. The Volume Transformation Law. It is convenient to introduce the following notations: $(u, v, w) = \mathbf{r}'$ and $(x, y, z) = \mathbf{r}$; that is, under the change of variables,

$$\mathbf{r} = \Big(x(\mathbf{r}'),\ y(\mathbf{r}'),\ z(\mathbf{r}')\Big) \quad \text{or} \quad \mathbf{r}' = \Big(u(\mathbf{r}),\ v(\mathbf{r}),\ w(\mathbf{r})\Big). \tag{14.20}$$

Let E'_0 be an infinitesimal rectangle in E', $u \in [u_0, u_0 + \Delta u]$, $v \in [v_0, v_0 + \Delta v]$, and $w \in [w_0, w_0 + \Delta w]$, where Δu, Δv, and Δw are infinitesimal variations that can be viewed as differentials of the new variables. This means that all algebraic expressions involving Δu, Δv, and Δw are to be linearized with respect to them, and their higher powers neglected. If E_0 is the image of E'_0, the volumes of E_0 and E'_0 are proportional:

$$\Delta V = J\,\Delta V'\,, \qquad \Delta V' = \Delta u\,\Delta v\,\Delta w\,.$$

The objective is to calculate J. By the examples of cylindrical and spherical coordinates, J is a function of the point (u_0, v_0, w_0) at which the rectangle E'_0 is taken. The derivation of J is fully analogous to the two-variable case.

Let O', A', B', and C' have the coordinates, respectively,

$$\begin{aligned}
\mathbf{r}'_0 &= (u_0, v_0, w_0)\,,\\
\mathbf{r}'_a &= (u_0 + \Delta u, v_0, w_0) = \mathbf{r}'_0 + \hat{\mathbf{e}}_1\,\Delta u,\\
\mathbf{r}'_b &= (u_0, v_0 + \Delta v, w_0) = \mathbf{r}'_0 + \hat{\mathbf{e}}_2\,\Delta v,\\
\mathbf{r}'_c &= (u_0, v_0, w_0 + \Delta w) = \mathbf{r}'_0 + \hat{\mathbf{e}}_3\,\Delta w,
\end{aligned}$$

where $\hat{\mathbf{e}}_{1,2,3}$ are unit vectors along the first, second, and third coordinate axes. In other words, the segments $O'A'$, $O'B'$, and $O'C'$ are the adjacent sides of the rectangle E'_0. Let O, A, B, and C be the images of O', A', B', and C' in the region E. The volume ΔV of E_0 can be approximated by the volume of the parallelepiped with adjacent sides $\mathbf{a} = \vec{OA}$, $\mathbf{b} = \vec{OB}$, and $\mathbf{c} = \vec{OC}$. Then

$$\begin{aligned}
\mathbf{a} &= \Big(x(\mathbf{r}'_a) - x(\mathbf{r}'_0),\ y(\mathbf{r}'_a) - y(\mathbf{r}'_0),\ z(\mathbf{r}'_a) - z(\mathbf{r}'_0)\Big) = (x'_u,\ y'_u,\ z'_u)\,\Delta u,\\
\mathbf{b} &= \Big(x(\mathbf{r}'_b) - x(\mathbf{r}'_0),\ y(\mathbf{r}'_b) - y(\mathbf{r}'_0),\ z(\mathbf{r}'_b) - z(\mathbf{r}'_0)\Big) = (x'_v,\ y'_v,\ z'_v)\,\Delta v,\\
\mathbf{c} &= \Big(x(\mathbf{r}'_c) - x(\mathbf{r}'_0),\ y(\mathbf{r}'_c) - y(\mathbf{r}'_0),\ z(\mathbf{r}'_c) - z(\mathbf{r}'_0)\Big) = (x'_w,\ y'_w,\ z'_w)\,\Delta w,
\end{aligned}$$

where all the differences have been linearized, for instance, $x(\mathbf{r}'_a) - x(\mathbf{r}'_0) = x(\mathbf{r}'_0 + \hat{\mathbf{e}}_1\,\Delta u) - x(\mathbf{r}'_0) = x'_u(\mathbf{r}'_0)\,\Delta u$, by the definition of the partial derivative of $x(u, v, w)$ with respect to the first variable u. The volume of the parallelepiped is given by the absolute value of the triple product:

(14.21)

$$\Delta V = |\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})| = \left|\det\begin{pmatrix} x'_u & y'_u & z'_u\\ x'_v & y'_v & z'_v\\ x'_w & y'_w & z'_w\end{pmatrix}\right|\Delta u\,\Delta v\,\Delta w = J\,\Delta V',$$

where the derivatives are evaluated at (u_0, v_0, w_0). The function J in (14.21) is the *absolute value* of the Jacobian. The first-order partial derivatives are continuous for a change of variables and so are the

Jacobian and its absolute value. Similarly to the two-dimensional case, it can also be proved that

$$J = \left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| = \frac{1}{\left|\frac{\partial(u,v,w)}{\partial(x,y,z)}\right|} = \left|\det\begin{pmatrix} u'_x & u'_y & u'_z \\ v'_x & v'_y & v'_z \\ w'_x & w'_y & w'_z \end{pmatrix}\right|^{-1} . \tag{14.22}$$

This expression defines J as a function of the old variables (x, y, z).

105.2. Triple Integral in Curvilinear Coordinates. Consider a partition of E' by equispaced planes $u = u_i$, $v = v_j$, and $w = w_k$ $u_{i+1} - u_i = \Delta u$, $v_{j+1} - v_j = \Delta v$, and $w_{k+1} - w_k = \Delta w$. The indices (i, j, k) enumerate planes that intersect E'. This rectangular partition of E' corresponds to a partition of E by the coordinate surfaces $u(\mathbf{r}) = u_i$, $v(\mathbf{r}) = v_j$, and $w(\mathbf{r}) = w_k$. If E'_{ijk} is the rectangle $u \in [u_i, u_{i+1}]$, $v_j \in [v_j, v_{j+1}]$, and $w \in [w_k, w_{k+1}]$, then its image, being the corresponding partition element of E, is denoted by E_{ijk}. A Riemann sum can be constructed for this partition of E (assuming as before that f is defined by zeros outside E). The triple integral of f over E is the limit (14.13) which is understood as the three-variable limit $(\Delta u, \Delta v, \Delta w) \to (0, 0, 0)$. The volume ΔV_{ijk} of E_{ijk} is related to the volume of the rectangle E'_{ijk} by (14.21). By continuity of J, its value in (14.21) can be taken at any sample point in E'_{ijk}. According to the definition of the triple integral, the limit of the Riemann sum is the triple integral of fJ over the region E'.

Theorem 14.11. (Change of Variables in a Triple Integral). *Let a continuously differentiable mapping $E' \to E$ have a non-vanishing Jacobian, except perhaps on the boundary of E'. Suppose that f is continuous on E and E is bounded by piecewise-smooth surfaces. Then*

$$\iiint_E f(\mathbf{r})\,dV = \iiint_{E'} f(x(\mathbf{r}'), y(\mathbf{r}'), z(\mathbf{r}'))J(\mathbf{r}')\,dV' ,$$
$$J(\mathbf{r}') = \left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| .$$

Evaluation of a triple integral in curvilinear coordinates follows the same steps as for a double integral in curvilinear coordinates.

Example 14.25. (Volume of an Ellipsoid). *Find the volume of a solid region E bounded by an ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$.*

Solution: The integration domain can be simplified by a scaling transformation $x = au$, $y = bv$, and $z = cw$ under which the ellipsoid is mapped onto a sphere of unit radius $u^2 + v^2 + w^2 = 1$. The image E' of E is a ball of unit radius. The Jacobian of this transformation is

$$J = \left|\det \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}\right| = abc.$$

Therefore,

$$\begin{aligned} V(E) = \iiint_E dV &= \iiint_{E'} J\, dV' = abc \iiint_{E'} dV' \\ &= abcV(E') = \frac{4\pi}{3}\, abc. \end{aligned}$$

□

When $a = b = c = R$, the ellipsoid becomes a ball of radius R, and a familiar expression for the volume is recovered: $V = (4\pi/3)R^3$.

105.3. Study Problems.

Problem 14.4. (Volume of a Tetrahedron).
A tetrahedron is a solid with four vertices and four triangular faces. Let the vectors **a***,* **b***, and* **c** *be three adjacent sides of the tetrahedron. Find its volume.*

Solution: Consider first a tetrahedron whose adjacent sides are along the coordinate axes and have the same length q. From the geometry, it is clear that six such tetrahedrons form a cube of volume q^3. Therefore, the volume of each tetrahedron is $q^3/6$ (if so desired this can also be established by evaluating the corresponding triple integral; this is left to the reader). The idea is to make a change of variables such that a generic tetrahedron is mapped onto a tetrahedron whose adjacent faces lie in the three coordinate planes. The adjacent faces are portions of the planes through the origin. The face containing vectors **a** and **b** is perpendicular to vector $\mathbf{n} = \mathbf{a} \times \mathbf{b}$ so the equation of this boundary is $\mathbf{n} \cdot \mathbf{r} = 0$. The other adjacent faces are similar:

$$\begin{aligned} \mathbf{n} \cdot \mathbf{r} = 0 \quad &\text{or} \quad n_1x + n_2y + n_3z = 0\,, & \mathbf{n} = \mathbf{a} \times \mathbf{b}, \\ \mathbf{l} \cdot \mathbf{r} = 0 \quad &\text{or} \quad l_1x + l_2y + l_3z = 0\,, & \mathbf{l} = \mathbf{c} \times \mathbf{a}, \\ \mathbf{m} \cdot \mathbf{r} = 0 \quad &\text{or} \quad m_1x + m_2y + m_3z = 0\,, & \mathbf{m} = \mathbf{b} \times \mathbf{c}, \end{aligned}$$

where $\mathbf{r} = (x, y, z)$. So, by putting $u = \mathbf{m} \cdot \mathbf{r}$, $v = \mathbf{l} \cdot \mathbf{r}$, and $w = \mathbf{n} \cdot \mathbf{r}$, the images of these planes become the coordinate planes, $w = 0$, $v = 0$, and $u = 0$. A linear equation in the old variables becomes a linear equation in the new variables under a linear transformation. Therefore, an image of a plane is a plane. So the fourth boundary of E' is a plane through the points $\mathbf{a}'$, $\mathbf{b}'$, and $\mathbf{c}'$, which are the images of $\mathbf{r} = \mathbf{a}$, $\mathbf{r} = \mathbf{b}$, and $\mathbf{r} = \mathbf{c}$, respectively. One has $\mathbf{a}' = (u(\mathbf{a}), v(\mathbf{a}), w(\mathbf{a})) = (q, 0, 0)$, where $q = \mathbf{a} \cdot \mathbf{m} = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$ because $\mathbf{a} \cdot \mathbf{n} = 0$ and $\mathbf{a} \cdot \mathbf{l} = 0$ by the geometrical properties of the cross product. Similarly, $\mathbf{b}' = (0, q, 0)$ and $\mathbf{c}' = (0, 0, q)$. Thus, the volume of the image region E' is $V(E') = |q|^3/6$ (the absolute value is needed because the triple product can be negative). To find the volume $V(E)$, the Jacobian of the transformation has to be found. It is convenient to use the representation (14.22):

$$J = \left| \det \begin{pmatrix} m_1 & m_2 & m_3 \\ l_1 & l_2 & l_3 \\ n_1 & n_2 & n_3 \end{pmatrix} \right|^{-1} = \frac{1}{|\mathbf{m} \cdot (\mathbf{n} \times \mathbf{l})|}.$$

Therefore,

$$V(E) = \iiint_E dV = \iiint_{E'} J\, dV' = J \iiint_{E'} dV' = JV(E') = \frac{|q|^3 J}{6}.$$

The volume $V(E)$ is independent of the orientation of the coordinate axes. It is convenient to direct the x axis along the vector $\mathbf{a}$. The y axis is directed so that $\mathbf{b}$ is in the xy plane. With this choice, $\mathbf{a} = (a_1, 0, 0)$, $\mathbf{b} = (b_1, b_2, 0)$, and $\mathbf{c} = (c_1, c_2, c_3)$. A straight forward calculation shows that $q = a_1 b_2 c_3$ and $J = (a_1^2 b_2^2 c_3^2)^{-1}$. Hence, $V(E) = |a_1 b_2 c_3|/6$. Finally, note that $|c_3| = h$ is the height of the tetrahedron, that is, the distance from a vertex $\mathbf{c}$ to the opposite face (to the xy plane). The area of that face is $A = \|\mathbf{a} \times \mathbf{b}\|/2 = |a_1 b_2|/2$. Thus,

$$V(E) = \frac{1}{3} hA;$$

that is, the volume of a tetrahedron is one-third the distance from a vertex to the opposite face, times the area of that face. □

106. Improper Multiple Integrals

In the case of one-variable integration, improper integrals occur when the integrand is not defined at a boundary point of the integration interval or the integration interval is not bounded. For example,

$$(14.23) \qquad \int_0^1 \frac{dx}{x^\nu} = \lim_{a \to 0} \int_a^1 \frac{dx}{x^\nu} = \lim_{a \to 0} \frac{1 - a^{1-\nu}}{1 - \nu} = \frac{1}{1 - \nu}, \qquad \nu < 1,$$

or

$$\int_0^\infty \frac{1}{1 + x^2}\, dx = \lim_{a \to \infty} \int_0^a \frac{1}{1 + x^2}\, dx = \lim_{a \to \infty} \tan^{-1} a = \frac{\pi}{2}.$$

Improper multiple integrals are quite common in many practical applications.

106.1. Multiple Integrals of Unbounded Functions. Suppose a function $f(\mathbf{r})$ is not defined at a point $\mathbf{r}_0$ that is a limit point of the domain of f (any neighborhood of $\mathbf{r}_0$ contains points of the domain of f). Here $\mathbf{r} = (x, y, z) \in E$ or $\mathbf{r} = (x, y) \in D$. If in any small ball (or disk) B_ε of radius ε centered at $\mathbf{r}_0$ the values of $|f(\mathbf{r})|$ are not bounded, then the function f is said to be *singular* at $\mathbf{r}_0$. In this case, the upper and lower sums cannot be defined because for some partition rectangles $\sup f$ or $\inf f$ or both do not exist, and neither is defined a multiple integral of f. If a region E (or D) contains such a point, define the region E_ε (or D_ε) by removing all points of E (or D) that also lie in the ball (disk) B_ε. Suppose that f is integrable on E_ε (or D_ε) for any $\varepsilon > 0$ (e.g., it is continuous). Then, by analogy with the one-variable case, a multiple integral of f over E (or D) is *defined* as the limit
(14.24)

$$\iiint_E f\,dV = \lim_{\varepsilon\to 0}\iiint_{E_\varepsilon} f\,dV \quad \text{or} \quad \iint_D f\,dA = \lim_{\varepsilon\to 0}\iint_{D_\varepsilon} f\,dA,$$

provided, of course, the limit exists. If f is singular in a point set S, then one can construct a set S_ε that is the union of balls of radius ε centered at each point of S. Then D_ϵ (or E_ε) is obtained by removing S_ε from D (or E).

Although this definition seems a rather natural generalization of the one-variable case, there are subtleties that are specific to multivariable integrals. This is illustrated by the following example. Suppose that

$$f(x, y) = \frac{y^2 - x^2}{(x^2 + y^2)^2} \tag{14.25}$$

is to be integrated over the sector $0 \le \theta \le \theta_0$ of a disk $x^2 + y^2 \le 1$, where θ is the polar angle. If the definition (14.24) is applied, then D_ε is the portion of the ring $\varepsilon^2 \le x^2 + y^2 \le 1$ corresponding to $0 \le \theta \le \theta_0$. Then, by evaluating the integral in polar coordinates, one finds that

$$\iint_{D_\varepsilon} \frac{y^2 - x^2}{(x^2 + y^2)^2} dA = -\int_0^{\theta_0} \cos(2\theta)\,d\theta \int_\varepsilon^1 \frac{dr}{r} = \frac{1}{2}\sin(2\theta_0)\ln\varepsilon\,.$$

The limit $\varepsilon \to 0$ does not exist for all θ_0 such that $\sin(2\theta_0) \ne 0$, whereas the integral vanishes if $\theta_0 = k\pi/2$, $k = 1, 2, 3, 4$, for any $\varepsilon > 0$. Let $\theta_0 = \pi/2$. The integral vanishes because of symmetry, $(x, y) \to (y, x)$, $f(y, x) = -f(x, y)$, while the integration region is invariant under this transformation. The integrand is positive in the part of the domain where $x^2 < y^2$ and negative if $y^2 > x^2$, and there is a mutual cancellation of contributions from these regions. If the improper integral of the absolute value $|f(x,y)|$ is considered, then no such cancellation can occur, and the improper integral always diverges.

Definition 14.13. (Conditional Convergence). *The improper integral is said to converge conditionally if the limit (14.24) exists.*

Definition 14.14. (Absolute Convergence and Integrability). *The improper integral of a function f is said to converge absolutely if the improper integral of the absolute value $|f|$ converges, and in this case, the function f is said to be absolutely integrable.*

Consider a Riemann sum for an improper integral over a bounded region where no sample points coincide with $\mathbf{r}_0$. The absolute integrability of f guarantees that all Riemann sums remain bounded, and hence they cannot diverge. Indeed,

$$|R(f,N)| = \left|\sum_p f(\mathbf{r}_p)\,\Delta V_p\right| \le \sum_p |f(\mathbf{r}_p)|\,\Delta V_p < \infty,$$

and the sum on the right side converges in the limit $N \to \infty$ by the convergence of the integral of $|f|$. This conclusion holds independently of the choice of a partition of the integration region and the choice of sample points (with the aforementioned restriction).

If a function is not absolutely integrable, its iterated integrals may still be well defined. However, the value of the iterated integral depends on the order of integration (e.g., Fubini's theorem may not hold). For example, consider the function (14.25) over the rectangle $D = [0,1] \times [0,1]$. It is not absolutely integrable on D as the improper integral of $|f|$ diverges as argued. On the other hand, consider a rectangular partition of D where each partition rectangle has the area $\Delta x\,\Delta y$ and the sample point in the partition rectangle $[0,\Delta x] \times [0,\Delta y]$ does not coincide with the origin (where f is not defined). The limit of the Riemann sum is the *two-variable* limit $(\Delta x, \Delta y) \to (0,0)$. By taking first $\Delta y \to 0$ and then $\Delta x \to 0$, one obtains an iterated integral in which the integration with respect to y is carried out first:

$$\begin{aligned}
\lim_{a\to 0}\int_a^1 \lim_{b\to 0}\int_b^1 \frac{x^2-y^2}{(x^2+y^2)^2}dy\,dx &= \lim_{a\to 0}\int_a^1 \lim_{b\to 0}\int_b^1 \frac{\partial}{\partial y}\frac{y}{x^2+y^2}dy\,dx \\
&= \lim_{a\to 0}\int_a^1 \lim_{b\to 0}\Big(\frac{1}{1+x^2} - \frac{b}{x^2+b^2}\Big)dx \\
&= \lim_{a\to 0}\int_a^1 \frac{dx}{1+x^2} = \int_0^1 \frac{dx}{1+x^2} = \frac{\pi}{4}.
\end{aligned}$$

Here $(a, b) = (\Delta x, \Delta y)$. Alternatively, the limit $\Delta x \to 0$ can be taken first and then $\Delta y \to 0$, which results in the iterated integral in the reverse order:

$$\begin{aligned}\lim_{b\to 0}\int_b^1 \lim_{a\to 0}\int_a^1 \frac{x^2-y^2}{(x^2+y^2)^2}dx\,dy &= -\lim_{b\to 0}\int_b^1 \lim_{a\to 0}\int_a^1 \frac{\partial}{\partial x}\frac{x}{x^2+y^2}dy\,dx \\ &= -\lim_{b\to 0}\int_b^1 \lim_{a\to 0}\Big(\frac{1}{1+y^2}-\frac{a}{y^2+a^2}\Big)dy \\ &= -\lim_{b\to 0}\int_b^1 \frac{dy}{1+y^2} \\ &= -\int_0^1 \frac{dy}{1+y^2} = -\frac{\pi}{4}.\end{aligned}$$

This shows that the limit of the Riemann sum as a function of *two variables* Δx and Δy does not exist because it depends on a path along which the limit point is approached.

Thus, when dealing with an improper integral, the absolute convergence (absolute integrability) must be established first. Then the improper integral can be evaluated by means of the limit procedure (14.24). An analogy can be made with the conditionally and absolutely convergent series studied in Calculus II. If a series converges absolutely, then the sum does not depend on the order of summation or rearrangement of the series terms. If a series converges conditionally, but not absolutely, then, by rearranging the terms, the sum can take any value or even diverge. Riemann sums of conditionally convergent integrals behave pretty much as conditionally convergent series. The following theorem is useful to assess the integrability.

THEOREM 14.12. (Absolute Integrability Test).
If $|f(\mathbf{r})| \leq g(\mathbf{r})$ for all $\mathbf{r}$ in D and $g(\mathbf{r})$ is integrable on D, then f is absolutely integrable on D.

EXAMPLE 14.26. *Evaluate the triple integral of $f(x, y, z) = (x^2 + y^2 + z^2)^{-1}$ over a ball of radius R centered at the origin if it exists.*

SOLUTION: The function is singular only at the origin. So the restricted region E_ε lies between two spheres: $\varepsilon^2 \leq x^2 + y^2 + z^2 \leq R^2$. Since $|f| = f > 0$ in D, the convergence of the integral over E_ε also implies the absolute integrability of f. By making use of the spherical coordinates, one obtains.

$$\iiint_{E_\varepsilon} \frac{dV}{x^2+y^2+z^2} = \int_0^{2\pi}\int_0^{\pi}\int_\varepsilon^R \frac{\rho^2\sin\theta}{\rho^2}d\rho\,d\phi\,d\theta = 4\pi(R-\varepsilon) \to 4\pi R$$

as $\varepsilon \to 0$. So the improper integral exists and equals $4\pi R$. □

Example 14.27. *Investigate the absolute integrability of $f(x,y) = x/(x^2+y^2)^{\nu/2}$, $\nu > 0$, on a bounded region D. Find the integral, if it exists, over D, the part of the disk of unit radius in the first quadrant.*

Solution: The function is singular at the origin. Since f is continuous everywhere except the origin, it is sufficient to investigate the integrability on a disk centered at the origin. Put $r = \sqrt{x^2+y^2}$ (the polar radial coordinate). Then $|x| \le r$ and hence $|f| \le r/r^\nu = r^{1-\nu} = g$. In the polar coordiantes, the improper integral (14.24) of g over a disk of unit radius is

$$\int_0^{2\pi} d\theta \int_\varepsilon^1 g(r) r\, dr = 2\pi \int_\varepsilon^1 r^{2-\nu}\, dr = 2\pi \begin{cases} -\ln\varepsilon, & \nu = 3 \\ 1 - \frac{\varepsilon^{3-\nu}}{3-\nu}, & \nu \ne 3 \end{cases} .$$

The limit $\varepsilon \to 0$ is finite if $\nu < 3$. By the integrability test (Theorem 14.12), the function f is absolutely integrable if $\nu < 3$. For $\nu < 3$ and D, the part of the unit disk in the first quadrant, one infers that

$$\lim_{\varepsilon\to 0} \iint_{D_\varepsilon} f\, dA = \lim_{\varepsilon\to 0} \int\limits_0^{\pi/2} \int\limits_\varepsilon^1 \frac{r\cos\theta}{r^\nu} r\, dr\, d\theta = \lim_{\varepsilon\to 0} \int\limits_\varepsilon^1 r^{2-\nu}\, dr = 1.$$

□

The two examples studied exhibit a common feature of how the function should change with the distance from the point of singularity in order to be integrable.

Theorem 14.13. *Let a function f be continuous on a bounded region D of a Euclidean space and let f be singular at a limit point $\mathbf{r}_0$ of D. Suppose that $|f(\mathbf{r})| \le M\|\mathbf{r}-\mathbf{r}_0\|^{-\nu}$ for all $\mathbf{r}$ in D such that $\|\mathbf{r}-\mathbf{r}_0\| < R$ for some $R > 0$ and $M > 0$. Then f is absolutely integrable on D if $\nu < n$, where n is the dimension of the space.*

Proof. One can always set the origin of the coordinate system at $\mathbf{r}_0$ by the shift transformation $\mathbf{r} \to \mathbf{r}-\mathbf{r}_0$. Evidently, its Jacobian is 1. So, without loss of generality, assume that f is singular at the origin. Let B_R be the ball $\|\mathbf{r}\| < R$ and let B_R^D be the intersection of B_R and D. For $n = 1$, the integrability follows from (14.23). In the two-variable case, the use of the polar coordinates yields $dA = r\, dr\, d\theta$, $\|\mathbf{r}\| = r$, and

$$\iint_{B_R^D} |f|\, dA \le M \iint_{B_R^D} \frac{dA}{\|\mathbf{r}\|^\nu} \le M \iint_{B_R} \frac{dA}{\|\mathbf{r}\|^\nu} = 2\pi M \int_0^R \frac{dr}{r^{\nu-1}},$$

which is finite if $\nu < 2$; the second inequality follows from that the part B_R^D is contained in B_R and the integrand is positive. In the three-variable case, the volume element in the spherical coordinates is $dV =$

$\rho^2 \sin\phi \, d\rho \, d\phi \, d\theta$ where $\|\mathbf{r}\| = \rho$. So a similar estimate of the improper triple integral of f over B_R^D yields an upper bound $4\pi M \int_0^R \rho^{2-\nu} \, d\rho$, which is finite if $\nu < 3$. □

106.2. Multiple Integrals Over Unbounded Regions. Suppose $f(\mathbf{r})$ is a continuous function on an unbounded planar (or spatial) region D (or E). Let D_R be the intersection of D with a disk of radius R centered at the origin and let E_R be the intersection of E with a ball of radius R centered at the origin. The integral of f over D (or E) is *defined* by

$$\iint_D f(\mathbf{r}) \, dA = \lim_{R\to\infty} \iint_{D_R} f(\mathbf{r}) \, dA \quad \text{or}$$
$$\iiint_E f(\mathbf{r}) \, dV = \lim_{R\to\infty} \iiint_{E_R} f(\mathbf{r}) \, dV.$$

The improper integral is said to converge absolutely if the limit of integrals of the absolute value $|f|$ exists as $R \to \infty$ and the function f is called absolutely integrable on D. The improper integral is called conditionally convergent if the limit of integrals of f exists, while f is not absolutely integrable. Since $f \leq |f|$, the absolute integrability implies the existence of the improper integral.

Example 14.28. *Evaluate the double integral of* $f(x, y) = \exp(-x^2 - y^2)$ *over the entire plane.*

Solution: The function is positive and $|f| = f$. So it is sufficient to investigate the convergence of the integral over D_R, which is the disk of radius R as $R \to 0$. By making use of the polar coordinates,

$$\iint_D e^{-x^2-y^2} dA = \lim_{R\to\infty} \int_0^{2\pi} \int_0^R e^{-r^2} r \, dr \, d\theta = \pi \lim_{R\to\infty} \int_0^{R^2} e^{-u} \, du$$
$$= \pi \lim_{R\to\infty} (1 - e^{-R^2}) = \pi,$$

where the substitution $u = r^2$ has been made. □

It is interesting to observe the following. Since the double integral exists, it can also be represented as an iterated integral in rectangular coordinates, which is the product of two improper ordinary integrals:

$$\iint_D e^{-x^2-y^2} dA = \int_{-\infty}^{\infty} e^{-x^2} dx \int_{-\infty}^{\infty} e^{-y^2} dy = I^2 \, ,$$
$$I = \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

because $I^2 = \pi$ by the value of the double integral. A direct evaluation of I by means of the fundamental theorem of calculus is problematic as an antiderivative of e^{-x^2} cannot be expressed in elementary function.

The integrability test (Theorem 14.12) holds for the case of unbounded regions. The asymptotic behavior of a function sufficient for integrability on an unbounded region is stated in the following theorem, which is an analog of Theorem 14.13.

THEOREM 14.14. *Suppose f is a continuous function on an unbounded region D of a Euclidean space such that $|f(\mathbf{r})| \le M\|\mathbf{r}\|^{-\nu}$ for all $\|\mathbf{r}\| \ge R$ in D and some $R > 0$ and $M \ge 0$. Then f is absolutely integrable on D if $\nu > n$, where n is the dimension of the space.*

PROOF. Let $R > 0$. Consider the following one-dimensional improper integral:

$$\int_R^\infty \frac{dx}{x^\nu} = \lim_{a\to\infty} \int_R^a \frac{dx}{x^\nu} = \lim_{a\to\infty} \frac{x^{1-\nu}}{1-\nu}\bigg|_R^a = -\frac{R^{1-\nu}}{1-\nu} + \lim_{a\to\infty} \frac{a^{1-\nu}}{1-\nu}$$

if $\nu \ne 1$. The limit is finite if $\nu > 1$. When $\nu = 1$, the integral diverges as $\ln a$. Let D'_R be the part of D that lies outside the ball B_R of radius R and let B'_B be the part of the space outside B_R. Note that B'_R includes D'_R. In the two-variable case, the use of the polar coordinates gives

$$\iint\limits_{D'_R} |f| dA \le \iint\limits_{B'_R} |f| dA \le \iint\limits_{B'_R} \frac{M dA}{\|\mathbf{r}\|^\nu} = M \int\limits_0^{2\pi} d\theta \int\limits_R^\infty \frac{r\,dr}{r^\nu}$$
$$= 2\pi M \int\limits_R^\infty \frac{dr}{r^{\nu-1}},$$

which is finite, provided $\nu - 1 > -1$ or $\nu > 2$. The case of triple integrals is proved similarly by means of the spherical coordinates. The volume element is $dV = \rho^2 \sin\phi\, d\rho\, d\phi\, d\theta$. The integration over the spherical angles yields the factor 4π as $0 \le \phi \le \pi$ and $0 \le \theta \le 2\pi$ for the region B'_R so that

$$\iiint\limits_{D'_R} |f| dV \le \iiint\limits_{B'_R} |f| dV \le \iiint\limits_{B'_R} \frac{M\, dV}{\|\mathbf{r}\|^\nu} = 4\pi M \int\limits_R^\infty \frac{\rho^2\, d\rho}{\rho^\nu}$$
$$= 4\pi M \int\limits_R^\infty \frac{d\rho}{\rho^{\nu-2}},$$

which converges if $\nu > 3$. □

106.3. Study Problems.

Problem 14.5. *Evaluate the triple integral of* $f(x,y,z) = (x^2 + y^2)^{-1/2}(x^2 + y^2 + z^2)^{-1/2}$ *over* E*, which is bounded by the cone* $z = \sqrt{x^2 + y^2}$ *and the sphere* $x^2 + y^2 + z^2 = 1$ *if it exists.*

Solution: The function is singular at all points on the z axis. Consider E_ε obtained from E by eliminating from the latter a cone $\phi \le \varepsilon$ and a ball $\rho \le \varepsilon$, where ρ and ϕ are spherical coordinates. To investigate the integrability, consider $|f|dV = f\,dV$ in the spherical coordinates: $f\,dV = (\rho^2 \sin\phi)^{-1}\rho^2 \sin\phi\, d\rho\, d\phi\, d\theta = d\rho\, d\phi\, d\theta$ which is regular. So the function f is integrable as the image E' of E in the spherical coordinate is a rectangle (i.e., it is bounded). Hence,

$$\lim_{\varepsilon\to 0}\iiint_{E_\varepsilon} f\,dV = \lim_{\varepsilon\to 0}\iiint_{E'_\varepsilon} d\rho\, d\phi\, d\theta = \int_0^{2\pi} d\theta \int_0^{\pi/4} d\phi \int_0^1 d\rho = \frac{\pi^2}{2}\,.$$

So the Jacobian cancels out all the singularities of the function. □

107. Line Integrals

Consider a wire made of a nonhomogeneous material. The inhomogeneity means that, if one takes a small piece of the wire of length Δs at a point $\mathbf{r}$, then its mass Δm depends on the point $\mathbf{r}$. It can therefore be characterized by a *linear* mass density (the mass per unit length at a point $\mathbf{r}$):

$$\sigma(\mathbf{r}) = \lim_{\Delta s\to 0}\frac{\Delta m(\mathbf{r})}{\Delta s}.$$

Suppose that the linear mass density is known as a function of $\mathbf{r}$. What is the total mass of the wire that occupies a space curve C? If the curve C has a length L, then it can be partitioned into N small segments of length $\Delta s = L/N$. If $\mathbf{r}_p^*$ is a sample point in the pth segment, then the total mass reads

$$M = \lim_{N\to\infty}\sum_{p=1}^{N}\sigma(\mathbf{r}_p^*)\,\Delta s,$$

where the mass of the pth segment is approximated by $\Delta m_p \approx \sigma(\mathbf{r}_p^*)\,\Delta s$ and the limit is required because this approximation becomes exact only in the limit $\Delta s \to 0$. The expression for M resembles the limit of a Riemann sum and leads to the concept of a *line integral* of σ along a curve C.

107.1. Line Integral of a Function. Let f be a bounded function in E and let C be a smooth (or piecewise-smooth) curve in E. Suppose C has a finite arc length. Consider a partition of C by its N pieces C_p of length Δs_p, $p = 1, 2, ..., N$, which is the arc length of C_p (it exists for a smooth curve!). Put $m_p = \inf_{C_p} f$ and $M_p = \sup_{C_P} f$; that is, m_p is the largest lower bound of values of f for all $\mathbf{r} \in C_p$, and M_p is the smallest upper bound on the values of f for all $\mathbf{r} \in C_p$. The upper and lower sums are defined by $U(f, N) = \sum_{p=1}^{N} M_p \, \Delta s_p$ and $L(f, N) = \sum_{p=1}^{N} m_p \, \Delta s_p$.

DEFINITION 14.15. (Line Integral of a Function).
The line integral of a function f along a piecewise-smooth curve C is

$$\int_C f(\mathbf{r}) \, ds = \lim_{N\to\infty} U(f, N) = \lim_{N\to\infty} L(f, N),$$

provided the limits of the upper and lower sums exist and coincide. The limit is understood in the sense that $\max \Delta s_p \to 0$ *as* $N \to \infty$ *(the partition element of the maximal length becomes smaller as N increases).*

The line integral can also be represented by the limit of a Riemann sum:

$$\int_C f(\mathbf{r}) \, ds = \lim_{N\to\infty} \sum_{p=1}^{N} f(\mathbf{r}_p^*) \, \Delta s_p = \lim_{N\to\infty} R(f, N).$$

If the line integral exists, it follows from the inequality $m_p \leq f(\mathbf{r}) \leq M_p$ for all $\mathbf{r} \in C_p$ that $L(f, N) \leq R(f, N) \leq U(f, N)$, and the limit of the Riemann sum is *independent* of the choice of sample points $\mathbf{r}_p^*$.

It is also interesting to establish a relation of the line integral with a triple (or double) integral. Suppose that f is integrable on a region E that looks like a wire of the shape C with a cross section of a small area ΔA at any point; that is, E is a "cylinder" whose axis is the curve C. Then, in the limit $\Delta A \to 0$ (in the sense that the diameter of the area element goes to 0),

$$\frac{1}{\Delta A} \iiint_E f(\mathbf{r}) \, dV \to \int_C f(\mathbf{r}) \, ds. \tag{14.26}$$

In other words, line integrals can be viewed as the limiting case of triple (or double) integrals when two (or one) dimensions of the integration region become infinitesimally small. This follows immediately from considering a partition of E by volume elements $\Delta V_p = \Delta \, A \Delta s_p$ in (14.13). In particular, it can be concluded that *the line integral exists for any f that is continuous or has only a finite number of bounded* *jump discontinuities along C.* Also, *the line integral inherits all the properties of multiple integrals.*

The evaluation of a line integral is based on the following theorem.

THEOREM 14.15. (Evaluation of a Line Integral).
Suppose that f is continuous in a region that contains a smooth curve C. Let a vector function $\mathbf{r}(t)$, $t \in [a, b]$, trace out the curve C just once. Then

$$\int_C f(\mathbf{r})\, ds = \int_a^b f(\mathbf{r}(t))\|\mathbf{r}'(t)\|\, dt. \tag{14.27}$$

PROOF. Consider a partition of $[a, b]$, $t_p = a + p\,\Delta t$, $p = 0, 1, 2, ..., N$, where $\Delta t = (b - a)/N$. It induces a partition of C by pieces C_p so that $\mathbf{r}(t)$ traces out C_p when $t \in [t_{p-1}, t_p]$, $p = 1, 2, ..., N$. The arc length of C_p is $\int_{t_{p-1}}^{t_p} \|\mathbf{r}'(t)\|\, dt = \Delta s_p$. Since C is smooth, the tangent vector $\mathbf{r}'(t)$ is a continuous function and so is its length $\|\mathbf{r}'(t)\|$. By the integral mean value theorem, there is $t_p^* \in [t_{p-1}, t_p]$ such that $\Delta s_p = \|\mathbf{r}'(t_p^*)\|\Delta t$. Since f is integrable along C, the limit of its Riemann sum is independent of the choice of sample points and a partition of C. Choose the sample points to be $\mathbf{r}_p^* = \mathbf{r}(t_p^*)$. Therefore,

$$\int_C f\, ds = \lim_{N\to\infty} \sum_{p=1}^{N} f(\mathbf{r}(t_p^*))\|\mathbf{r}'(t_p^*)\|\Delta t = \int_a^b f(\mathbf{r})\|\mathbf{r}'(t)\|dt.$$

Note that the Riemann sum for the line integral becomes a Riemann sum of the function $F(t) = f(\mathbf{r}(t))\|\mathbf{r}'(t)\|$ over an interval $t \in [a, b]$. Its limit exists by the continuity of F and equals the integral of F over $[a, b]$. □

The conclusion of the theorem still holds if f has a finite number of bounded jump discontinuities and C is piecewise smooth. The latter implies that the tangent vector may only have a finite number of discontinuities and so does $\|\mathbf{r}'\|$. Therefore, $F(t)$ has only a finite number of bounded jump discontinuities and hence is integrable.

107.2. Evaluation of a Line Integral.

Step 1. Find the parametric equation of a curve C, $\mathbf{r}(t) = (x(t), y(t), z(t))$.
Step 2. Restrict the range of the parameter t to an interval $[a, b]$ so that $\mathbf{r}(t)$ traces out C only once when $t \in [a, b]$.
Step 3. Calculate the derivative $\mathbf{r}'(t)$ and its norm $\|\mathbf{r}'(t)\|$.

Step 4. Substitute $x = x(t)$, $y = y(t)$, and $z = z(t)$ into $f(x, y, z)$ and evaluate the integral (14.27).

Remark. A curve C may be traced out by different vector functions. The value of the line integral is *independent* of the choice of parametric equations because its definition is given only in parameterization-invariant terms (the arc length and values of the function on the curve). The integrals (14.27) written for two different parameterizations of C are related by a change of the integration variable (recall the concept of reparameterization of a spatial curve).

EXAMPLE 14.29. *Evaluate the line integral of* $f(x, y) = x^2y$ *over a circle of radius* R *centered at the point* $(0, a)$.

SOLUTION: The equation of a circle of radius R centered at the origin is $x^2 + y^2 = R^2$. It has familiar parametric equations $x = R\cos t$ and $y = R\sin t$, where t is the angle between $\mathbf{r}(t)$ and the positive x axis counted counterclockwise. The equation of the circle in question is $x^2+(y-a)^2 = R^2$. So, by analogy, one can put. $x = R\cos t$ and $y-a = R\sin t$ (by shifting the origin to the point $(0, a)$). Parametric equation of the circle can be taken in the form $\mathbf{r}(t) = (R\cos t, a + R\sin t)$. The range of t must be restricted to the interval $t \in [0, 2\pi]$ so that $\mathbf{r}(t)$ traces the circle only once. Then $\mathbf{r}'(t) = (-R\sin t, R\cos t)$ and $\|\mathbf{r}'(t)\| = \sqrt{R^2\sin^2 t + R^2\cos^2 t} = R$. Therefore,

$$\int_C x^2y\,ds = \int_0^{2\pi} (R\cos t)^2(a+R\sin t)R\,dt = R^2a\int_0^{2\pi} \cos^2 t\,dt = \pi R^2a,$$

where the integral of $\cos^2 t\sin t$ over $[0, 2\pi]$ vanishes by the periodicity of the cosine function. The last integral is evaluated with the help of the double-angle formula $\cos^2 t = (1 + \cos(2t))/2$. □

EXAMPLE 14.30. *Evaluate the line integral of* $f(x, y, z) = \sqrt{3x^2 + 3y^2 - z^2}$ *over the curve of intersection of the cylinder* $x^2+y^2 = 1$ *and the plane* $x + y + z = 1$.

SOLUTION: Since the curve lies on the cylinder, one can always put $x = \cos t$, $y = \sin t$, and $z = z(t)$, where $z(t)$, is to be found from the condition that the curve also lies in the plane: $x(t) + y(t) + z(t) = 0$ or $z(t) = -\cos t - \sin t$. The interval of t is $[0, 2\pi]$ as the curve winds about the cylinder. Therefore, $\mathbf{r}'(t) = (-\sin t, \cos t, \sin t - \cos t)$ and $\|\mathbf{r}'(t)\| = \sqrt{2 - 2\sin t\cos t} = \sqrt{2 - \sin(2t)}$. The values of the function along the curve are $f = \sqrt{3 - (\cos t + \sin t)^2} = \sqrt{2 - \sin(2t)}$. Note that the function is defined only in the region $3(x^2 + y^2) \geq z^2$ (outside the double cone). It happens that the curve C lies in the domain of f

because its values along C are well defined as $2 > \sin(2t)$ for any t. Hence,

$$\int_C f\, ds = \int_0^{2\pi} \sqrt{2-\sin(2t)}\sqrt{2-\sin(2t)}dt = \int_0^{2\pi} (2-\sin(2t))\, dt = 4\pi$$

□

108. Surface Integrals

108.1. Surface Area. Let S be a surface in space. Suppose that it admits an algebraic description as a graph of a function of two variables, $z = f(x, y)$, where $(x, y) \in D$, or, at least, it can be viewed as a union of a few graphs. For example, a sphere $x^2 + y^2 + z^2 = 1$ is the union of two graphs, $z = \sqrt{1 - x^2 - y^2}$ and $z = -\sqrt{1 - x^2 - y^2}$, where (x, y) are in the disk D of unit radius, $x^2 + y^2 \le 1$. What is the area of the surface?

The question can be answered by the standard trick of integral calculus. Consider a rectangular partition of D. Let ΔS_{ij} be the area of the part of the graph that lies above the rectangle $(x, y) \in [x_i, x_i + \Delta x] \times [y_j, y_j + \Delta y] = R_{ij}$. The total surface area is the sum of all ΔS_{ij}. If the graph is a smooth surface (i.e., the function f is differentiable on D) then ΔS_{ij} can be approximated by the area of the parallelogram that lies above R_{ij} in the tangent plane to the graph through the point (x_i^*, y_j^*, z_{ij}^*), where $z_{ij}^* = f(x_i^*, y_j^*)$ and $(x_i^*, y_j^*) \in R_{ij}$ is any sample point. Recall that the differentiability of f means that the linearization of f (or the tangent plane approximation) becomes more and more accurate as $(\Delta x, \Delta y) \to (0, 0)$. Therefore, in this limit, Δx and Δy can be viewed as the differentials dx and dy, and the areas ΔS_{ij} and $\Delta A = \Delta x\, \Delta y$ must be proportional:

$$\Delta S_{ij} = J_{ij}\, \Delta A.$$

The coefficient J_{ji} is found by comparing the area of the parallelogram in the tangent plane above R_{ij} with the area ΔA of R_{ij}. Think of the roof of a building of shape $z = f(x, y)$ covered by shingles of area ΔS_{ij}. The equation of the tangent plane is

$$z = z_{ij}^* + f_x'(x_i^*, y_j^*)(x - x_i^*) + f_y'(x_i^*, y_j^*)(y - y_j^*) = L(x, y).$$

Let O', A', and B' be, respectively, the vertices $(x_i, y_j, 0)$, $(x_i{+}\Delta x, y_j, 0)$, and $(x_i, y_j + \Delta y, 0)$ of the rectangle R_{ij}; that is, the segments $O'A'$ and $O'B'$ are the adjacent sides of R_{ij}. If O, A, and B are the points in the tangent plane above O', A', and B', respectively, then the adjacent sides of the parallelogram in question are $\mathbf{a} = \vec{OA}$ and $\mathbf{b} = \vec{OB}$ and $\Delta S_{ij} = \|\mathbf{a} \times \mathbf{b}\|$.

By substituting O' into the tangent plane equation, the coordinates of the point O are found, $(x_i, y_j, L(x_i, y_j))$. By substituting A' into the tangent plane equation, the coordinates of the point A are found, $(x_i+\Delta x, y_j, L(x_i+\Delta x, y_j))$. By the linearity of the function L, $L(x_i+\Delta x, y_j) - L(x_i, y_j) = f'_x(x_i^*, y_j^*)\,\Delta x$ and $\mathbf{a} = (\Delta x, 0, f'_x\,\Delta x)$. Similarly, $\mathbf{b} = (0, \Delta y, f'_y\,\Delta y)$. Hence,

$$\mathbf{a}\times\mathbf{b} = (-f'_x,\ -f'_y,\ 1)\ \Delta x\,\Delta y\,,$$

$$\Delta S_{ij} = \|\mathbf{a}\times\mathbf{b}\| = \sqrt{1+(f'_x)^2+(f'_y)^2}\ \Delta A = J(x_i^*, y_j^*)\,\Delta A,$$

where $J(x,y) = \sqrt{1+(f'_x)^2+(f'_y)^2}$. Thus, the surface area is given by

$$A(S) = \lim_{(\Delta x,\Delta y)\to(0,0)} \sum_{ij} J(x_i^*, y_j^*)\ \Delta A.$$

Since the derivatives of f are continuous, the function $J(x,y)$ is continuous on D, and the Riemann sum converges to the double integral of J over D.

Theorem 14.16. (Surface Area).
Suppose that $f(x,y)$ has continuous first-order partial derivatives on D. Then the surface area of the graph $z = f(x,y)$ is given by

$$A(S) = \iint_D \sqrt{1+(f'_x)^2+(f'_y)^2}\ dA.$$

If $z = \text{const}$, then $f'_x = f'_y = 0$ and $A(S) = A(D)$ as required because S is D moved parallel into the plane $z = \text{const}$.

Example 14.31. *Prove that the surface area of a sphere of radius R is $4\pi R^2$.*

Solution: The hemisphere is the graph $z = f(x,y) = \sqrt{R^2-x^2-y^2}$ on the disk $x^2+y^2 \le R^2$ of radius R. The area of the sphere is twice the area of this graph. One has $f'_x = -x/f$ and $f'_y = -y/f$. Therefore, $J = (1+x^2/f^2+y^2/f^2)^{1/2} = (f^2-x^2-y^2)^{1/2}/f = R/f$. Hence,

$$A(S) = 2R\iint_D \frac{dA}{\sqrt{R^2-x^2-y^2}} = 2R\int_0^{2\pi} d\theta \int_0^R \frac{r\,dr}{\sqrt{R^2-r^2}}$$

$$= 4\pi R\int_0^R \frac{r\,dr}{\sqrt{R^2-r^2}} = 2\pi R\int_0^{R^2} \frac{du}{\sqrt{u}} = 4\pi R^2,$$

where the double integral has been converted to polar coordinates and the substitution $u = R^2 - r^2$ has been used to evaluate the last integral. □

EXAMPLE 14.32. *Find the area of the part of the paraboloid $z = x^2 + y^2$ in the first octant and below the plane $z = 4$.*

SOLUTION: The surface in question is the graph $z = f(x, y) = x^2 + y^2$. Next, the region D must be specified (it determines the part of the graph whose area is to be found). One can view D as the vertical projection of the surface onto the xy plane. The plane $z = 4$ intersects the paraboloid along the circle $4 = x^2 + y^2$ of radius 2. Since the surface also lies in the first octant, D is the part of the disk $x^2 + y^2 \leq 4$ in the first quadrant. Then $f'_x = 2x$, $f'_y = 2y$, and $J = (1 + 4x^2 + 4y^2)^{1/2}$. The surface area is

$$\begin{aligned} A(S) &= \iint_D \sqrt{1 + 4x^2 + 4y^2}\, dA = \int_0^{\pi/2} d\theta \int_0^2 \sqrt{1 + 4r^2}\, r\, dr \\ &= \frac{\pi}{2} \int_0^2 \sqrt{1 + 4r^2}\, r\, dr = \frac{\pi}{16} \int_1^{17} \sqrt{u}\, du = \frac{\pi}{24}(17^{3/2} - 1), \end{aligned}$$

where the double integral has been converted to polar coordinates and the substitution $u = 1 + 4r^2$ has been used to evaluate the last integral. □

108.2. Surface Integral of a Function. An intuitive idea of the concept of the surface integral of a function can be understood from the following example. Suppose one wants to find the total human population on the globe. The data about the population is usually supplied as the population *density* (i.e., the number of people per unit area). The population density is not a constant function on the globe. It is high in cities and low in deserts and jungles. Therefore, the surface of the globe must be partitioned by surface elements of area ΔS_p. If $\sigma(\mathbf{r})$ is the population density as a function of position $\mathbf{r}$ on the globe, then the population on each partition element is approximately $\sigma(\mathbf{r}_p^*)\, \Delta S_p$, where $\mathbf{r}_p^*$ is a sample point in the partition element. The approximation neglects variations of σ within each partition element. The total population is approximately the Riemann sum $\sum_p \sigma(\mathbf{r}_p^*)\, \Delta S_p$. To get an exact value, the partition has to be refined so that the size of each partition element becomes smaller. The limit is the surface integral of σ over the surface of the globe, which is the total population. In general, one can think of some quantity distributed over a surface with some density (the amount of this quantity per unit area as a function of position on the surface). The total amount is the surface integral of the density over the surface.

Let f be a bounded function in E and let S be a surface in E that has a finite surface area. Consider a partition of S by N pieces S_p,

$p = 1, 2, ..., N$, which have surface area ΔS_p. Put $m_p = \inf_{S_p} f$ and $M_p = \sup_{S_P} f$; that is, m_p is the largest lower bound of values of f for all $\mathbf{r} \in S_p$ and M_p is the smallest upper bound on the values of f for all $\mathbf{r} \in S_p$. The upper and lower sums are defined by $U(f, N) = \sum_{p=1}^{N} M_p \, \Delta S_p$ and $L(f, N) = \sum_{p=1}^{N} m_p \, \Delta S_p$. Let R_p be the radius of the smallest ball that contains S_p and $\max_p R_p = R_N$. A partition of S is said to be refined if $R_{N'} < R_N$ for $N' > N$. In other words, under the refinement, the sizes R_p of each partition element become uniformly smaller.

Definition 14.16. (Surface Integral of a Function). *The surface integral of a function f over a surface S is*

$$\iint_S f(\mathbf{r}) \, dS = \lim_{N\to\infty} U(f, N) = \lim_{N\to\infty} L(f, N),$$

provided the limits of the upper and lower sums exist and coincide. The limit is understood in the sense $R_N \to 0$ as $N \to \infty$.

The surface integral can also be represented by the limit of a Riemann sum:

$$\iint_S f(\mathbf{r}) \, dS = \lim_{N\to\infty} \sum_{p=1}^{N} f(\mathbf{r}_p^*) \, \Delta S_p = \lim_{N\to\infty} R(f, N). \tag{14.28}$$

If the surface integral exists, it follows from the inequality $m_p \leq f(\mathbf{r}) \leq M_p$ for all $\mathbf{r} \in S_p$ that $L(f, N) \leq R(f, N) \leq U(f, N)$, and the limit of the Riemann sum is *independent* of the choice of sample points $\mathbf{r}_p^*$. Riemann sums can be used in numerical approximations of the surface integral.

Similar to line integrals, surface integrals are related to triple integrals. Suppose that f is integrable on a region E that looks like a shell of the shape S with a constant small thickness Δh. Then, in the limit $\Delta h \to 0$,

$$\frac{1}{\Delta h} \iiint_E f(\mathbf{r}) \, dV \to \iint_S f(\mathbf{r}) \, dS. \tag{14.29}$$

This follows immediately by considering a partition of E by volume elements $\Delta V_p = \Delta h \, \Delta S_p$ in (14.13). Hence, *the surface integral exists for any f that is continuous or has bounded jump discontinuities along a finite number of smooth curves on S, and it inherits all the properties of multiple integrals.*

108.3. Evaluation of a Surface Integral.

Theorem 14.17. (Evaluation of a Surface Integral).
Suppose that f is continuous in a region that contains a surface S defined by the graph $z = g(x, y)$ on D. Suppose that g has continuous first-order partial derivatives on D. Then
(14.30)

$$\iint_S f(x, y, z)\, dS = \iint_D f(x, y, g(x, y))\sqrt{1 + (g'_x)^2 + (g'_y)^2}\, dA.$$

Consider a partition of D by elements D_p of area ΔA_p, $p = 1, 2, ...,$ N. Let $J(x, y) = \sqrt{1 + (g'_x)^2 + (g'_y)^2}$. By the continuity of g'_x and g'_y, J is continuous on D. By the integral mean value theorem, the area of the part of the graph $z = g(x, y)$ over D_p is given by

$$\Delta S_p = \iint_{D_p} J(x, y)\, dA = J(x^*_p, y^*_p)\, \Delta A_p$$

for some $(x^*_p, y^*_p) \in D_p$. In the Riemann sum for the surface integral (14.28), take the sample points to be $\mathbf{r}^*_p = (x^*_p, y^*_p, g(x^*_p, y^*_p)) \in S_p$. The Riemann sum becomes the Riemann sum (14.3) of the function $F(x, y) = f(x, y, g(x, y))J(x, y)$ on D. By the continuity of F, it converges to the double integral of F over D. The argument given here is based on a tacit assumption that the surface integral exists according to Definition 14.16, and hence the limit of the Riemann sum exists and is independent of the choice of sample points. It can be proved that under the hypothesis of the theorem the surface integral exists.

The evaluation of the surface integral involves the following steps:

Step 1. Represent S as a graph $z = g(x, y)$ (i.e., find the function g using a geometrical description of S).
Step 2. Find the region D that defines the part of the graph that coincides with S (if S is not the entire graph).
Step 3. Calculate the derivatives g'_x and g'_y and the area transformation function J, $dS = J\, dA$.
Step 4. Evaluate the double integral (14.30).

Example 14.33. *Evaluate the integral of $f(x, y, z) = z$ over the part of the saddle surface $z = xy$ that lies inside the cylinder $x^2+y^2 = 1$ in the first octant.*

Solution: The surface is a part of the graph $z = g(x, y) = xy$. Since it lies within the cylinder, its projection onto the xy plane is bounded by the circle of unit radius, $x^2 + y^2 = 1$. Thus, D is the quarter of the disk $x^2 + y^2 \leq 1$ in the first quadrant. One has $g'_x = y$, $g'_y = x$, and $J(x, y) = (1 + x^2 + y^2)^{1/2}$. The surface integral is

$$\begin{aligned}
\iint_S z\,dS &= \iint_D xy\sqrt{1 + x^2 + y^2}\,dA \\
&= \int_0^{\pi/2} \cos\theta \sin\theta\,d\theta \int_0^1 r^2\sqrt{1 + r^2}\;r\,dr \\
&= \frac{\sin^2\theta}{2}\Big|_0^{\pi/2} \frac{1}{2}\int_1^2 (u - 1)\sqrt{u}\,du \\
&= \frac{1}{2}\Big(\frac{u^{5/2}}{5} - \frac{u^{3/2}}{3}\Big)\Big|_1^2 = \frac{2(4\sqrt{2} + 1)}{15},
\end{aligned}$$

where the double integral has been converted to polar coordinates and the last integral is evaluated by the substitution $u = 1 + r^2$. □

108.4. Parametric Equations of a Surface. The graph $z = g(x, y)$, where $(x, y) \in D$ defines a surface S in space. Consider the vectors $\mathbf{r}(u, v) = (u, v, g(u, v))$ where the pair of parameters (u, v) spans the region D. The vector function $\mathbf{r}(u, v)$ of two variables defines a one-to-one mapping of the region D into space so that the image of D is the surface S. Consider now a region D spanned by the ordered pairs (u, v). Three functions $x(u, v)$, $y(u, v)$, and $z(u, v)$ on D define a mapping of D into space $\mathbf{r}(u, v) = (x(u, v), y(u, v), z(u, v))$. The range of this mapping is called a *surface in space*, and the equations $x = x(u, v)$, $y = y(u, v)$, and $z = z(u, v)$ are called *parametric equations* of this surface.

For example, the equations

$$x = R\cos v \sin u\,, \quad y = R\sin v \sin u\,, \quad z = R\cos u \tag{14.31}$$

are parametric equations of a sphere of radius R. Indeed, by comparing these equations with the spherical coordinates, one finds that $(\rho, \phi, \theta) = (R, u, v)$; that is, when (u, v) range over the rectangle $[0, \pi] \times [0, 2\pi]$, the vector $(x, y, z) = \mathbf{r}(u, v)$ traces out the sphere $\rho = R$. An apparent advantage of using parametric equations of a surface is that the surface no longer needs be represented as a graph. Here a sphere is described by one vector-valued function of two variables.

Definition 14.17. *Let $\mathbf{r}(u, v)$ be a vector function on an open region D that has continuous partial derivatives $\mathbf{r}'_u$ and $\mathbf{r}'_v$ on D. The range S of the vector function is called a smooth surface if S is covered just once as (u, v) ranges throughout D and the vector $\mathbf{r}'_u \times \mathbf{r}'_v$ is not zero.*

An analogy can be made with parametric equations of a curve in space. A curve in space is a mapping of an *interval* $[a, b]$ into space defined by a vector function of *one variable* $\mathbf{r}(t)$. If $\mathbf{r}'(t)$ is continuous and $\mathbf{r}'(t) \neq \mathbf{0}$, then the curve has a continuous tangent vector and the curve is smooth. Similarly, the condition $\mathbf{r}'_u \times \mathbf{r}'_v \neq \mathbf{0}$ ensures that the surface has a continuous normal vector just like a graph of a continuously differentiable function of two variables. This will be explained shortly after the discussion of a few examples.

EXAMPLE 14.34. *Find the parametric equations of the double cone* $z^2 = x^2 + y^2$.

SOLUTION: Suppose $z \neq 0$. Then $(x/z)^2 + (y/z)^2 = 1$. The solution of this equation is $x/z = \cos u$ and $y/z = \sin u$, where $u \in [0, 2\pi)$. Therefore, the parametric equations are

$$x = v \cos u\,, \quad y = v \sin u\,, \quad z = v,$$

where $(u, v) \in [0, 2\pi) \times (-\infty, \infty)$ for the whole double cone. Of course, there are many different parameterizations of the same surface. They are related by a change of variables $(u, v) \in D \leftrightarrow (s, t) \in D'$, where $s = s(u, v)$ and $t = t(u, v)$. □

EXAMPLE 14.35. *A torus is a surface obtained by rotating a circle about an axis outside the circle and parallel to its diameter. Find the parametric equations of a torus.*

SOLUTION: Let the rotation axis be the z axis. Let R be the distance from the z axis to the center of the rotated circle and let a be the radius of the latter, $a \leq R$. In the xz plane, the rotated circle is $z^2 + (x - a)^2 = R^2$. Let $(x_0, 0, z_0)$ be a solution to this equation. The point $(x_0, 0, z_0)$ traces out the circle of radius x_0 upon the rotation about the z axis. All such points are $(x_0 \cos v, x_0 \sin v, z_0)$, where $v \in [0, 2\pi)$. All that is left is to parameterize all solutions $(x_0, 0, z_0)$, which is simply $z_0 = R \sin u$ and $x_0 - a = R \cos u$. Thus, the parametric equations of a torus are

$$x = (R + a \cos u) \cos v\,, \quad y = (R + a \cos u) \sin v\,, \quad z = R \sin u\,, \tag{14.32}$$

where $(u, v) \in [0, 2\pi) \times [0, 2\pi)$ □

The parametric equations of a surface are convenient for evaluating the surface integrals. If the region D spanned by the parameters (u, v) is partitioned by rectangles of area $\Delta A = \Delta u\, \Delta v$, then the mapping $\mathbf{r}(u, v)$ defines a partition of the surface. So the surface integral can be converted into a double integral over D if one establishes the area transformation law $\Delta S = J\, \Delta A$. Consider a rectangle $(u, v) \in [u_0, u_0+$

$\Delta u] \times [v_0, v_0 + \Delta v] = R_0$. Let its vertices O', A', and B' have the coordinates (u_0, v_0), $(u_0 + \Delta u, v_0)$, and $(u_0, v_0 + \Delta v)$, respectively. The segments $O'A'$ and $O'B'$ are the adjacent sides of the rectangle R_0. Let O, A, and B be the images of these points in the surface. Their position vectors are $\mathbf{r}_0 = \mathbf{r}(u_0, v_0)$, $\mathbf{r}_a = \mathbf{r}(u_0 + \Delta u, v_0)$, and $\mathbf{r}_a = \mathbf{r}(u_0, v_0 + \Delta v)$, respectively. The area ΔS of the image of the rectangle R_0 can be approximated by the area of the parallelogram with adjacent sides:

$$\mathbf{a} = \vec{OA} = \mathbf{r}_a - \mathbf{r}_0 = \mathbf{r}(u_0 + \Delta u, v_0) - \mathbf{r}(u_0, v_0) = \mathbf{r}'_u(u_0, v_0)\,\Delta u,$$
$$\mathbf{b} = \vec{OB} = \mathbf{r}_b - \mathbf{r}_0 = \mathbf{r}(u_0, v_0 + \Delta v) - \mathbf{r}(u_0, v_0) = \mathbf{r}'_v(u_0, v_0)\,\Delta v,$$

which hold in the limit $(\Delta u, \Delta v) \to (0, 0)$ (when $du = \Delta u$ and $dv = \Delta v$) under the assumption that the components of $\mathbf{r}(u, v)$ are *continuously differentiable*. Note that if the surface is a graph $z = g(x, y)$, then $\mathbf{r}(u, v) = (u, v, g(u, v))$, and the vectors $\mathbf{a}$ and $\mathbf{b}$ are given by the familiar expressions $\mathbf{a} = (1, 0, g'_u)\,\Delta u$ and $\mathbf{b} = (0, 1, g'_v)\,\Delta v$. The vector $\mathbf{r}(u, v_0)$ (one argument is fixed, $v = v_0$) traces out a curve in the surface. The derivative $\mathbf{r}'_u(u, v_0)$ is tangent to the curve and hence to the surface. A similar argument applies to $\mathbf{r}'_v$. Thus, the derivatives $\mathbf{r}'_u$ and $\mathbf{r}'_v$ are tangent to the surface and, hence, their cross product must be normal to it.

Corollary 14.5. (Normal to a Parametric Surface).
Let a smooth surface be described by the parametric equations $\mathbf{r} = \mathbf{r}(u, v)$*. Then the vector* $\mathbf{n} = \mathbf{r}'_u \times \mathbf{r}'_v$ *is normal to the surface.*

The area transformation law is now easy to find: $\Delta S = \|\mathbf{a} \times \mathbf{b}\| = \|\mathbf{r}'_u \times \mathbf{r}'_v\| \Delta A$. The surface integral can be written as the double integral

$$\iint_S f(\mathbf{r})\,dS = \iint_D f(\mathbf{r}(u, v))\|\mathbf{r}'_u \times \mathbf{r}'_v\|\,dA$$

and, in particular,

$$A(S) = \iint_D \|\mathbf{r}'_u \times \mathbf{r}'_v\|\,dA.$$

Example 14.36. *Find the surface area of the torus (14.32).*

Solution: To simplify the notation, put $w = R + a\cos u$. One has

$$\begin{aligned}
\mathbf{r}'_u &= (-a\sin u\cos v,\ -a\sin u\sin v,\ R\cos u),\\
\mathbf{r}'_v &= (-(R + a\cos u)\sin v,\ (R + a\cos u)\cos v,\ 0)\\
&= w(-\sin v,\ \cos v,\ 0),\\
\mathbf{n} &= \mathbf{r}'_u \times \mathbf{r}'_v = w(-a\cos v\cos u,\ -a\cos v\cos u,\ -a\sin u),\\
J &= \|\mathbf{r}'_u \times \mathbf{r}'_v\| = aw = a(R + a\cos u).
\end{aligned}$$

The surface area is

$$A(S) = \iint_D J(u,v)\,dA = \int_0^{2\pi}\int_0^{2\pi} a(R + a\cos u)\,dv\,du = 4\pi^2 Ra$$

□

EXAMPLE 14.37. *Evaluate the surface integral of* $f(x,y,z) = z^2(x^2+y^2)$ *over a sphere of radius* R *centered at the origin.*

SOLUTION: Using the parametric equations (14.31), one finds

$$\begin{aligned}
\mathbf{r}'_u &= (R\cos v\cos u,\ R\sin v\cos u,\ -R\sin u),\\
\mathbf{r}'_v &= (-R\sin v\sin u,\ R\cos v\sin u,\ 0)\\
&= R\sin u(-\sin v,\ \cos v,\ 0),\\
\mathbf{n} &= \mathbf{r}'_u \times \mathbf{r}'_v = R\sin u(R\sin u\cos v,\ R\sin u\sin v,\ R\cos u)\\
&= R\sin u\ \mathbf{r}(u,v),\\
J &= \|\mathbf{r}'_u \times \mathbf{r}'_v\| = R^2\sin u,\\
f(\mathbf{r}(u,v)) &= (R\cos u)^2R^2\sin^2 u = R^4\cos^2 u(1-\cos^2 u).
\end{aligned}$$

Note that $\sin u \geq 0$ because $u \in [0,\pi]$ ($u = \phi$ and $v = \theta$). Therefore, the normal vector $\mathbf{n}$ is outward (parallel to the position vector; the inward normal would be opposite to the position vector.) The surface integral is

$$\begin{aligned}
\iint_S f\,dS &= \iint_D f(\mathbf{r}(u,v))J(u,v)\,dA\\
&= R^6\int_0^{2\pi}dv\int_0^{\pi}\cos^2 u(1-\cos^2 u)\sin u\,du\\
&= 2\pi R^6\int_{-1}^{1} w^2(1-w^2)\,dw = \frac{8\pi}{15}R^6,
\end{aligned}$$

where the substitution $w = \cos u$ has been made to evaluate the last integral. □

109. Moments of Inertia and Center of Mass

An important application of multiple integrals is finding the *center of mass* and *moments of inertia* of an extended object. The laws of mechanics say that the center of mass of an extended object on which no external force acts moves along a straight line with a constant speed. In other words, the center of mass is a particular point of an extended object that defines the trajectory of the object as a whole. The motion of an extended object can be viewed as a combination of the motion of its center of mass and rotation about its center of mass. The kinetic energy of the object is

$$K = \frac{Mv^2}{2} + K_{\text{rot}},$$

where M is the total mass of the object, v is the speed of its center of mass, and K_{rot} is the kinetic energy of rotation of the object about its center of mass; the latter quantity is determined by moments of inertia. For example, when docking a spacecraft to a space station, one needs to know exactly how long the engine should be fired to achieve the required position of its center of mass and the orientation of the craft relative to it, that is, how exactly its kinetic energy has to be changed by firing the engines. So its center of mass and moments of inertia must be known to accomplish the task.

109.1. Center of Mass. Consider a point mass m fixed at an endpoint of a rod that can rotate about its other end. If the rod has length L and the gravitational force is normal to the rod, then the quantity gmL is called the *rotational moment* of the gravitational force mg, where g is the free acceleration. If the rotation is clockwise (the mass is at the right endpoint), the moment is assumed to be positive, and it is negative, $-gmL$, for a counterclockwise rotation (the mass is at the left endpoint). More generally, if the mass has a position x on the x axis, then its rotation moment about a point x_c is $M = (x - x_c)m$ (omitting the constant g). It is negative if $x < x_c$ and positive when $x > x_c$. The center of mass is understood through the concept of rotational moments.

The simplest extended object consists of two point masses m_1 and m_2 connected by a massless rod. Suppose that one point of the rod is fixed so that it can only rotate about that point. The center of mass is the point on the rod such that the object would not rotate about it under a uniform gravitational force applied along the direction perpendicular to the rod. Evidently, the position of the center of mass is determined by the condition that the total rotational moment about it vanishes. Suppose that the rod lies on the x axis so that the masses have the coordinates x_1 and x_2. The total rotational moment of the object about the point x_c is $M = M_1 + M_2 = (x_1 - x_c)m_1 + (x_2 - x_c)m_2$. If x_c is such that $M = 0$, then

$$m_1(x_1 - x_c) + m_2(x_2 - x_c) = 0 \quad \Longrightarrow \quad x_c = \frac{m_1x_1 + m_2x_2}{m_1 + m_2}.$$

The center of mass (x_c, y_c) of point masses m_i, $i = 1, 2, ..., N$, positioned on a plane at (x_i, y_i) can be understood as follows. Think of the plane as a plate on which the masses are positioned. The gravitational force is normal to the plane. If a rod is put underneath the plane, then due an even distribution of masses the plane can rotate about the rod. When the rod is aligned along either the line $x = x_c$ or the line $y = y_c$, the plane with distributed masses on it does not rotate under the gravitational pull. In other words, the rotational moments about the lines $x = x_c$ and $y = y_c$ vanish. The rotational moment about the line $x = x_c$ or $y = y_c$ is determined by the distances of the masses from this line:

$$\sum_{i=1}^{N}(x_i - x_c)m_i = 0 \quad\Longrightarrow\quad x_c = \frac{1}{m}\sum_{i=1}^{N} m_i x_i = \frac{M_y}{m}\,, \quad m = \sum_{i=1}^{N} m_i,$$

$$\sum_{i=1}^{N}(y_i - y_c)m_i = 0 \quad\Longrightarrow\quad y_c = \frac{1}{m}\sum_{i=1}^{N} m_i y_i = \frac{M_x}{m}\,, \quad m = \sum_{i=1}^{N} m_i,$$

where m is the total mass. The quantity M_y is the moment about the y axis (the line $x = 0$), whereas M_x is the moment about the x axis (the line $y = 0$).

The center of mass of a general extended object is defined similarly by demanding that the total moments about either of the planes $x = x_c$, or $y = y_c$, or $z = z_c$ vanish. Thus, if $\mathbf{r}_c$ is the position vector of the center of mass, it satisfies the condition:

$$\sum_i m_i(\mathbf{r}_i - \mathbf{r}_c) = \mathbf{0},$$

where the vectors $\mathbf{r}_i - \mathbf{r}_c$ are position vectors of masses *relative to the center of mass.*

Definition 14.18. (Center of Mass).
Suppose that an extended object consists of N point masses m_i, $i = 1, 2, ..., N$, whose position vectors are $\mathbf{r}_i$. Then its center of mass is a point with the position vector

$$\mathbf{r}_c = \frac{1}{m}\sum_{i=1}^{N} m_i \mathbf{r}_i\,, \qquad m = \sum_{i=1}^{N} m_i, \tag{14.33}$$

where m is the total mass of the object.

If an extended object contains continuously distributed masses, then the object can be partitioned into N small pieces. Let B_i be the smallest ball of radius R_i within which the ith partition piece lies.

Although all the partition pieces are small, they still have finite sizes R_i, and the definition (14.33) cannot be used because the point $\mathbf{r}_i$ could be any point in B_i. By making the usual trick of integral calculus, this uncertainty can be eliminated by taking the limit $N \to \infty$ in the sense that all the partition sizes tend to 0 uniformly, $R_i \leq \max_i R_i = R_N \to 0$ as $N \to \infty$. In this limit, the position of each partition piece can be described by any sample point $\mathbf{r}_i^* \in B_i$. The limit of the Riemann sum is given by the integral over the region E in space occupied by the object. If $\sigma(\mathbf{r})$ is the mass density of the object, then $\Delta m_i = \sigma(\mathbf{r}_i^*)\,\Delta V_i$, where ΔV_i is the volume of the ith partition element and

$$\mathbf{r}_c = \frac{1}{m}\lim_{N\to\infty}\sum_{i=1}^{N}\mathbf{r}_i^*\,\Delta m_i = \frac{1}{m}\iiint_E \mathbf{r}\,\sigma(\mathbf{r})\,dV\,, \qquad m = \iiint_E \sigma(\mathbf{r})\,dV.$$

In practical applications, one often encounters extended objects whose one or two dimensions are small relative to the other (e.g., shell-like objects or wirelike objects). In this case, the triple integral is simplified to either a surface (or double) integral for shell-like E, according to (14.29), or to a line integral, according to (14.26). For two- and one-dimensional extended objects, the center of mass can be written as, respectively,

$$\mathbf{r}_c = \frac{1}{m}\iint_S \mathbf{r}\,\sigma(\mathbf{r})\,dS\,, \qquad m = \iint_S \sigma(\mathbf{r})\,dS,$$
$$\mathbf{r}_c = \frac{1}{m}\int_C \mathbf{r}\,\sigma(\mathbf{r})\,ds\,, \qquad m = \int_C \sigma(\mathbf{r})\,ds,$$

where, accordingly, σ is the surface mass density or the line mass density for two- or one-dimensional objects. In particular, when S is a planar, flat surface, the surface integral turns into a double integral.

The concept of rotational moments is also useful for finding the center of mass using the symmetries of the mass distribution of an extended object. For example, the center of mass of a disk with a uniform mass distribution apparently coincides with the disk center (the disk would not rotate about its diameter under the gravitational pull).

Example 14.38. *Find the center of mass of the half-disk $x^2 + y^2 \leq R^2$, $y \geq 0$, if the mass density at any point is proportional to the distance of that point from the x axis.*

Solution: The mass is distributed evenly to the left and right from the y axis because the mass density is independent of x, $\sigma(x, y) = ky$

(k is a constant). So, the rotational moment about the y axis vanishes; $M_y = 0$ by symmetry and hence $x_c = M_y/m = 0$. The total mass is

$$m = \iint_D \sigma\, dA = k \iint_D y\, dA = k \int_0^{\pi} \int_0^R r \sin\theta \; r\, dr\, d\theta$$
$$= 2k \int_0^R r^2\, dr = \frac{2kR^3}{3},$$

where the integral has been converted to polar coordinates. The moment about the x axis (about the line $y = 0$) is

$$M_x = \iint_D y\sigma\, dA = \int_0^{\pi} \int_0^R k(r\sin\theta)^2 r\, dr\, d\theta = \frac{\pi k}{2} \int_0^R r^3\, dr = \frac{\pi k R^4}{8}.$$

So $y_c = M_x/m = 3\pi R/16$. □

Example 14.39. *Find the center of mass of the solid that lies between spheres of radii $a < b$ centered at the origin and is bounded by the cone $z = \sqrt{x^2 + y^2}/\sqrt{3}$ if the mass density is constant.*

Solution: The mass is evenly distributed about the xz and yz planes. So the moments M_{xz} and M_{yz} about them vanish, and hence $y_c = M_{xz}/m = 0$ and $x_c = M_{yz}/m = 0$. The center of mass lies on the z axis. Put $\sigma = k =$ const. The total mass is

$$m = \iiint_E \sigma\, dV = k \int_0^{2\pi} \int_0^{\pi/3} \int_a^b \rho^2 \sin\phi\, d\rho\, d\phi\, d\theta = \frac{\pi k}{3}\,(b^3 - a^3),$$

where the triple integral has been converted to spherical coordinates. The boundaries of E are the spheres $\rho = a$ and $\rho = b$ and the cone defined by the condition $\cot\phi = 1/\sqrt{3}$ or $\phi = \pi/3$. Therefore, the image E' of E under the transformation to spherical coordinates is the rectangle $(\rho, \phi, \theta) \in E' = [a, b] \times [0, \pi/3] \times [0, 2\pi]$. The full range is taken for the polar angle θ as the equations of the boundaries impose no condition on it. The moment about the xy plane is

$$M_{xy} = \iiint_E z\sigma\, dV = k \int_0^{2\pi} \int_0^{\pi/3} \int_a^b \rho\cos\phi\, \rho^2 \sin\phi\, d\rho\, d\phi\, d\theta$$
$$= \frac{3\pi k}{16}\,(b^4 - a^4).$$

So $z_c = M_{xy}/m = (9/16)(a + b)(a^2 + b^2)/(a^2 + ab + b^2)$. □

109.2. Moments of Inertia. Consider a point mass m rotating about an axis γ at a constant rate of ω rad/s (called the *angular velocity*). If the radius of the circular trajectory is R, then the linear velocity of the object is $v = \omega R$. The object has the kinetic energy

$$K_{\rm rot} = \frac{mv^2}{2} = \frac{mR^2\omega^2}{2} = \frac{I_\gamma \omega^2}{2}.$$

The constant I_γ is called the *moment of inertia* of the point mass m about the axis γ. Similarly, consider an extended object consisting of N point masses. The relative positions of the masses do not change when the object moves. So, if the object rotates about an axis γ at a constant rate ω, then each point mass rotates at the same rate and hence has kinetic energy $m_i R_i^2 \omega^2/2$, where R_i is the distance from the mass m_i to the axis γ. The total kinetic energy is $K_{\rm rot} = I_\gamma \omega^2/2$, where the constant

$$I_\gamma = \sum_{i=1}^{N} m_i R_i^2$$

is called the *moment of inertia of the object about the axis* γ. It is independent of the motion itself and determined solely by the mass distribution and distances of the masses from the rotation axis.

Suppose that the mass is continuously distributed in a region E with the mass density $\sigma(\mathbf{r})$. Let $R_\gamma(\mathbf{r})$ be the distance from a point $\mathbf{r} \in E$ to an axis (line) γ. Consider a partition of E by small elements E_i of volume ΔV_i. The mass of each partition element is $\Delta m_i = \sigma(\mathbf{r}_i^*)\,\Delta V$ for some sample point $\mathbf{r}_i^* \in E_i$ in the limit when all the sizes of partition elements tend to 0 uniformly. The moment of inertia about the axis γ is

$$I_\gamma = \lim_{N\to\infty} \sum_{i=1}^{N} R_\gamma^2(\mathbf{r}_i)\sigma(\mathbf{r}_i^*)\,\Delta V_i = \iiint_E R_\gamma^2(\mathbf{r})\sigma(\mathbf{r})\,dV$$

in accordance with the Riemann sum for triple integrals (14.13). In particular, the distance of a point (x, y, z) from the x-, y-, and z axes is, respectively, $R_x = \sqrt{y^2 + z^2}$, $R_y = \sqrt{x^2 + z^2}$, and $R_z = \sqrt{x^2 + y^2}$. So the moments of inertia about the coordinate axes are

$$I_x = \iiint_E (y^2 + z^2)\sigma\,dV\,, \quad I_y = \iiint_E (x^2 + z^2)\sigma\,dV\,,$$
$$I_z = \iiint_E (x^2 + y^2)\sigma\,dV.$$

In general, if the axis γ goes through the origin parallel to a unit vector $\hat{\mathbf{u}}$, then by the distance formula between a point $\mathbf{r}$ and the line,

(14.34)

$$R_\gamma^2(\mathbf{r}) = \|\hat{\mathbf{u}} \times \mathbf{r}\|^2 = (\hat{\mathbf{u}} \times \mathbf{r}) \cdot (\hat{\mathbf{u}} \times \mathbf{r}) = \hat{\mathbf{u}} \cdot (\mathbf{r} \times (\hat{\mathbf{u}} \times \mathbf{r})) = \mathbf{r}^2 - (\hat{\mathbf{u}} \cdot \mathbf{r})^2,$$

where the $bac - cab$ rule (see Study Problem 11.16) has been used to transform the double cross product.

If one or two dimensions of the object are small relative to the other, the triple integral is reduced to either a surface integral or a line integral, respectively, in accordance with (14.29) or (14.26); that is, for two- or one-dimensional objects, the moment of inertia becomes, respectively,

$$I_\gamma = \iint_S R_\gamma^2(\mathbf{r})\sigma(\mathbf{r})\, dS\,, \qquad I_\gamma = \int_C R_\gamma^2(\mathbf{r})\sigma(\mathbf{r})\, ds,$$

where σ is either the surface or linear mass density.

Example 14.40. *A rocket tip is made of thin plates with a constant surface mass density $\sigma = k$. It has a circular conic shape with base diameter $2a$ and distance h from the tip to the base. Find the moment of inertia of the tip about its axis of symmetry.*

Solution: Set up the coordinate system so that the tip is at the origin and the base lies in the plane $z = h$; that is the symmetry axis coincides with the z axis. If ϕ is the angle between the z axis and the surface of the cone, then $\cot\phi = h/a$ and the equation of the cone is $z = \cot\phi\sqrt{x^2+y^2}$. Thus, the object in question is the surface (graph) $z = g(x,y) = (h/a)\sqrt{x^2+y^2}$ over the region D: $x^2+y^2 \le a^2$. To evaluate the needed surface integral, the area transformation law $dS = J\,dA$ should be established. One has $g'_x = (hx/a)(x^2+y^2)^{-1/2}$ and $g'_x = (hy/a)(x^2+y^2)^{-1/2}$ so that

$$J = \sqrt{1+(g'_x)^2+(g'_y)^2} = \sqrt{1+(h/a)^2} = \frac{\sqrt{h^2+a^2}}{a}.$$

The moment of inertia about the z axis is

$$\begin{aligned} I_z &= \iint_S (x^2+y^2)\sigma\, dS = k\iint_D (x^2+y^2)J\, dA \\ &= kJ\int_0^{2\pi} d\theta \int_0^a r^3\, dr = \frac{\pi k}{2}a^3\sqrt{h^2+a^2} \end{aligned}$$

□

Example 14.41. *Find the moment of inertia of a homogeneous ball of radius a and mass m about its diameter.*

Solution: Set up the coordinate system so that the origin is at the center of the ball. Then the moment of inertia about the z axis has to be evaluated. Since the ball is homogeneous, its mass density is constant, $\sigma = m/V$, where $V = 4\pi a^3/3$ is the volume of the ball. One has

$$I_z = \iiint_E (x^2+y^2)\sigma\, dV = \frac{3m}{4\pi a^3}\int_0^{2\pi}\int_0^{\pi}\int_0^{a}(\rho\sin\phi)^2\rho^2\sin\phi\, d\rho\, d\phi\, d\theta$$
$$= \frac{3}{10}ma^2\int_0^{\pi}\sin^3\phi\, d\phi = \frac{3}{10}ma^2\int_{-1}^{1}(1-u^2)\, du = \frac{2}{5}ma^2,$$

where the substitution $u = \cos\phi$ has been made to evaluate the integral. It is noteworthy that the problem admits a smarter solution by noting that $I_z = I_x = I_y$ owing to the rotational symmetry of the mass distribution. By the identity $I_z = (I_x + I_y + I_z)/3$, the triple integral can be simplified:

$$I_z = \frac{1}{3}\sigma\iiint_E 2(x^2+y^2+z^2)\, dV = \frac{1}{3}\sigma 8\pi\int_0^a \rho^4\, d\rho = \frac{2}{5}ma^2$$

□

109.3. Study Problems.

Problem 14.6. *Find the center of mass of the shell described in Example 14.40.*

Solution: By the symmetry of the mass distribution about the axis of the conic shell, the center of mass must be on that axis. Using the algebraic description of a shell given in Example 14.41, the total mass of the shell is

$$m = \iint_S \sigma\, dS = k\iint_S dS = kJ\iint_D dA = kJA(D) = \pi ka\sqrt{h^2+a^2}.$$

The moment about the xy plane is

$$M_{xy} = \iint_S z\sigma\, dS = k\iint_D (h/a)\sqrt{x^2+y^2}J\, dA = \frac{kJh}{a}\iint_D \sqrt{x^2+y^2}dA$$
$$= \frac{kJh}{a}\int_0^{2\pi}\int_0^a r^2\, dr\, d\theta = \frac{2\pi kha}{3}\sqrt{h^2+a^2}.$$

Thus, the center of mass is at the distance $z_c = M_{xy}/m = 2h/3$ from the tip of the cone. □

Problem 14.7. (Parallel Axis Theorem).
Let I_γ be the moment of inertia of an extended object about an axis γ and let γ_c be a parallel axis through the center of mass of the object. Prove that

$$I_\gamma = I_{\gamma_c} + mR_c^2,$$

where R_c is the distance between the axis γ and the center of mass, and m is the total mass.

Solution: Choose the coordinate system so that the axis γ goes through the origin. Let it be parallel to a unit vector $\hat{\mathbf{u}}$. The difference $I_\gamma - I_{\gamma_c}$ is to be investigated. If $\mathbf{r}_c$ is the position vector of the center of mass, then the axis γ_c is obtained from γ by parallel transport of the latter along the vector $\mathbf{r}_c$. Therefore, the distance $R^2_{\gamma_c}(\mathbf{r})$ is obtained from $R^2_\gamma(\mathbf{r})$ (see (14.34)) by changing the position vector $\mathbf{r}$ in the latter to the position vector relative to the center of mass, $\mathbf{r} - \mathbf{r}_c$. In particular, $R^2_\gamma(\mathbf{r}_c) = R_c^2$ by the definition of R_γ. Hence,

$$\begin{aligned} R^2_\gamma(\mathbf{r}) - R^2_{\gamma_c}(\mathbf{r}) &= R^2_\gamma(\mathbf{r}) - R^2_\gamma(\mathbf{r} - \mathbf{r}_c) \\ &= 2\mathbf{r_c}\cdot\mathbf{r} - \mathbf{r}_c^2 - (\hat{\mathbf{u}}\cdot\mathbf{r}_c)(2\hat{\mathbf{u}}\cdot\mathbf{r} - \hat{\mathbf{u}}\cdot\mathbf{r}_c) \\ &= \mathbf{r}_c^2 - (\hat{\mathbf{u}}\cdot\mathbf{r}_c)^2 + 2\mathbf{r}_c\cdot(\mathbf{r}-\mathbf{r}_c) - 2(\hat{\mathbf{u}}\cdot\mathbf{r}_c)\hat{\mathbf{u}}\cdot(\mathbf{r}-\mathbf{r}_c) \\ &= R_c^2 - 2\mathbf{a}\cdot(\mathbf{r}-\mathbf{r}_c), \end{aligned}$$

where $\mathbf{a} = \mathbf{r}_c - (\hat{\mathbf{u}}\cdot\mathbf{r}_c)\hat{\mathbf{u}}$. Therefore,

$$\begin{aligned} I_\gamma - I_{\gamma_c} &= \iiint_E \Big(R^2_\gamma(\mathbf{r}) - R^2_{\gamma_c}(\mathbf{r})\Big)\sigma(\mathbf{r})\,dV \\ &= R_c^2 \iiint_E \sigma(\mathbf{r})\,dV - 2\mathbf{a}\ \cdot \iiint_E (\mathbf{r}-\mathbf{r}_c)\sigma(\mathbf{r})\,dV = R_c^2 m, \end{aligned}$$

where the second integral vanishes by the definition of the center of mass. □

Problem 14.8. *Find the moment of inertia of a homogeneous ball of radius a and mass m about an axis that is at a distance R from the ball center.*

Solution: The center of mass of the ball coincides with its center because the mass distribution is invariant under rotations about the center. The moment of inertia of the ball about its diameter is $I_{\gamma_c} = (2/5)ma^2$ by Example 14.41. By the parallel axis theorem, for any axis γ at a distance R from the center of mass, $I_\gamma = I_{\gamma_c} + mR^2 = m(R^2 + 2a^2/5)$. □

CHAPTER 15

Vector Calculus

110. Line Integrals of a Vector Field

110.1. Vector Fields. Consider an air flow in the atmosphere. The air velocity varies from point to point. In order to describe the motion of the air, the air velocity must be defined as a function of position, which means that a velocity *vector* has to be assigned to every point in space. In other words, in contrast to ordinary functions, the air velocity is a *vector-valued* function of the position vector in space.

DEFINITION 15.1. (Vector Field).
Let E be a subset in space. A vector field on E is a function $\mathbf{F}$ that assigns to each point $\mathbf{r} = (x, y, z)$ a vector $\mathbf{F}(\mathbf{r}) = (F_1(\mathbf{r}), F_2(\mathbf{r}), F_3(\mathbf{r}))$. The functions F_1, F_2, and F_3 are called the components of the vector field $\mathbf{F}$.

A vector field is *continuous* if its components are continuous. A vector field is *differentiable* if its components are differentiable.

A simple example of a vector field is the gradient of a function, $\mathbf{F}(\mathbf{r}) = \nabla f(\mathbf{r})$. The components of this vector field are the first-order partial derivatives:

$$\mathbf{F}(\mathbf{r}) = \nabla f(\mathbf{r}) \quad \Longleftrightarrow \quad F_1(\mathbf{r}) = f'_x(\mathbf{r}), \quad F_2(\mathbf{r}) = f'_y(\mathbf{r}), \quad F_3(\mathbf{r}) = f'_z(\mathbf{r}).$$

Many physical quantities are described by vector fields. Electric and magnetic fields are vector fields. Light waves, radio waves, TV waves, and waves used in cell phone communications are all electromagnetic waves that are alternating electromagnetic fields. The propagation of electromagnetic waves in space is described by differential equations that relate electromagnetic fields at each point in space and each moment of time to a distribution of electric charges and currents (e.g., antennas). The gravitational force looks constant near the surface of the Earth, but on the scale of the solar system this is not so. The gravitational force exerted by a planet of mass M on a spacecraft of mass m depends on the position of the spacecraft relative to the planet center according to Newton's law of gravity:

$$\mathbf{F}(\mathbf{r}) = -\frac{GMm}{r^3}\,\mathbf{r} = \Big(-GMm\frac{x}{r^3},\ -GMm\frac{y}{r^3},\ -GMm\frac{z}{r^3}\Big),$$

where G is Newton's gravitational constant, $\mathbf{r}$ is the position vector relative to the planet center, and $r = \|\mathbf{r}\|$ is its length (the distance between the planet center and the spacecraft). The force is proportional to the position vector and hence parallel to it for each point in space. The minus sign indicates that $\mathbf{F}$ is directed opposite to $\mathbf{r}$, that is, the force is *attractive*; the gravitational force pulls toward its source (the planet). The magnitude $\|\mathbf{F}\| = GMmr^{-2}$ decreases with increasing distance r. So the gravitational vector field can be visualized by plotting vectors of length $\|\mathbf{F}\|$ at each point in space pointing toward the origin. The magnitudes of these vectors become smaller for points farther away from the origin. This observation leads to the concept of *flow lines* of a vector field.

110.2. Flow Lines of a Vector Field.

DEFINITION 15.2. (Flow Lines of a Vector Field).
The flow line of a vector field $\mathbf{F}$ *is a curve in space such that, at any point* $\mathbf{r}$*, the vector field* $\mathbf{F}(\mathbf{r})$ *is tangent to it.*

The direction of $\mathbf{F}$ defines the *orientation* of flow lines; that is, the direction of a tangent vector $\mathbf{F}$ is shown by arrows on the flow lines. For example, the flow lines of the planet's gravitational field are straight lines oriented toward the center of the planet. Flow lines of a gradient vector field $\mathbf{F} = \nabla f$ are normal to level surfaces of f and oriented in the direction in which f increases (most rapidly). They are the curves of steepest ascent of the function f. Flow lines of the air velocity vector field are often shown in weather forecasts to indicate the wind direction over large areas. For example, flow lines of the air velocity of a hurricane would look like closed loops around the eye of the hurricane.

The qualitative behavior of flow lines may be understood by plotting vectors $\mathbf{F}$ at several points $\mathbf{r}_i$ and sketching curves through them so that the vectors $\mathbf{F}_i = \mathbf{F}(\mathbf{r}_i)$ are tangent to the curves. Finding the exact shape of the flow lines requires solving differential equations. If $\mathbf{r} = \mathbf{r}(t)$ is a parametric equation of a flow line, then $\mathbf{r}'(t)$ is parallel to $\mathbf{F}(\mathbf{r}(t))$. So the derivative $\mathbf{r}'(t)$ must be proportional to $\mathbf{F}(\mathbf{r}(t))$, which defines a system of differential equations for the components of the vector function $\mathbf{r}(t)$, for example, $\mathbf{r}'(t) = \mathbf{F}(\mathbf{r}(t))$.

EXAMPLE 15.1. *Analyze flow lines of the planar vector field* $\mathbf{F} = (-y, x, 0)$.

SOLUTION: By noting that $\mathbf{F} \cdot \mathbf{r} = 0$, it is concluded that at any point $\mathbf{F}$ is perpendicular to the position vector $\mathbf{r} = (x, y, 0)$ in the plane. So flow lines are curves whose tangent vector is perpendicular to the position vector. If $\mathbf{r} = \mathbf{r}(t)$ is a parametric equation of such a curve, then $\mathbf{r}(t) \cdot \mathbf{r}'(t) = 0$ or $(d/dt)\mathbf{r}^2(t) = 0$ and hence $\mathbf{r}^2(t) = \text{const}$, which is a circle centered at the origin. So flow lines are concentric circles. At the point $(1, 0, 0)$, the vector field is directed along the y axis: $\mathbf{F}(1, 0, 0) = (0, 1, 0) = \hat{\mathbf{e}}_2$. Therefore, the flow lines are oriented counterclockwise. The magnitude $\|\mathbf{F}\| = \sqrt{x^2 + y^2}$ remains constant on each circle and increases with increasing circle radius. □

110.3. Line Integral of a Vector Field. The work done by a constant force $\mathbf{F}$ in moving an object along a straight line is given by

$$W = \mathbf{F} \cdot \mathbf{d},$$

where $\mathbf{d}$ is the displacement vector. Suppose that the force varies in space and the displacement trajectory is no longer a straight line. What is the work done by the force? This question leads to the concept of the line integral of a vector field.

Let C be a smooth curve that goes from a point $\mathbf{r}_a$ to a point $\mathbf{r}_b$ and has a length L. Consider a partition of C by segments C_i, $i = 1, 2, ..., N$, of length $\Delta s = L/N$. Since the curve is smooth, each segment can be approximated by a straight line segment of length Δs oriented along the unit tangent vector $\hat{\mathbf{T}}(\mathbf{r}_i^*)$ at a sample point $\mathbf{r}_i^* \in C_i$. The work along the segment C_i can therefore be approximated by $\Delta W_i = \mathbf{F}(\mathbf{r}_i^*) \cdot \hat{\mathbf{T}}(\mathbf{r}_i^*)\, \Delta s$ so that the total work is approximately the sum $W = \Delta W_1 + \Delta W_2 + \cdots + \Delta W_N$. The actual work should not depend on the choice of sample points. This problem is resolved by the usual trick of integral calculus, that is, by refining a partition, finding the low and upper sums, and taking their limits. If these limits exist and coincide, the limiting value is the sought-for work. The technicalities involved may be spared by noting that $\Delta W_i = f(\mathbf{r}_i^*)\, \Delta s$, where $f(\mathbf{r}) = \mathbf{F}(\mathbf{r}) \cdot \hat{\mathbf{T}}(\mathbf{r})$ and $\hat{\mathbf{T}}(\mathbf{r})$ denotes the unit tangent vector at a point $\mathbf{r} \in C$. The approximate total work appears to be a Riemann sum for the line integral of f along C. So the *work is the line integral with respect to the arc length of the tangential component* $\mathbf{F} \cdot \hat{\mathbf{T}}$ *of the force.*

Definition 15.3. (Line Integral of a Vector Field).
The line integral of a vector field $\mathbf{F}$ *along a smooth curve* C *is*

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

where $\hat{\mathbf{T}}$ *is the unit tangent vector to* C*, provided the tangential component* $\mathbf{F} \cdot \hat{\mathbf{T}}$ *of the vector field is integrable on* C*.*

The integrability of $\mathbf{F}\cdot\hat{\mathbf{T}}$ is defined in the sense of line integrals for ordinary functions (see Definition 14.15)

110.4. Evaluation of Line Integrals of Vector Fields. The line integral of a vector field is evaluated in much the same way as the line integral of a function.

THEOREM 15.1. (Evaluation of Line Integrals).
Let $\mathbf{F} = (F_1, F_2, F_3)$ be a continuous vector field on E and let C be a smooth curve C in E that originates from a point $\mathbf{r}_a$ and terminates at a point $\mathbf{r}_b$. Suppose that $\mathbf{r}(t) = (x(t), y(t), z(t))$, $t \in [a, b]$, is a vector function that traces out the curve C so that $\mathbf{r}(a) = \mathbf{r}_a$ and $\mathbf{r}(b) = \mathbf{r}_b$. Then

$$\int_C \mathbf{F}(\mathbf{r})\cdot d\mathbf{r} = \int_C \mathbf{F}\cdot\hat{\mathbf{T}}\,ds = \int_a^b \mathbf{F}(\mathbf{r}(t))\cdot\mathbf{r}'(t)\,dt$$

$$= \int_a^b \Big(F_1(\mathbf{r}(t))x'(t) + F_2(\mathbf{r}(t))y'(t) + F_3(\mathbf{r}(t))z'(t)\Big)\,dt. \tag{15.35}$$

PROOF. The unit tangent vector reads $\hat{\mathbf{T}} = \mathbf{r}'/\|\mathbf{r}'\|$ and $ds = \|\mathbf{r}'\|\,dt$. Therefore, $\hat{\mathbf{T}}\,ds = \mathbf{r}'(t)\,dt$. The function $f(t) = \mathbf{F}(\mathbf{r}(t))\cdot\mathbf{r}'(t)$ is continuous on $[a, b]$, and the conclusion of the theorem follows from Theorem 14.15. □

Equation (15.35) also holds if C is piecewise smooth and $\mathbf{F}$ has a finite number of bounded jump discontinuities along C much like in the case of the line integral of ordinary functions.

In contrast to the line integral of ordinary functions, the line integral of a vector field depends on the orientation of C. The orientation of C is fixed by the conditions $\mathbf{r}(a) = \mathbf{r}_a$ and $\mathbf{r}(b) = \mathbf{r}_b$ for a vector function $\mathbf{r}(t)$, where $a \le t \le b$, provided the vector function traces out the curve only once. If $\mathbf{r}(t)$ traces out C from $\mathbf{r}_b$ to $\mathbf{r}_a$, then the orientation is reversed, and such a curve is denoted by $-C$. The line integral changes its sign when the orientation of the curve is reversed:

$$\int_{-C} \mathbf{F}\cdot d\mathbf{r} = -\int_C \mathbf{F}\cdot d\mathbf{r} \tag{15.36}$$

because the direction of the derivative $\mathbf{r}'(t)$ is reversed for all t.

The evaluation of a line integral includes the following steps:

Step 1. If the curve C is defined as a point set in space by some geometrical means, then find its parametric equations $\mathbf{r} = \mathbf{r}(t)$ that agree with the orientation of C. Here it is useful to remember that, if $\mathbf{r}(t)$ corresponds to the opposite orientation, then it can still be used according to (15.36).

Step 2. Restrict the range of t to an interval $[a, b]$ so that C is traced

out only once by $\mathbf{r}(t)$.
Step 3. Substitute $\mathbf{r} = \mathbf{r}(t)$ into the arguments of $\mathbf{F}$ to obtain the values of $\mathbf{F}$ on C and calculate the derivative $\mathbf{r}'(t)$ and the dot product $\mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t)$.
Step 4. Evaluate the (ordinary) integral (15.35).

Remark. If C is piecewise smooth (e.g., the union of smooth curves C_1 and C_2), then the additivity of the integral should be used:

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

Remark. If a curve is *defined* as a vector function on $[a, b]$ (see Section 79.3), then $\mathbf{r}(t)$ may trace its range (as a point set in space) or some parts of it several times as t changes from a to b. If two vector functions $\mathbf{r}_1(t)$ and $\mathbf{r}_2(t)$ on $[a, b]$ have the same range but $\mathbf{r}_1(t) \neq \mathbf{r}_2(t)$ for some values of $t \in [a, b]$, they are considered *different* curves. For example, $\mathbf{r}_1 = (\cos t, \sin t, 0)$ and $\mathbf{r}_1(t) = (\cos(2t), \sin(2t), 0)$ have the same range on $[0, 2\pi]$, which is the circle of unit radius, but $\mathbf{r}_2(t)$ traces out the circle twice. Note that these curves have different lengths, $L_1 = 2\pi$ and $L_2 = 4\pi$. So the line integral (15.35) may be different for two curves *defined* as two vector functions, even though the ranges of these functions coincide as point sets in space. The curve defined by a vector is much like the trajectory of a particle that can pass through the same points multiple times.

EXAMPLE 15.2. *Evaluate the line integral of* $\mathbf{F} = (-y, x, z^2)$ *along a closed contour* C *that consists of one turn of a helix of radius* R*, which begins at the point* $\mathbf{r}_a = (R, 0, 0)$ *and ends at the point* $\mathbf{r}_b = (R, 0, 2\pi h)$*, and a straight line segment from* $\mathbf{r}_b$ *to* $\mathbf{r}_a$*.*

SOLUTION: Let C_1 be one turn of the helix and let C_2 be the straight line segment. Two line integrals have to be evaluated. The parametric equations of the helix are $\mathbf{r}(t) = (R\cos t, R\sin t, ht)$ so that $\mathbf{r}(0) = (R, 0, 0)$ and $\mathbf{r}(2\pi) = (R, 0, h)$ as required by the orientation of C_1. The range of t has to be restricted to $[0, 2\pi]$. Then $\mathbf{r}'(t) = (-R\sin t, R\cos t, h)$. Therefore,

$$\begin{aligned}
\mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) &= (-R\sin t, R\cos t, h^2t^2) \cdot (-R\sin t, R\cos t, h) \\
&= R^2 + h^3t^2, \\
\int_{C_1} \mathbf{F} \cdot d\mathbf{r} &= \int_0^{2\pi} \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t)\, dt = \int_0^{2\pi} (R^2 + h^3t^2)\, dt \\
&= 2\pi R^2 + \frac{(2\pi h)^3}{3}.
\end{aligned}$$

The parametric equations of the line through two points $\mathbf{r}_a$ and $\mathbf{r}_b$ are $\mathbf{r}(t) = \mathbf{r}_a + \mathbf{v}t$, where $\mathbf{v} = \mathbf{r}_b - \mathbf{r}_a$ is the vector parallel to the line, or in the components $\mathbf{r} = (R, 0, 0) + t(0, 0, 2\pi h) = (R, 0, 2\pi ht)$. Then $\mathbf{r}(0) = \mathbf{r}_a$ and $\mathbf{r}(1) = \mathbf{r}_b$ so that the orientation is reversed if $t \in [0, 1]$. The found parametric equations describe the curve $-C_2$. One has $\mathbf{r}'(t) = (0, 0, 2\pi h)$ and hence

$$\mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = (0, R, (2\pi h)^2 t^2) \cdot (0, 0, 2\pi h) = (2\pi h)^3 t^2,$$

$$\int_{C_2} \mathbf{F} \cdot d\mathbf{r} = -\int_{-C_2} \mathbf{F} \cdot d\mathbf{r} = -(2\pi h)^3 \int_0^1 t^2\, dt = -\frac{(2\pi h)^3}{3}.$$

The line integral along C is the sum of these integrals, which is equal to $2\pi R^2$. □

111. Fundamental Theorem for Line Integrals

Recall the fundamental theorem of calculus, which asserts that, if the derivative $f'(x)$ is continuous on an interval $[a, b]$, then

$$\int_a^b f'(x)\, dx = f(b) - f(a).$$

It appears that there is an analog of this theorem for line integrals.

111.1. Conservative Vector Fields.

Definition 15.4. (Conservative Vector Field and Its Potential). *A vector field* $\mathbf{F}$ *in a region* E *is said to be conservative if there is a function* f*, called a* potential *of* $\mathbf{F}$*, such that* $\mathbf{F} = \nabla f$ *in* E*.*

Conservative vector fields play a significant role in many practical applications. It has been proved earlier (see Study Problem 13.9) that if a particle moves along a trajectory $\mathbf{r} = \mathbf{r}(t)$ under the force $\mathbf{F} = -\nabla U$, then its energy $E = m\mathbf{v}^2/2 + U(\mathbf{r})$, where $\mathbf{v} = \mathbf{r}'$ is the velocity, is conserved along the trajectory, $dE/dt = 0$. In particular, Newton's gravitational force is conservative, $\mathbf{F} = -\nabla U$, where $U(\mathbf{r}) = -GMm\|\mathbf{r}\|^{-1}$. A static electric field (the Coulomb field) created by a distribution of static electric charges is also conservative. Conservative vector fields have a remarkable property.

Theorem 15.2. (Fundamental Theorem for Line Integrals). *Let* C *be a smooth curve in a region* E *with initial and terminal points* $\mathbf{r}_a$ *and* $\mathbf{r}_b$*, respectively. Let* f *be a function on* E *whose gradient* ∇f *is continuous on* C*. Then*

$$\int_C \nabla f \cdot d\mathbf{r} = f(\mathbf{r}_b) - f(\mathbf{r}_a). \tag{15.37}$$

PROOF. Let $\mathbf{r} = \mathbf{r}(t)$, $t \in [a, b]$, be the parametric equations of C such that $\mathbf{r}(a) = \mathbf{r}_a$ and $\mathbf{r}(b) = \mathbf{r}_b$. Then, by (15.35) and the chain rule,

$$\int_C \nabla f \cdot d\mathbf{r} = \int_a^b (f'_x x' + f'_y y' + f'_z z')\, dt = \int_a^b \frac{d}{dt} f(\mathbf{r}(t))\, dt = f(\mathbf{r}_b) - f(\mathbf{r}_a).$$

The latter equality holds by the fundamental theorem of calculus and the continuity of the first-order derivatives of f and $\mathbf{r}'(t)$ for a smooth curve. □

111.2. Path Independence of Line Integrals.

DEFINITION 15.5. (Path Independence of Line Integrals).
A continuous vector field $\mathbf{F}$ *has path-independent line integrals if*

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

for any two simple, piecewise-smooth curves in the domain of $\mathbf{F}$ *with the same endpoints.*

Recall that a curve is simple if it does not intersect itself (see Section 79.3). An important consequence of the fundamental theorem for line integrals is that the work done by a conservative force, $\mathbf{F} = \nabla f$, is *path-independent*. So a criterion for a vector field to be conservative would be advantageous for evaluating line integrals because for a conservative vector field a curve may be deformed at convenience without changing the value of the integral.

THEOREM 15.3. (Path-Independent Property).
Let $\mathbf{F}$ *be a continuous vector field on an open region* E*. Then* $\mathbf{F}$ *has path-independent line integrals if and only if its line integral vanishes along every piecewise-smooth, simple, closed curve* C *in* E*. In that case, there exists a function* f *such that* $\mathbf{F} = \nabla f$*:*

$$\mathbf{F} = \nabla f \quad \Longleftrightarrow \quad \oint_C \mathbf{F} \cdot d\mathbf{r} = 0.$$

The symbol $\oint_C$ is often used to denote line integrals along a closed path.

PROOF. Pick a point $\mathbf{r}_0$ in E and consider any smooth curve C from $\mathbf{r}_0$ to a point $\mathbf{r} = (x, y, z) \in E$. The idea is to prove that the function

$$f(\mathbf{r}) = \int_C \mathbf{F} \cdot d\mathbf{r} \qquad (15.38)$$

is a potential of $\mathbf{F}$, that is, to prove that $\nabla f = \mathbf{F}$ under the condition that the line integral of $\mathbf{F}$ vanishes for every closed curve in E. This "guess" for f is motivated by the fundamental theorem for line integrals (15.37), where $\mathbf{r}_b$ is replaced by a generic point $\mathbf{r} \in E$. The potential is defined up to an additive constant ($\nabla(f + \text{const}) = \nabla f$) so the choice of a fixed point $\mathbf{r}_0$ is irrelevant. First, note that the value of f is independent of the choice of C. Consider two such curves C_1 and C_2. Then the union of C_1 and $-C_2$ (the curve C_2 whose orientation is reversed) is a closed curve, and the line integral along it vanishes by the hypothesis. On the other hand, this line integral is the sum of line integrals along C_1 and $-C_2$. By the property (15.36), the line integrals along C_1 and C_2 coincide. To calculate the derivative $f'_x(\mathbf{r}) = \lim_{h\to 0}(f(\mathbf{r} + h\hat{\mathbf{e}}_1) - f(\mathbf{r}))/h$, where $\hat{\mathbf{e}}_1 = (1, 0, 0)$, let us express the difference $f(\mathbf{r} + h\hat{\mathbf{e}}_1) - f(\mathbf{r})$ via a line integral. Note that E is open, which means that a ball of sufficiently small radius centered at any point in E is contained in E (i.e., $\mathbf{r} + h\hat{\mathbf{e}}_1 \in E$ for a sufficiently small h). Since the value of f is path-independent, for the point $\mathbf{r} + h\hat{\mathbf{e}}_1$, the curve can be chosen so that it goes from $\mathbf{r}_0$ to $\mathbf{r}$ and then from $\mathbf{r}$ to $\mathbf{r} + h\hat{\mathbf{e}}_1$ along the straight line segment. Denote the latter by ΔC. Therefore,

$$f(\mathbf{r} + h\hat{\mathbf{e}}_1) - f(\mathbf{r}) = \int_{\Delta C} \mathbf{F} \cdot d\mathbf{r}$$

because the line integral of $\mathbf{F}$ from $\mathbf{r}_0$ to $\mathbf{r}$ is path-independent. A vector function that traces out ΔC is $\mathbf{r}(t) = (t, y, z)$ if $x \le t \le x + h$. Therefore, $\mathbf{r}'(t) = \hat{\mathbf{e}}_1$ and $\mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = F_1(t, y, z)$. Thus,

$$\begin{aligned} f'_x(\mathbf{r}) &= \lim_{h\to 0} \frac{1}{h} \int_x^{x+h} F_1(t, y, z)\,dt = \lim_{h\to 0} \frac{1}{h} \Big(\int_a^{x+h} - \int_a^x \Big) F_1(t, y, z)\,dt \\ &= \frac{\partial}{\partial x} \int_a^x F_1(t, y, z)\,dt = F_1(x, y, z) = F_1(\mathbf{r}) \end{aligned}$$

by the continuity of F_1. The equalities $f'_y = F_2$ and $f'_z = F_3$ are established similarly. The details are omitted. □

Although the path independence property does provide a necessary and sufficient condition for a vector field to be conservative, it is rather impractical to verify (one cannot evaluate line integrals along every closed curve!). A more feasible and practical criterion is needed, which is established next. It is worth noting that (15.38) gives a practical method of finding a potential if the vector field is found to be conservative (see the study problems at the end of this section).

111.3. The Curl of a Vector Field. According to the rules of vector algebra, the product of a vector $\mathbf{a} = (a_1, a_2, a_3)$ and a number s is defined by $s\mathbf{a} = (sa_1, sa_2, sa_3)$. By analogy, the gradient ∇f can be viewed as the product of the vector $\nabla = (\partial/\partial x,\ \partial/\partial y,\ \partial/\partial z)$ and a scalar f:

$$\nabla f = \Big(\frac{\partial}{\partial x},\ \frac{\partial}{\partial y},\ \frac{\partial}{\partial z}\Big) f = \Big(\frac{\partial f}{\partial x},\ \frac{\partial f}{\partial y},\ \frac{\partial f}{\partial z}\Big).$$

The components of ∇ are not ordinary numbers, but rather they are *operators* (i.e., symbols standing for a specified operation that has to be carried out). For example, $(\partial/\partial x)f$ means that the operator $\partial/\partial x$ is applied to a function f and the result of its action on f is the partial derivative of f with respect to x. The directional derivative $D_{\mathbf{u}}f$ can be viewed as the result of the action of the operator $D_{\mathbf{u}} = \hat{\mathbf{u}} \cdot \nabla = u_1(\partial/\partial x) + u_2(\partial/\partial y) + u_3(\partial/\partial z)$ on a function f. In what follows, the formal vector ∇ is viewed as an operator whose action obeys the rules of vector algebra.

Definition 15.6. (Curl of a Vector Field). *The curl of a differentiable vector field* $\mathbf{F}$ *is*

$$\operatorname{curl} \mathbf{F} = \nabla \times \mathbf{F}.$$

The curl of a vector field is also a vector field whose components can be computed according to the definition of the cross product:

$$\begin{aligned}\nabla \times \mathbf{F} &= \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_1 & F_2 & F_3 \end{pmatrix} \\ &= \Big(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\Big)\hat{\mathbf{e}}_1 + \Big(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}\Big)\hat{\mathbf{e}}_2 + \Big(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\Big)\hat{\mathbf{e}}_3.\end{aligned}$$

When calculating the components of the curl, the product of a component of ∇ and a component of $\mathbf{F}$ means that the component of ∇ operates on the component of $\mathbf{F}$, producing the corresponding partial derivative. Of course, it is assumed that partial derivatives of components of $\mathbf{F}$ exist in order for the curl to exist.

Example 15.3. *Find the curl of the vector field* $\mathbf{F} = (yz, xyz, x^2)$.

Solution:

$$\begin{aligned}\nabla \times \mathbf{F} &= \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ yz & xyz & x^2 \end{pmatrix} \\ &= \Big((x^2)'_y - (xyz)'_z,\ -(x^2)'_x + (yz)'_z,\ (xyz)'_x - (yz)'_y\Big) \\ &= (-xy,\ y - 2x,\ yz - z).\end{aligned}$$

□

The geometrical significance of the curl of a vector field will be discussed later. Here the curl is used to formulate sufficient conditions for a vector field to be conservative.

111.4. On the Use of the Operator ∇. The rules of vector algebra are useful to simplify algebraic operations involving the operator ∇. For example,

$$\operatorname{curl}\nabla f = \nabla \times (\nabla f) = (\nabla \times \nabla) f = \mathbf{0}$$

because the cross product of a vector with itself vanishes. However, this formal algebraic manipulation should be adopted with precaution because it contains a tacit assumption that the action of the components of $\nabla \times \nabla$ on f vanishes. The latter imposes conditions on the class of functions for which such formal algebraic manipulations are justified. Indeed, according to the definition,

$$\nabla \times \nabla f = \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ f'_x & f'_y & f'_z \end{pmatrix} = (f''_{zy} - f''_{yz},\ f''_{zx} - f''_{xz},\ f''_{xy} - f''_{yx}).$$

This vector vanishes, provided the order of differentiation does not matter (i.e., Clairaut's theorem holds for f). Thus, *the rules of vector algebra can be used to simplify the action of an operator involving ∇ if the partial derivatives of a function on which this operator acts are continuous up to the order determined by that action.*

111.5. Test for a Vector Field to Be Conservative. A conservative, continuously differentiable vector field in an region E has been shown to have the vanishing curl:

$$\mathbf{F} = \nabla f \quad \Longrightarrow \quad \operatorname{curl}\mathbf{F} = \mathbf{0}.$$

Unfortunately, the converse is *not* true in general. In other words, the vanishing of the curl of a vector field does *not* guarantee that the vector field is conservative. The converse is true only if the region in which the curl vanishes belongs to a special class. A region E is said to be *connected* if any two points in it can be connected by a path that lies in E. In other words, a connected region cannot be represented as the union of two or more non-intersecting (disjoint) regions.

DEFINITION 15.7. (Simply Connected Region).
A connected region E is simply connected if every simple closed curve in E can be continuously shrunk to a point in E while remaining in E throughout the deformation.

Naturally, the entire Euclidean space is simply connected. A ball in space is also simply connected. If E is the region outside a ball, then it is also simply connected. However, if E is obtained by removing a line (or a cylinder) from the entire space, then E is not simply connected. Indeed, take a circle such that the line pierces through the disk bounded by the circle. There is no way this circle can be continuously contracted to a point of E without crossing the line. A solid torus is not simply connected. (Explain why!) A simply connected region D in a plane cannot have "holes" in it.

Theorem 15.4. (Test for a Vector Field to Be Conservative). *Suppose that a vector field* $\mathbf{F}$ *is continuously differentiable on a simply connected open region* E*. Then* $\mathbf{F}$ *is conservative in* E *if and only if its curl vanishes for all points of* E*:*

$$\operatorname{curl}\mathbf{F} = \mathbf{0} \text{ on simply connected } E \quad \Longleftrightarrow \quad \mathbf{F} = \nabla f \text{ on } E.$$

This theorem follows from Stokes' theorem discussed later and has two useful consequences. First, the test for the path independence of line integrals:

$$\operatorname{curl}\mathbf{F} = \mathbf{0} \text{ on simply connected } E \quad \Longleftrightarrow \quad \int_{C_1} \mathbf{F}\cdot d\mathbf{r} = \int_{C_2} \mathbf{F}\cdot d\mathbf{r}$$

for any two paths C_1 and C_2 in E originating from a point $\mathbf{r}_a \in E$ and terminating at another point $\mathbf{r}_b \in E$. Second, the test for vanishing line integrals along closed paths:

$$\operatorname{curl}\mathbf{F} = \mathbf{0} \text{ on simply connected } E \quad \Longleftrightarrow \quad \oint_C \mathbf{F}\cdot d\mathbf{r} = 0,$$

where C is a closed curve in E. The condition that E is simply connected is crucial here. Even if $\operatorname{curl}\mathbf{F} = \mathbf{0}$, but E is not simply connected, the line integral of $\mathbf{F}$ may still depend on the path and the line integral along a closed path may not vanish! An example is given in one of the study problems at the end of this section.

Newton's gravitational force can be written as the gradient $\mathbf{F} = -\nabla U$, where $U(\mathbf{r}) = -GMm\|\mathbf{r}\|^{-1}$ everywhere except the origin. Therefore, its curl vanishes in E that is the entire space with one point removed; it is simply connected. Hence, the work done by the gravitational force is *independent* of the path traveled by the object and determined by the difference of values of its potential U at the initial and terminal points of the path.

Example 15.4. *Evaluate the line integral of the vector field* $\mathbf{F} = (F_1, F_2, F_3) = (yz,\ xz{+}z{+}2y,\ xy{+}y{+}2z)$ *along the path* C *that consists* *of straight line segments AB_1, B_1B_2, and B_2D, where the initial point is $A = (0,0,0)$, $B_1 = (2010, 2011, 2012)$, $B_2 = (102, 1102, 2102)$, and the terminal point is $D = (1,1,1)$.*

Solution: The path looks complicated enough to check whether **F** is conservative before evaluating the line integral using the parametric equations of C. First, note that the components of **F** are polynomials and hence continuously differentiable in the entire space. Therefore, if its curl vanishes, then **F** is conservative in the entire space as the entire space is simply connected:

$$\begin{aligned}\nabla\times\mathbf{F} &= \det\begin{pmatrix}\hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3\\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z}\\ F_1 & F_2 & F_3\end{pmatrix} = \det\begin{pmatrix}\hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3\\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z}\\ yz & xz+z+2y & xy+y+2z\end{pmatrix}\\ &= \Big((F_3)'_y-(F_2)'_z,\ -(F_3)'_x+(F_1)'_z,\ (F_2)'_x-(F_1)'_y\Big)\\ &= (x+1-(x+1),\ -y+y,\ z-z) = (0,0,0).\end{aligned}$$

Thus, **F** is conservative. Now there are two options to finish the problem.

Option 1. One can use the path independence of the line integral, which means that one can pick any other path C_1 connecting the initial point A and the terminal point D to evaluate the line integral in question. For example, a straight line segment connecting A and D is a simple enough to evaluate the line integral. Its parametric equations are $\mathbf{r} = \mathbf{r}(t) = (t,t,t)$, where $t \in [0,1]$. Therefore,

$$\mathbf{F}(\mathbf{r}(t))\cdot\mathbf{r}'(t) = (t^2,\ t^2+3t,\ t^2+3t)\cdot(1,\ 1,\ 1) = 3t^2+6t$$

and hence

$$\int_C \mathbf{F}\cdot d\mathbf{r} = \int_{C_1}\mathbf{F}\cdot d\mathbf{r} = \int_0^1 (3t^2+6t)\,dt = 4.$$

Option 2. The procedure of Section 89.1 may be used to find a potential f of **F** (see also the study problems at the end of this section for an alternative procedure). The line integral is then found by the fundamental theorem for line integrals. Put $\nabla f = \mathbf{F}$. Then the problem is reduced to finding f from its first-order partial derivatives (the existence of f has already been established). Following the procedure of Section 89.1,

$$f'_x = F_1 = yz \quad\Longrightarrow\quad f(x,y,z) = xyz + g(y,z),$$

where $g(y,z)$ is arbitrary. The substitution of f into the second equation $f'_y = F_2$ yields

$$xz + g'_y(y,z) = xz + z + 2y \quad \Longrightarrow \quad g(y,z) = y^2 + zy + h(z),$$

where $h(z)$ is arbitrary. The substitution of $f = xyz + y^2 + zy + h(z)$ into the third equation $f'_z = F_3$ yields

$$xy + y + h'(z) = xy + y + 2z \quad \Longrightarrow \quad h(z) = z^2 + c,$$

where c is a constant. Thus, $f(x,y,z) = xyz + yz + z^2 + y^2 + c$ and

$$\int_C \mathbf{F} \cdot d\mathbf{r} = f(1,1,1) - f(0,0,0) = 4$$

by the fundamental theorem for line integrals. □

111.6. Study Problems.

Problem 15.1. *Verify that*

$$\mathbf{F} = \nabla f = \Big(-\frac{y}{x^2+y^2},\ \frac{x}{x^2+y^2},\ 2z\Big), \qquad f(x,y,z) = \tan^{-1}(y/x) + z^2$$

and curl $\mathbf{F} = \mathbf{0}$ *in the domain of* $\mathbf{F}$. *Evaluate the line integral of* $\mathbf{F}$ *along the circular path* C*:* $x^2 + y^2 = R^2$ *in the plane* $z = a$*. The path is oriented counterclockwise as viewed from the top of the* z *axis. Does the result contradict to the fundamental theorem for line integrals? Explain.*

Solution: A straightforward differentiation of f shows that indeed $\nabla f = \mathbf{F}$ and therefore curl $\mathbf{F} = \mathbf{0}$ everywhere except the line $x = y = 0$ where $\mathbf{F}$ is not defined. The path C is traced out by $\mathbf{r}(t) = (R\cos t,\ R\sin t,\ a)$, where $t \in [0, 2\pi]$. Then $\mathbf{F}(\mathbf{r}(t)) = (-R^{-1}\sin t, R^{-1}\cos t, 2a)$ and $\mathbf{r}'(t) = (-R\sin t, R\cos t, 0)$. Therefore, $\mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = 1$ and

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = \int_0^{2\pi} dt = 2\pi.$$

So the integral over the closed contour does not vanish despite the fact that $\mathbf{F} = \nabla f$, which seems to be in conflict with the fundamental theorem for line integrals as by the latter the integral should have vanished.

Consider the values of f along the circle. By construction, $f(x,y,a) = \theta(x,y) + a^2$, where $\theta(x,y)$ is the polar angle in any plane $z = a$. It is 0 on the positive x axis and increases as the point moves about the origin. As the point arrives back to the positive x axis, the angle reaches the value 2π; that is, f is not really a function on the closed contour because it takes *two* values, 0 and 2π, at the same point on the positive x axis.

The only way to make f a function is to remove the half-plane $\theta = 0$ from the domain of f. Think of a cut in space along the half-plane. But in this case, any closed path that intersects the half-plane becomes nonclosed as it has two *distinct* endpoints on the opposite edges of the cut. If the fundamental theorem for line integrals is applied to such a path, then no contradiction arises because the values of f on the edges of the cut differ exactly by 2π in full accordance with the conclusion of the theorem.

Alternatively, the issue can be analyzed by studying whether $\mathbf{F}$ is conservative in its domain E. The vector field is defined everywhere in space except the line $x = y = 0$ (the z axis). So E is not simply connected. Therefore, the condition $\operatorname{curl} \mathbf{F} = \mathbf{0}$ is not sufficient to claim that the vector field is conservative on its domain. Indeed, the evaluated line integral along the closed path (which cannot be continuously contracted, staying within E, to a point in E) shows that the vector field cannot be conservative on E. If the half-plane $\theta = 0$ is removed from E, then $\mathbf{F}$ is conservative on this "reduced" region because the latter is simply connected. Naturally, the line integral along any closed path that does not cross the half-plane $\theta = 0$ (i.e., it lies within the reduced domain) vanishes. □

Problem 15.2. *Prove that if* $\mathbf{F} = (F_1, F_2, F_3)$ *is conservative, then its potential is*

$$f(x, y, z) = \int_{x_0}^{x} F_1(t, y_0, z_0)\, dt + \int_{y_0}^{y} F_2(x, t, z_0)\, dt + \int_{z_0}^{z} F_3(x, y, t)\, dt,$$

where (x_0, y_0, z_0) *is any point in the domain of* $\mathbf{F}$. *Use this equation to find a potential of* $\mathbf{F}$ *from Example 15.4.*

SOLUTION: In (15.38), take C that consists of three straight line segments, $(x_0, y_0, z_0) \to (x, y_0, z_0) \to (x, y, z_0) \to (x, y, z)$. The parametric equation of the first line C_1 is $\mathbf{r}(t) = (t, y_0, z_0)$, where $x_0 \le t \le x$. Therefore, $\mathbf{r}'(t) = (1, 0, 0)$ and $\mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = F_1(t, y_0, z_0)$. So the line integral of $\mathbf{F}$ along C_1 gives the first term in the above expression for f. Similarly, the second term is the line integral of $\mathbf{F}$ along the second line $\mathbf{r}(t) = (x, t, z_0)$, where $y_0 \le t \le y$, so that $\mathbf{r}'(t) = (0, 1, 0)$. The third term is the line integral of $\mathbf{F}$ along the third line $\mathbf{r}(t) = (x, y, t)$, where $z_0 \le t \le z$. In Example 15.4, it was established that $\mathbf{F} = (F_1, F_2, F_3) = (yz,\ xz + z + 2y,\ xy + y + 2z)$ is conservative. For simplicity, choose

$(x_0, y_0, z_0) = (0, 0, 0)$. Then

$$\begin{aligned} f(x,y,z) &= \int_0^x F_1(t,0,0)\,dt + \int_0^y F_2(x,t,0)\,dt + \int_0^z F_3(x,y,t)\,dt \\ &= 0 + y^2 + (xyz + yz + z^2) = xyz + yz + z^2 + y^2, \end{aligned}$$

which naturally coincides with f found by a different (longer) method. □

112. Green's Theorem

Green's theorem should be regarded as the counterpart of the fundamental theorem of calculus for the double integral.

Definition 15.8. (Orientation of Planar Closed Curves). *A simple closed curve C in a plane whose single traversal is counterclockwise (clockwise) is said to be positively (negatively) oriented.*

A simple closed curve divides the plane into two connected regions. If a planar region D is bounded by a simple closed curve, then the positively oriented boundary of D is denoted by the symbol ∂D.

Recall that a simple closed curve can be regarded as a continuous vector function $\mathbf{r}(t) = (x(t), y(t))$ on $[a, b]$ such that $\mathbf{r}(a) = \mathbf{r}(b)$ and, for any $t_1 \neq t_2$ in the open interval (a, b), $\mathbf{r}(t_1) \neq \mathbf{r}(t_2)$; that is, $\mathbf{r}(t)$ traces out C only once without self-intersection. A positive orientation means that $\mathbf{r}(t)$ traces out its range counterclockwise. For example, the vector functions $\mathbf{r}(t) = (\cos t, \sin t)$ and $\mathbf{r}(t) = (\cos t, -\sin t)$ on the interval $[0, 2\pi]$ define the positively and negatively oriented circles of unit radius, respectively.

Theorem 15.5. (Green's Theorem).
Let C be a positively oriented, piecewise-smooth, simple, closed curve in the plane and let D be the region bounded by $C = \partial D$. If the functions F_1 and F_2 have continuous partial derivatives on an open region that contains D, then

$$\iint_D \Big(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\Big) dA = \oint_{\partial D} F_1\,dx + F_2\,dy.$$

Just like the fundamental theorem of calculus, Green's theorem relates the derivatives of F_1 and F_2 in the integrand to the values of F_1 and F_2 on the boundary of the integration region. A proof of Green's theorem is rather involved. Here it is limited to the case when the region D is simple.

Proof (for simple regions). A simple region D admits two equivalent algebraic descriptions:

$$D = \{(x,y)\,|\,y_{\rm bot}(x) \le y \le y_{\rm top}(x)\,,\ x \in [a,b]\}, \tag{15.39}$$

$$D = \{(x,y)\,|\,x_{\rm bot}(y) \le x \le x_{\rm top}(y)\,,\ y \in [c,d]\}. \tag{15.40}$$

The idea of the proof is to establish the equalities

$$\oint_{\partial D} F_1\,dx = -\iint_D \frac{\partial F_1}{\partial y}dA\,, \qquad \oint_{\partial D} F_2\,dy = \iint_D \frac{\partial F_2}{\partial x}dA \tag{15.41}$$

using, respectively, (15.39) and (15.40). The conclusion of the theorem is then obtained by adding these equations.

The line integral is transformed into an ordinary integral first. The boundary ∂D contains four curves, denoted C_1, C_2, C_3, and C_4. The curve C_1 is the graph $y = y_{\rm bot}(x)$ whose parametric equations are $\mathbf{r} = (t, y_{\rm bot}(t))$, where $t \in [a,b]$. So C_1 is traced out from left to right as required by the positive orientation of ∂D. The curve C_3 is the top boundary $y = y_{\rm top}(x)$, and, similarly, its parametric equations $\mathbf{r}(t) = (t, y_{\rm top}(t))$, where $t \in [a,b]$. Since C_3 is traced out from left to right, the orientation of C_3 must be reversed; that is, ∂D contains the curve $-C_3$. The boundary curves C_2 and C_4 (the sides of D) are segments of the vertical lines $x = b$ (oriented upward) and $x = a$ (oriented downward), which may collapse to a single point if the graphs $y = y_{\rm bot}(x)$ and $y = y_{\rm top}(x)$ intersect at $x = a$ or $x = b$ or both. The line integrals along C_2 and C_4 do not contribute to the line integral with respect to x along ∂D because $dx = 0$ along C_2 and C_4. By construction, $x = t$ and $dx = dt$ for the curves C_1 and C_2. Hence,

$$\begin{aligned}\oint_{\partial D} F_1\,dx &= \int_{C_1} F_1\,dx + \int_{-C_2} F_1\,dx\\ &= \int_a^b \Big(F(x, y_{\rm bot}(x)) - F(x, y_{\rm top}(x))\Big)\,dx,\end{aligned}$$

where the property (15.36) has been used. Next, the double integral is transformed into an ordinary integral by converting it to an iterated integral:

$$\begin{aligned}\iint_D \frac{\partial F_1}{\partial y}dA &= \int_a^b \int_{y_{\rm bot}(x)}^{y_{\rm top}(x)} \frac{\partial F_1}{\partial y}\,dy\,dx\\ &= \int_a^b \Big(F(x, y_{\rm top}(x)) - F(x, y_{\rm bot}(x))\Big)dx,\end{aligned}$$

where the latter equality follows from the fundamental theorem of calculus and the continuity of F_1 on an open interval that contains

$[y_{\mathrm{bot}}(x), y_{\mathrm{top}}(x)]$ for any $x \in [a,b]$ (the hypothesis of Green's theorem). Comparing the expression of the line and double integrals via ordinary integrals, the validity of the first relation in (15.41) is established. The second equality in (15.41) is proved analogously by using (15.40). The details are omitted. □

Remark. Suppose that a smooth, oriented curve C divides a region D into two *simple* regions D_1 and D_2. If the boundary ∂D_1 contains C (i.e., the orientation of C coincides with the positive orientation of ∂D_1), then ∂D_2 must contain the curve $-C$ and vice versa. Using the conventional notation $F_1\,dx + F_2\,dy = \mathbf{F}\cdot d\mathbf{r}$, where $\mathbf{F} = (F_1, F_2)$, one infers that

$$\begin{aligned}\oint_{\partial D} \mathbf{F}\cdot d\mathbf{r} &= \oint_{\partial D_1} \mathbf{F}\cdot d\mathbf{r} + \oint_{\partial D_2} \mathbf{F}\cdot d\mathbf{r}\\ &= \iint_{D_1}\Big(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\Big)dA + \iint_{D_2}\Big(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\Big)dA\\ &= \iint_{D}\Big(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\Big)dA.\end{aligned}$$

The first equality holds because of the cancellation of the line integrals along C and $-C$ according to (15.36). The validity of the second equality follows from the proof of Green's theorem for simple regions. Finally, the equality is established by the additivity property of double integrals. By making use of similar arguments, the proof can be extended to a region D that can be represented as the union of a finite number of simple regions.

Remark. Let the regions D_1 and D_2 be bounded by simple, piecewise-smooth, closed curves and let D_2 lie in the interior of D_1. Consider the region D that was obtained from D_1 by removing D_2 (the region D has a hole of the shape D_2). Making use of Green's theorem, one finds

$$\begin{aligned}\iint_{D}\Big(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\Big)dA &= \iint_{D_1}\Big(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\Big)dA - \iint_{D_2}\Big(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\Big)dA\\ &= \oint\limits_{\partial D_1} \mathbf{F}\cdot d\mathbf{r} - \oint\limits_{\partial D_2} \mathbf{F}\cdot d\mathbf{r} = \oint\limits_{\partial D_1} \mathbf{F}\cdot d\mathbf{r} + \oint\limits_{-\partial D_2} \mathbf{F}\cdot d\mathbf{r}\\ &= \oint_{\partial D} \mathbf{F}\cdot d\mathbf{r}. \qquad (15.42)\end{aligned}$$

This establishes the validity of Green's theorem for not simply connected regions. The boundary ∂D consists of ∂D_1 and $-\partial D_2$; that is, the outer boundary has a positive orientation, while the inner boundary is negatively oriented. A similar line of reasoning leads to the conclusion that this holds for any number of holes in D: all inner boundaries of D must be negatively oriented. Such orientation of the boundaries can also be understood as follows. Let a curve C connects a point of the outer boundary with a point of the inner boundary. Let us make a cut of the region D along C. Then the region D becomes simply connected and ∂D consists of a *continuous* curve (the inner and outer boundaries, and the curves C and $-C$). The boundary ∂D can always be positively oriented. The latter requires that the outer boundary be traced counterclockwise, while the inner boundary is traced clockwise (the orientation of C and $-C$ is chosen accordingly). By applying Green's theorem to ∂D, one can see that the line integrals over C and $-C$ are cancelled and (15.42) follows from the additivity of the double integral. Evidently, the same argument can be used to establish Green's theorem for a region with multiple holes (all inner boundaries must be oriented clockwise).

112.1. Evaluating Line Integrals via Double Integrals. Green's theorem provides a technically convenient tool to evaluate line integrals along planar closed curves. It is especially beneficial when the curve consists of several smooth pieces that are defined by different vector functions; that is, the line integral must be split into a sum of line integrals to be converted into ordinary integrals. Sometimes, the line integral turns out to be much more difficult to evaluate than the double integral.

Example 15.5. *Evaluate the line integral of* $\mathbf{F} = (y^2 + e^{\cos x},\ 3xy - \sin(y^4))$ *along the curve* C *that is the boundary of the half of the ring:* $1 \leq x^2 + y^2 \leq 4$ *and* $y \geq 0$*;* C *is oriented clockwise.*

The curve C consists of four smooth pieces, the half-circles of radii 1 and 2 and two straight line segments of the x axis, $[-2, -1]$ and $[1, 2]$. Each curve can be easily parameterized and the line integral in question can be transformed into the sum of four ordinary integrals which are then evaluated. The reader is advised to pursue this avenue of actions to appreciate the following alternative way based on Green's theorem (this is not impossible to accomplish if one figures out how to handle the integration of the functions $e^{\cos x}$ and $\sin(y^4)$ whose anti-derivatives are not expressible in elementary functions).

Solution: The curve C is a simple, piecewise-smooth, closed curve and the vector field $\mathbf{F}$ is continuously differentiable. Thus, Green's theorem applies if $\partial D = -C$ (because the orientation of C is negative) and D is the half-ring. One has $\partial F_1/\partial y = 2y$ and $\partial F_2/\partial x = 3y$. By

Green's theorem,

$$\oint_C \mathbf{F}\cdot d\mathbf{r} = -\oint_{\partial D} \mathbf{F}\cdot d\mathbf{r} = -\iint_D \Big(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\Big) dA = -\iint_D y\, dA$$
$$= -\int_0^{\pi}\int_1^2 r\sin\theta \; r\,dr\,d\theta = -\int_0^{\pi} \sin\theta\, d\theta \int_1^2 r^2 dr = -\frac{14}{3},$$

where the double integral has been transformed into polar coordinates. □

112.2. Area of a Planar Region as a Line Integral. Put $F_2 = x$ and $F_1 = 0$. Then

$$\iint_D \Big(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\Big) dA = \iint_D dA = A(D).$$

The area $A(D)$ can also be obtained if $\mathbf{F} = (-y, 0)$ or $\mathbf{F} = (-y/2, x/2)$. By Green's theorem, the area of D can be expressed by line integrals:

$$A(D) = \oint_{\partial D} x\,dy = -\oint_{\partial D} y\,dx = \frac{1}{2}\oint_{\partial D} x\,dy - y\,dx, \tag{15.43}$$

assuming, of course, that the boundary of D is a simple, piecewise-smooth, closed curve (or several such curves if D has holes). The reason the values of these line integrals coincide is simple. The difference of any two vector fields involved is a conservative vector field whose line integral along a closed curve vanishes. For example, for $\mathbf{F} = (0, x)$ and $\mathbf{G}(-y, 0)$, the difference is $\mathbf{F} - \mathbf{G} = (y, x) = \nabla f$, where $f(x, y) = xy$, so that

$$\oint_{\partial D} \mathbf{F}\cdot d\mathbf{r} - \oint_{\partial D} \mathbf{G}\cdot d\mathbf{r} = \oint_{\partial D} (\mathbf{F} - \mathbf{G})\cdot d\mathbf{r} = \oint_{\partial D} \nabla f\cdot d\mathbf{r} = 0.$$

The representation (15.43) of the area of a planar region as the line integral along its boundary is quite useful when the shape of D is too complicated to be computed using a double integral (e.g., when D is not simple and/or a representation of boundaries of D by graphs becomes technically difficult).

Example 15.6. *Consider an arbitrary polygon whose vertices in counterclockwise order are (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). Find its area.*

Solution: Evidently, a generic polygon is not a simple region (e.g., it may have a starlike shape). So the double integral is not at all suitable for finding the area. In contrast, the line integral approach seems far more feasible as the boundary of the polygon consists of n straight line segments connecting neighboring vertices. If C_i is such a segment oriented from (x_i, y_i) to (x_{i+1}, y_{i+1}) for $i = 1, 2, ..., n-1$, then C_n goes from (x_n, y_n) to (x_1, y_1). A vector function that traces out a straight line segment from a point $\mathbf{r}_a$ to a point $\mathbf{r}_b$ is $\mathbf{r}(t) = \mathbf{r}_a + (\mathbf{r}_b - \mathbf{r}_a)t$, where $0 \leq t \leq 1$. For the segment C_i, take $\mathbf{r}_a = (x_i, y_i)$ and $\mathbf{r}_b = (x_{i+1}, y_{i+1})$. Hence, $x(t) = x_i - (x_{i+1} - x_i)t = x_i + \Delta x_i\, t$ and $y(t) = y_i + (y_{i+1} - y_i)t = y_i + \Delta y_i\, t$. For the vector field $\mathbf{F} = (-y, x)$ on C_i, one has

$$\begin{aligned}\mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = (-y(t), x(t)) \cdot (\Delta x_i, \Delta y_i) &= x_i\, \Delta y_i - y_i\, \Delta x_i \\ &= x_i y_{i+1} - y_i x_{i+1};\end{aligned}$$

that is, the t dependence cancels out. Therefore, taking into account that C_n goes from (x_n, y_n) to (x_1, y_1), the area is

$$\begin{aligned}A &= \frac{1}{2}\oint_{\partial D} x\, dy - y\, dx = \frac{1}{2}\sum_{i=1}^{n}\int_{C_i} x\, dy - y\, dx \\ &= \frac{1}{2}\sum_{i=1}^{n-1}\int_0^1 (x_i y_{i+1} - y_i x_{i+1})\, dt + \frac{1}{2}\int_0^1 (x_n y_1 - y_n x_1)\, dt \\ &= \frac{1}{2}\Big(\sum_{i=1}^{n-1}(x_i y_{i+1} - y_i x_{i+1}) + (x_n y_1 - y_n x_1)\Big).\end{aligned}$$

□

So Green's theorem offers an elegant way to find the area of a general polygon if the coordinates of its vertices are known. A simple, piecewise-smooth, closed curve C in a plane can always be approximated by a polygon. The area of the region enclosed by C can therefore be approximated by the area of a polygon with a large enough number of vertices, which is often used in many practical applications.

112.3. The Test for Planar Vector Fields to Be Conservative. Green's theorem can be used to prove Theorem 15.4 for planar vector fields. Consider a planar vector field $\mathbf{F} = (F_1(x, y), F_2(x, y), 0)$. Its curl has only one component:

$$\nabla \times \mathbf{F} = \det\begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_1(x, y) & F_2(x, y) & 0 \end{pmatrix} = \hat{\mathbf{e}}_3\Big(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\Big).$$

Suppose that the curl of $\mathbf{F}$ vanishes throughout a simply connected open region D, $\nabla \times \mathbf{F} = \mathbf{0}$. By definition, any simple closed curve C in a simply connected region D can be shrunk to a point of D while remaining in D throughout the deformation (i.e., any such C bounds a subregion D_s of D). By Green's theorem, where $C = \partial D_s$,

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = \iint_{D_s}\Big(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\Big)dA = \iint_{D_s} 0\; dA = 0$$

for any closed simple curve C in D. By the path-independence property, the vector field $\mathbf{F}$ is conservative in D.

112.4. Study Problems.

Problem 15.3. *Evaluate the line integral of* $\mathbf{F} = (y+e^{x^2},\, 3x-\sin(y^2))$ *along the counterclockwise-oriented boundary of D that is enclosed by the parabolas $y = x^2$ and $x = y^2$.*

Solution: One has $\partial F_1/\partial y = 1$ and $\partial F_2/\partial x = 3$. By Green's theorem,

$$\oint_{\partial D} \mathbf{F}\cdot d\mathbf{r} = \iint_D 2\,dA = 2\int_0^1 \int_{x^2}^{\sqrt{x}} dy\,dx = 2\int_0^1 (\sqrt{x} - x^2)\,dx = \frac{1}{3}$$

□

Problem 15.4. *Prove that the line integral of the planar vector field*

$$\mathbf{F} = \Big(-\frac{y}{x^2+y^2},\ \frac{x}{x^2+y^2}\Big)$$

along any positively oriented, simple, smooth, closed curve C that encircles the origin is 2π and that it vanishes for any such curve that does not encircle the origin.

Solution: It has been established (see Study Problem 15.1) that the curl of this vector field vanishes in the domain that is the entire plane with the origin removed. If C does not encircle the origin, then $\partial F_2/\partial x - \partial F_1/\partial y = 0$ throughout the region encircled by C, and the line integral along C vanishes by Green's theorem. Given a closed curve C that encircles the origin, but does not go through it, one can always find a disk of a small enough radius a such that the curve C does not intersect it. Let D_a be the region bounded by the circle C_a of radius a and the curve C. Then $\partial F_2/\partial x - \partial F_1/\partial y = 0$ throughout D_a. Let C be oriented counterclockwise, while C_a is oriented clockwise. Then ∂D_a is the union of C and C_a. By Green's theorem,

$$\oint_{\partial D} \mathbf{F}\cdot d\mathbf{r} = 0 \quad\Rightarrow\quad \oint_C \mathbf{F}\cdot d\mathbf{r} = -\oint_{C_a} \mathbf{F}\cdot d\mathbf{r} = \oint_{-C_a} \mathbf{F}\cdot d\mathbf{r} = 2\pi$$

because $-C_a$ is the circle oriented counterclockwise and for such a circle the line integral has been found to be 2π (see Study Problem 15.1). □

113. Flux of a Vector Field

The idea of a flux of a vector field stems from an engineering problem of mass transfer across a surface. Suppose there is a flow of a fluid or gas with a constant velocity $\mathbf{v}$ and a constant mass density σ (mass per unit volume). Let ΔA be a planar area element placed into the flow. At what rate is the fluid or gas carried by the flow across the area ΔA? In other words, what is the mass of fluid transferred across ΔA per unit time? This quantity is called a *flux* of the mass flow across the area ΔA.

Suppose first that the mass flow is normal to the area element. Consider the cylinder with an axis parallel to $\mathbf{v}$ with cross section area ΔA and height $h = v\,\Delta t$, where $v = \|\mathbf{v}\|$ is the flow speed and Δt is a time interval. The volume of the cylinder is $\Delta V = h\,\Delta A = v\,\Delta t\,\Delta A$. In time Δt, all the mass stored in this cylinder is transferred by the flow across ΔA. This mass is $\Delta m = \sigma\,\Delta V = \sigma v\,\Delta t\,\Delta A$, and the flux is

$$\Delta\Phi = \frac{\Delta m}{\Delta t} = \sigma v\,\Delta A.$$

The flux depends on the orientation of an area element relative to the flow. If the flow is parallel to the area element, then no mass is transferred across it. The velocity vector can be viewed as the sum of a vector normal to the area element and a vector tangential to it. Only the normal component of the flow contributes to the flux. If $\hat{\mathbf{n}}$ is the unit normal vector to the area element and θ is the angle between $\mathbf{v}$ and $\hat{\mathbf{n}}$, then the normal component of the velocity is $v_n = v\cos\theta = \mathbf{v}\cdot\hat{\mathbf{n}}$ and

$$\Delta\Phi = \sigma v_n\,\Delta A = \sigma\mathbf{v}\cdot\hat{\mathbf{n}}\,\Delta A = \mathbf{F}\cdot\hat{\mathbf{n}}\,\Delta A = F_n\,\Delta A, \tag{15.44}$$

where the vector $\mathbf{F} = \sigma\mathbf{v}$ characterizes the mass flow ("how much" (σ) and "how fast" ($\mathbf{v}$)) and F_n is its component normal to the area element.

If now the mass flow is not constant (i.e., $\mathbf{F}$ becomes a vector field), then its flux across a surface S can be defined by partitioning S into small surface area elements S_i, $i = 1, 2, ..., N$, whose surface areas are ΔS_i. Let $\mathbf{r}_i^*$ be a sample point in S_i and let $\hat{\mathbf{n}}_i$ be the unit vector normal to S_i at $\mathbf{r}_i^*$. If the size (the radius of the smallest ball containing S_i) is small, then, by neglecting variations in $\mathbf{F}$ and the normal $\hat{\mathbf{n}}$ within S_i, the flux across S_i can be approximated by (15.44), $\Delta\Phi_i \approx \mathbf{F}(\mathbf{r}_i^*)\cdot\hat{\mathbf{n}}_i\,\Delta S_i$. The approximation becomes better when $N \to \infty$ so that the sizes of S_i decrease to 0 uniformly and hence the total flux is

$$\Phi = \lim_{N\to\infty}\sum_{i=1}^{N}\Delta\Phi_i = \lim_{N\to\infty}\sum_{i=1}^{N}\mathbf{F}(\mathbf{r}_i^*)\cdot\hat{\mathbf{n}}_i\,\Delta S_i = \lim_{N\to\infty}\sum_{i=1}^{N}F_n(\mathbf{r}_i^*)\,\Delta S_i.$$

The sum in this equation is nothing but the Riemann sum of the function $F_n(\mathbf{r})$ over a partition of the surface S. Naturally, its limit is the surface integral of $F_n(\mathbf{r})$ over S. Thus, *the flux of a vector field across a surface is the surface integral of the normal component of the vector field.*

113.1. Orientable Surfaces. The above definition of the flux sounds rather plausible. However, it contains a tacit assumption that the normal component of a vector field can always be *uniquely* defined as a continuous function on a smooth surface. It appears that there are smooth surfaces for which this cannot be done!

The normal $\hat{\mathbf{n}} = \hat{\mathbf{n}}(\mathbf{r})$ depends on the point of a surface. So it is a vector field on S. In order for the normal component F_n to be uniquely defined, the rule $\hat{\mathbf{n}} = \hat{\mathbf{n}}(\mathbf{r})$ should assign just one $\hat{\mathbf{n}}$ for every point of S. Furthermore, $\hat{\mathbf{n}}(\mathbf{r})$ should be continuous on S and hence along every closed curve C in a smooth surface S. In other words, if $\hat{\mathbf{n}}$ is transported along a closed curve C in S, the initial $\hat{\mathbf{n}}$ must coincide with the final $\hat{\mathbf{n}}$. Since, at every point of S, there are only two possibilities to direct the unit normal vector, by continuity the direction of $\hat{\mathbf{n}}(\mathbf{r})$ defines one side of S, while the direction of $-\hat{\mathbf{n}}(\mathbf{r})$ defines the other side. Thus, the normal component of a vector field is well defined for two-sided surfaces. For example, the outward normal of a sphere is continuous along any closed curve on the sphere (it remains outward along any closed curve) and hence defines the outer side of the sphere. If the normal on the sphere is chosen to be inward, then it is also continuous and defines the inner side of the sphere.

Are there one-sided surfaces? If such a surface exists, it should have quite remarkable properties. Take a point on it. In a neighborhood of this point, one always thinks about two sides (a surface is smooth). One side is defined by a normal $\hat{\mathbf{n}}$ (face-up patch), while the other has the same shape but its normal is $-\hat{\mathbf{n}}$ (face-down patch). For a one-sided surface, the face-up and face-down patches must be on the same side of the surface. This implies that there should exist a curve on the surface that starts at a point on one side and can reach the very same point but from the other side *without* crossing the surface boundaries (if any) or piercing the surface. By moving the face-up patch along such a curve, it becomes the face-down patch. Thus, the normal cannot be uniquely defined on a one-sided S.

113.1.1. Examples of One-Sided Surfaces. One-sided surfaces do exist. To construct an example, take a rectangular piece of paper. Put upward arrows on its vertical sides and glue these sides so that the arrows remain parallel. In doing so, a cylinder is obtained, which is a two-sided surface (there is no curve that traverses from one side to the other without crossing the boundary circles formed by the horizontal sides of the rectangle). The gluing can be done differently. Before gluing the vertical sides, twist the rectangle so that the arrows on them become opposite and then glue them. The resulting surface is the famous *Möbius strip* (named after the German mathematician August Möbius). It is one-sided. All curves winding about it traverse both sides of the glued rectangle without crossing its boundaries (the horizontal edges).

There are one-sided surfaces without boundaries (like a sphere). The most famous one is a *Klein bottle.* Take a bottle. Drill a hole on the side surface and in the bottom of the bottle. Suppose the neck of the bottle is flexible (a "rubber" bottle). Bend its neck and pull it through the hole on the bottle's side surface (so that neck fits tightly into the hole). Finally, attach the edge of the bottle's neck to the edge of the hole in the bottle bottom. The result is a surface without boundaries and it is one-sided. A bug can crawl along this surface and get in and out of the bottle.

113.1.2. Flux and One-Sided Surfaces. The flux makes sense only for two-sided surfaces. Indeed, the flux means that something is being transferred from one side to the other side of the surface (i.e., *across* it) at a certain rate. If the surface is one-sided, then one can get "across it" by merely sliding along it! For example, a mass flow *tangential* to a one-sided surface can transfer mass across the surface.

Definition 15.9. (Orientable Surface).
A smooth surface is called orientable *if there is no closed curve in it such that the normal vector is reversed when moved around this curve.*

So orientable surfaces are two-sided surfaces. The flux of a vector field can only be defined across an orientable surface.

113.2. Flux as a Surface Integral.

Definition 15.10. (Flux of a Vector Field).
Let S be an orientable smooth surface and let $\hat{\mathbf{n}}$ be the unit normal vector on S. The flux of a vector field $\mathbf{F}$ across S is the surface integral

$$\Phi = \iint_S \mathbf{F} \cdot \hat{\mathbf{n}}\, dS,$$

provided the normal component $\mathbf{F} \cdot \hat{\mathbf{n}}$ of the vector field is integrable on S.

The integrability of the normal component $F_n(\mathbf{r}) = \mathbf{F}\cdot\hat{\mathbf{n}}$ is defined in the sense of surface integrals of ordinary functions (see Definition 14.16).

113.3. Evaluation of the Flux of a Vector Field. Suppose that a surface S is a graph $z = g(x, y)$ over a region $(x, y) \in D$. There are two possible orientations of S. The normal vector to the tangent plane at a point of S is $\mathbf{n} = (-g'_x, -g'_y, 1)$ (see Section 91.1). Its z component is positive. For this reason, the graph is said to be *oriented upward*. Alternatively, one can take the normal vector in the opposite direction, $\mathbf{n} = (g'_x, g'_y, -1)$. In this case, the graph is said to be *oriented downward*. Accordingly, the *upward* (*downward*) flux, denoted $\Phi_\uparrow$ ($\Phi_\downarrow$), of a vector field is associated with the upward (downward) orientation of the graph. When the orientation of a surface is reversed, the flux changes its sign:

$$\Phi_\uparrow = -\Phi_\downarrow.$$

Consider the upward-oriented graph $z = g(x, y)$. The unit normal vector reads

$$\hat{\mathbf{n}} = \frac{1}{\|\mathbf{n}\|}\,\mathbf{n} = \frac{1}{J}\left(-g'_x,\ g'_y,\ 1\right), \quad J = \sqrt{1 + (g'_x)^2 + (g'_y)^2}.$$

Recall that the area transformation law for a graph is $dS = J\,dA$. Therefore, in the infinitesimal flux across the surface area, dS can be written in the form

$$\mathbf{F}\cdot\hat{\mathbf{n}}\,dS = \mathbf{F}\cdot\mathbf{n}\,\frac{1}{J}\,J\,dA = \mathbf{F}\cdot\mathbf{n}\,dA,$$

where the vector field must be evaluated on S, that is, $\mathbf{F} = \mathbf{F}(x, y, g(x, y))$ (the variable z is replaced by $g(x, y)$ because $z = g(x, y)$ for any point $(x, y, z) \in S$). If the dot product $\mathbf{F}\cdot\mathbf{n}$ is an integrable function on D, the flux exists and is given by the double integral over D. The following theorem has been proved.

THEOREM 15.6. (Evaluation of the Flux Across a Graph). *Suppose that S is a graph $z = g(x, y)$ of a function g whose first-order partial derivatives are continuous on D. Let S be oriented upward by the normal vector $\mathbf{n} = (-g'_x, -g'_y, 1)$ and let $\mathbf{F}$ be a continuous vector field on S. Then*

$$\Phi_\uparrow = \iint_S \mathbf{F}\cdot\hat{\mathbf{n}}\,dS = \iint_D F_n(x, y)\,dA,$$

$$F_n(x, y) = \mathbf{F}\cdot\mathbf{n}\Big|_{z=g(x,y)} = -g'_x F_1(x, y, g) - g'_y F_2(x, y, g) + F_3(x, y, g).$$

The evaluation of the surface integral involves the following steps:

Step 1. Represent S as a graph $z = g(x, y)$ (i.e. find the function g using a geometrical description of S). If S cannot be represented as graph of a single function, then it has to be split into pieces so that each piece can be described as a graph. By the additivity property, the surface integral over S is the sum of integrals over each piece.
Step 2. Find the region D that defines the part of the graph that coincides with S (if S is not the graph on the whole domain of g).
Step 3. Determine the orientation of S (upward or downward) from the problem description. The sign of the flux is determined by the orientation. Calculate the normal component $F_n(x, y)$ of the vector field as a function on D.
Step 4. Evaluate the double integral of F_n over D.

EXAMPLE 15.7. *Evaluate the downward flux of the vector field* $\mathbf{F} = (xz, yz, z)$ *across the part of the paraboloid* $z = 1 - x^2 - y^2$ *in the first octant.*

SOLUTION: The surface is the part of the graph $z = g(x, y) = 1-x^2-y^2$ in the first octant. The paraboloid intersects the xy plane ($z = 0$) along the circle $x^2 + y^2 = 1$. Therefore, the region D is the quarter of the disk bounded by this circle in the first quadrant ($x, y \geq 0$). Since S is oriented downward, $\mathbf{n} = (g'_x, g'_y, -1) = (-2x, -2y, -1)$ and the normal component of $\mathbf{F}$ is

$$F_n(x, y) = (xg, yg, g) \cdot (-2x, -2y, -1) = -(1 - x^2 - y^2)(1 + 2x^2 + 2y^2).$$

Converting the double integral of F_n to polar coordinates,

$$\Phi_\downarrow = \iint_D F_n(x, y)\, dA = -\int_0^{\pi/2} \int_0^1 (1 + r^2)(1 + 2r^2)\, r\, dr\, d\theta = -\frac{19\pi}{24}.$$

The negative value of the downward flux means that the actual transfer of a quantity (like a mass), whose flow is described by the vector field $\mathbf{F}$, occurs in the upward direction across S. □

113.4. Parametric Surfaces. If the surface S in the flux integral is defined by the parametric equations $\mathbf{r} = \mathbf{r}(u, v)$, where $(u, v) \in D$, then, by Corollary 14.5, the normal vector to S is $\mathbf{n} = \mathbf{r}'_u \times \mathbf{r}'_v$ (or $-\mathbf{n}$; the sign is chosen according to the geometrical description of the orientation of S). Since $\|\mathbf{n}\| = J$, where J determines the area transformation law $dS = J\, dA$ ($dA = du\, dv$), the flux of a vector field $\mathbf{F}$ across the surface area dS reads:

$$\begin{aligned}\mathbf{F}(\mathbf{r}(u,v))\cdot\hat{\mathbf{n}}\,dS &= \mathbf{F}(\mathbf{r}(u,v))\cdot\mathbf{n}\,dA = \mathbf{F}(\mathbf{r}(u,v))\cdot(\mathbf{r}'_u\times\mathbf{r}'_v)\,dA\\ &= F_n(u,v)\,dA\end{aligned}$$

and the flux is given by the double integral

$$\Phi = \iint_F \mathbf{F}\cdot\hat{\mathbf{n}}\,dS = \iint_D \mathbf{F}(\mathbf{r}(u,v))\cdot(\mathbf{r}'_u\times\mathbf{r}'_v)\,dA = \iint_D F_n(u,v)\,dA.$$

Naturally, a graph $z = g(x,y)$ is described by the parametric equations $\mathbf{r}(u,v) = (u,v,g(u,v))$, which is a particular case of the above expression; it coincides with that given in Theorem 15.6 ($x = u$ and $y = v$). A description of surfaces by parametric equations is especially convenient for closed surfaces (i.e., when the surface cannot be represented as a graph of a single function).

Example 15.8. *Evaluate the outward flux of the vector field* $\mathbf{F} = (z^2x, z^2y, z^3)$ *across the sphere of unit radius centered at the origin.*

Solution: The parametric equations of the sphere of radius $R = 1$ are given in (14.31), and the normal vector is computed in Example 14.38: $\mathbf{n} = \sin(u)\mathbf{r}(u,v)$, where $\mathbf{r}(u,v) = (\cos v\sin u, \sin v\sin u, \cos u)$ and $(u,v)\in D = [0,\pi]\times[0,2\pi]$; it is an outward normal because $\sin u \geq 0$. It is convenient to represent $\mathbf{F} = z^2\mathbf{r}$ so that

$$\begin{aligned}F_n(u,v) &= \mathbf{F}(\mathbf{r}(u,v))\cdot\mathbf{n} = \cos^2 u\sin u\ \mathbf{r}(u,v)\cdot\mathbf{r}(u,v)\\ &= \cos^2 u\sin u\ \|\mathbf{r}(u,v)\|^2 = \cos^2 u\sin u\end{aligned}$$

because $\|\mathbf{r}(u,v)\|^2 = R^2 = 1$. The outward flux reads

$$\begin{aligned}\Phi &= \iint_S \mathbf{F}\cdot\hat{\mathbf{n}}\,dS = \iint_D \cos^2 u\sin u\,dA\\ &= \int_0^{2\pi} dv\int_0^{\pi}\cos^2 u\sin u\,du = \frac{4\pi}{3}.\end{aligned}$$

□

Nonorientable Surfaces. Nonorientable surfaces can be described by the parametric equations $\mathbf{r} = \mathbf{r}(u,v)$ or by an algebraic equation $F(x,y,z) = 0$ (as a level surface of a function). For example, a Möbius strip of width $2h$ with midcircle of radius R and height $z = 0$ is defined by the parametric equations

(15.45)

$$\mathbf{r}(u,v) = \Big([R + u\cos(v/2)]\cos v,\ [R + u\cos(v/2)]\sin v,\ u\sin(v/2)\Big),$$

where $(u, v) \in D = [-h, h] \times [0, 2\pi]$. It also follows from these parametric equations that the Möbius strip is defined by a *cubic* surface:

$$-R^2y + x^2y + y^3 - 2Rxz - 2x^2z - 2y^2z + yz^3 = 0.$$

This is verified by substituting the parametric equations into this algebraic equation and showing that the left side vanishes for all $(u, v) \in D$.

Let us prove that the surface defined by the parametric equations (15.45) is not orientable. To do so, one should analyze the behavior of a normal vector when the latter is moved around a closed curve in the surface. Consider the circle in the xy plane defined by the condition $u = 0$: $\mathbf{r}(0, v) = (R\cos v, R\sin v, 0)$. It is easy to show that

$$\begin{aligned} \mathbf{r}'_u(0, v) &= (\cos(v/2)\cos v,\ \cos(v/2)\sin v,\ \sin(v/2)), \\ \mathbf{r}'_v(0, v) &= (-R\sin v,\ R\cos v,\ 0). \end{aligned}$$

P When $\mathbf{r}(0, v)$ returns to the initial point, that is, $\mathbf{r}(0, v{+}2\pi) = \mathbf{r}(0, v)$, the normal vector is reversed. Indeed, $\mathbf{r}'_u(0, v + 2\pi) = -\mathbf{r}'_u(0, v)$ and $\mathbf{r}'_v(0, v + 2\pi) = \mathbf{r}'_v(0, v)$. Hence,

$$\begin{aligned} \mathbf{n}(0, v + 2\pi) &= \mathbf{r}'_u(0, v + 2\pi) \times \mathbf{r}'_v(0, v + 2\pi) = -\mathbf{r}'_u(0, v) \times \mathbf{r}'_v(0, v) \\ &= -\mathbf{n}(0, v); \end{aligned}$$

that is, the surface defined by these parametric equations is *not* orientable because the normal vector is reversed when moved around a closed curve.

So, if a surface S is defined by parametric or algebraic equations, one still has to verify that it is orientable (i.e., it is two-sided!), when evaluating the flux across it; otherwise, the flux makes no sense.

114. Stokes' Theorem

114.1. Vector Form of Green's Theorem. It was shown in Section 112.3 that the curl of a planar vector field $\mathbf{F}(x, y) = (F_1(x, y), F_2(x, y), 0)$ is parallel to the z axis, $\nabla \times F = (\partial F_2/\partial - \partial F_1/\partial x)\hat{\mathbf{e}}_3$. This observation allows us to reformulate Green's theorem in the following vector form:

$$\oint_{\partial D} \mathbf{F} \cdot d\mathbf{r} = \iint_D (\operatorname{curl} \mathbf{F}) \cdot \hat{\mathbf{e}}_3\ dA.$$

Thus, *the line integral of a vector field along a closed simple curve is determined by the flux of the curl of the vector field across the surface bounded by this curve.* It turns out that this statement holds not only in a plane, but also in space. It is known as *Stokes' theorem.*

114.2. Stokes' Theorem.

114.2.1. Positive (Induced) Orientation of a Closed Curve. Suppose S is a smooth surface oriented by its normal vector $\mathbf{n}$ and bounded by a closed simple curve C. Consider a tangent plane at a point $\mathbf{r}_0$ of S. Any circle in the tangent plane centered at $\mathbf{r}_0$ can always be oriented counterclockwise as viewed from the top of the normal vector $\mathbf{n} = \mathbf{n}_0$ at $\mathbf{r}_0$. This circle is said to be *positively oriented relative to the orientation of* S. Since the surface is smooth, a circle of a sufficiently small radius can always be projected onto a closed simple curve in S by moving each point of the circle parallel to $\mathbf{n}_0$. This curve is also *positively oriented* relative to $\mathbf{n}_0$. It can then be continuously (i.e., without breaking) deformed along S so that its part lies on the boundary C after the deformation. The orientation is preserved throughout the deformation, and hence it induces a *positive orientation* of the boundary C. The positively oriented boundary of S is denoted by ∂S.

In other words, the positive (or induced) orientation of C means that if one walks in the positive direction along C with one's head pointing in the direction of $\mathbf{n}$, then the surface will always be on one's left. Let S be a graph $z = g(x, y)$ over D oriented upward. Then ∂S is obtained from ∂D (a positively oriented boundary of D) by lifting points of ∂D to S parallel to the z axis.

Theorem 15.7. (Stokes' Theorem).
Let S be an oriented, piecewise-smooth surface that is bounded by a simple, closed, piecewise-smooth curve C with positive orientation $C = \partial S$. Let $\mathbf{F}$ be a continuously differentiable vector field on an open spatial region that contains S. Then

$$\oint_{\partial S} \mathbf{F} \cdot d\mathbf{r} = \iint_S \operatorname{curl} \mathbf{F} \cdot \hat{\mathbf{n}}\, dS,$$

where $\hat{\mathbf{n}}$ is the unit normal vector on S.

Stokes' theorem is difficult to prove in general. Here it is proved for a particular case when S is a graph of a function.

Proof (for S being a graph). Let S be the upward-oriented graph $z = g(x, y)$, $(x, y) \in D$, where g is twice continuously differentiable on D and D is a simple planar region whose boundary ∂D corresponds to the boundary ∂S. In this case, the normal vector $\mathbf{n} = (-g'_x, -g'_y, 1)$ and the upward flux of $\operatorname{curl} \mathbf{F}$ across S can be evaluated according to

Theorem 15.6, where $\mathbf{F}$ is replaced by $\nabla \times \mathbf{F}$:

$$\iint_S \operatorname{curl}\mathbf{F}\cdot\hat{\mathbf{n}}\,dS = \iint_D (\operatorname{curl}\mathbf{F})_n\,dA,$$

$$(\operatorname{curl}\mathbf{F})_n = -\Big(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\Big)\frac{\partial z}{\partial x} - \Big(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial z}\Big)\frac{\partial z}{\partial y} + \\ +\Big(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\Big),$$

where $\partial z/\partial x = g'_x$ and $\partial z/\partial y = g'_y$. Let $x = x(t)$ and $y = y(t)$, $t \in [a, b]$, be parametric equations of ∂D so that $x(a) = x(b)$ and $y(a) = y(b)$ (∂D is a closed curve). Then the vector function

$$\mathbf{r}(t) = (x(t),\ y(t),\ g(x(t), y(t)))\,, \quad t \in [a, b],$$

traces out the boundary ∂S, $\mathbf{r}(a) = \mathbf{r}(b)$. Making use of Theorem 15.1, the line integral of $\mathbf{F}$ along ∂S can be evaluated. One has $\mathbf{r}' = (x'\,, y'\,, g'_x x' + g'_y y')$. Therefore, $\mathbf{F}\cdot\mathbf{r}' = (F_1 + F_3 g'_x)x' + (F_2 + F_3 g'_y)y'$ and hence

$$\oint_{\partial S} \mathbf{F}\cdot d\mathbf{r} = \int_a^b [(F_1 + F_3 g'_x)x' + (F_2 + F_3 g'_y)y']\,dt \\ = \oint_{\partial D} \Big(F_1 + F_3\frac{\partial z}{\partial x}\Big)dx + \Big(F_2 + F_3\frac{\partial z}{\partial y}\Big)dy$$

because $x'\,dt = dx$ and $y'\,dt = dy$ along ∂D, where $z = g(x, y)$ in all components of $\mathbf{F}$. The latter line integral can be transformed into the double integral over D by Green's theorem:

$$\oint_{\partial S} \mathbf{F}\cdot d\mathbf{r} = \iint_D \Big[\frac{\partial}{\partial x}\Big(F_2 + F_3\frac{\partial z}{\partial y}\Big) - \frac{\partial}{\partial y}\Big(F_1 + F_3\frac{\partial z}{\partial x}\Big)\Big]dA \\ = \iint_D (\operatorname{curl}\mathbf{F})_n\,dA = \iint_S \operatorname{curl}\mathbf{F}\cdot\hat{\mathbf{n}}\,dS,$$

where the middle equality is verified by the direct evaluation of the partial derivatives using the chain rule. For example, $(\partial/\partial x)F_2(x, y, g(x, y)) = \partial F_2/\partial x + (\partial F_2/\partial z)(\partial z/\partial x)$. The terms containing the mixed derivatives $\partial^2 z/\partial x\,\partial y = g''_{xy} = g''_{yx}$ are cancelled out by Clairaut's theorem, while the other terms can be arranged to coincide with the expression for the normal component $(\operatorname{curl}\mathbf{F})_n$ found above. The last equality holds by Theorem 14.17 ($dS = J\,dA$ and $\mathbf{n} = J\hat{\mathbf{n}}$). □

114.3. Use of Stokes' Theorem. Stokes' theorem is very helpful for evaluating line integrals along closed curves of complicated shapes when a direct use of Theorem 15.1 is technically too involved. The procedure includes a few basic steps.

Step 1. Given a closed simple curve C, choose *any* smooth orientable surface S whose boundary is C. Note that, according to Stokes' theorem, the value of the line integral is independent of the choice of S. This freedom should be used to make S as simple as possible.
Step 2. Find the orientation of S (the direction of the normal vector) so that the orientation of C is positive relative to the normal of S, that is, $C = \partial S$.
Step 3. Evaluate $\mathbf{B} = \operatorname{curl}\mathbf{F}$ and calculate the flux of $\mathbf{B}$ across S.

EXAMPLE 15.9. *Evaluate the line integral of* $\mathbf{F} = (xy, yz, xz)$ *along the curve of intersection of the cylinder* $x^2 + y^2 = 1$ *and the plane* $x + y + z = 1$. *The curve is oriented clockwise as viewed from above.*

SOLUTION: The curve C lies in the plane $x + y + z = 1$. Therefore, the simplest choice of S is the portion of this plane that lies within the cylinder: $z = g(x, y) = 1 - x - y$, where $(x, y) \in D$ and D is the disk $x^2 + y^2 \le 1$. Since C is oriented clockwise as viewed from above, the orientation of S must be downward to make the orientation positive relative to the normal on S, that is, $\mathbf{n} = (g'_x, g'_y, -1) = (-1, -1, -1)$. Next,

$$\mathbf{B} = \nabla \times \mathbf{F} = \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ xy & yz & xz \end{pmatrix} = (-y, -z, -x).$$

Therefore, $B_n(x, y) = \mathbf{B} \cdot \mathbf{n} = (-y, -g, -x) \cdot (-1, -1, -1) = g(x, y) + y + x = 1$, and hence

$$\begin{aligned}\int_C \mathbf{F} \cdot d\mathbf{r} &= \int_{\partial S} \mathbf{F} \cdot d\mathbf{r} = \iint_S \mathbf{B} \cdot \hat{\mathbf{n}}\, dS = \iint_D B_n(x, y)\, dA \\ &= \iint_D dA = A(D) = \pi\end{aligned}$$

□

EXAMPLE 15.10. *Evaluate the line integral of* $\mathbf{F} = (z^2 y\,, -z^2 x\,, z)$ *along the curve* C *that is the boundary of the part of the paraboloid* $z = 1 - x^2 - y^2$ *in the first octant. The curve* C *is oriented counterclockwise as viewed from above.*

SOLUTION: Choose S to be the specified part of the paraboloid $z = g(x, y) = 1 - x^2 - y^2$, where $(x, y) \in D$ and D is the part of the disk $x^2 + y^2 \le 1$ in the first quadrant. The paraboloid must be oriented upward so that the given orientation of C is positive relative to the normal on S. Therefore, the normal vector is $\mathbf{n} = (-g'_x, -g'_y, 1) =$

$(2x, 2y, 1)$. Next,

$$\mathbf{B} = \nabla \times \mathbf{F} = \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ z^2 y & -z^2 x & z \end{pmatrix} = (2zx, 2zy, -2z^2)$$

so that $B_n(x, y) = \mathbf{B} \cdot \mathbf{n} = (2gx, 2gy, -2g^2) \cdot (2x, 2y, 1) = 4g(x^2 + y^2) - 2g^2 = 4g(1 - g) - 2g^2 = 4g - 6g^2$. Thus,

$$\begin{aligned} \int_C \mathbf{F} \cdot d\mathbf{r} &= \iint_S \mathbf{B} \cdot \hat{\mathbf{n}}\, dS = \iint_D B_n(x, y)\, dA \\ &= \int_0^{\pi/2} \int_0^1 [4(1 - r^2) - 6(1 - r^2)^2] r\, dr\, d\theta = \frac{7\pi}{15}, \end{aligned}$$

where the double integral has been converted to polar coordinates, $g(x, y) = 1 - r^2$. □

114.4. Geometrical Significance of the Curl. Stokes' theorem reveals the geometrical significance of the curl of a vector field. The line integral of a vector field along a closed curve C is often called the *circulation* of a vector field along C. Let $\mathbf{B} = \operatorname{curl} \mathbf{F}$ and let $\mathbf{B}_0 = \mathbf{B}(\mathbf{r}_0)$ at some point $\mathbf{r}_0$. Consider a plane through $\mathbf{r}_0$ normal to a unit vector $\hat{\mathbf{n}}$. Let C_a be a positively oriented simple, closed, smooth curve in the plane that encircles a region S_a of the plane and $\mathbf{r}_0 \in S_a$. Let a be the radius of the smallest disk centered at $\mathbf{r}_0$ that contains S_a. Consider the circulation of a vector field *per unit area* at a point $\mathbf{r}_0$ defined by

$$\lim_{a \to 0} \frac{1}{\Delta S} \oint_{C_a} \mathbf{F} \cdot d\mathbf{r} = \lim_{a \to 0} \frac{1}{\Delta S} \iint_{S_a} \mathbf{B} \cdot \hat{\mathbf{n}}\, dS = \mathbf{B}_0 \cdot \hat{\mathbf{n}} = (\operatorname{curl} \mathbf{F})_0 \cdot \hat{\mathbf{n}}.$$

This follows from the integral mean value theorem. Since the function $f(\mathbf{r}) = \mathbf{B} \cdot \hat{\mathbf{n}}$ is continuous on S_a, there is a point $\mathbf{r}_a \in S_a$ such that the surface integral of f equals $\Delta S\, f(\mathbf{r}_a)$. As $a \to 0$, $\mathbf{r}_a \to \mathbf{r}_0$ and, by the continuity of f, $f(\mathbf{r}_a) \to f(\mathbf{r}_0)$. This relation has the following mechanical interpretation. Let $\mathbf{F}$ describe a fluid flow $\mathbf{F} = \mathbf{v}$, where $\mathbf{v}$ is the fluid velocity vector field. Imagine a tiny paddle wheel in the fluid at a point $\mathbf{r}_0$ whose axis is directed along $\mathbf{n}$. The fluid exerts pressure on the paddles, causing the paddle wheel to rotate. The more work done by the pressure force along the loop C_a, the faster the wheel rotates. The wheel rotates fastest (maximal work) when its axis $\mathbf{n}$ is parallel to $\operatorname{curl} \mathbf{v}$ because, in this case, the normal component of the curl $\operatorname{curl} \mathbf{v} \cdot \hat{\mathbf{n}} = \|\operatorname{curl} \mathbf{v}\|$ is maximal. For this reason, the curl is often called the *rotation* of a vector field.

DEFINITION 15.11. (Rotational Vector Field).
A vector field $\mathbf{F}$ *that can be represented as the curl of another vector field* $\mathbf{A}$*, that is,* $\mathbf{F} = \nabla \times \mathbf{A}$*, is called a* rotational vector field.

The following theorem holds (the proof is omitted).

THEOREM 15.8. (Helmholtz's Theorem).
Let $\mathbf{F}$ *be a vector field on a bounded domain* E*, which is twice continuously differentiable. Then* $\mathbf{F}$ *can be decomposed into the sum of conservative and rotational vector field; that is, there is a function* f *and a vector field* $\mathbf{A}$ *such that*

$$\mathbf{F} = \nabla f + \nabla \times \mathbf{A}.$$

For example, electromagnetic waves are rotational components of electromagnetic fields, while the Coulomb field created by static charges is conservative.

114.5. Test for a Vector Field to Be Conservative. The test for a vector field to be conservative (Theorem 15.4) follows from Stokes' theorem. Indeed, in a simply connected region E, any simple, closed curve can be shrunk to a point while remaining in E through the deformation. Therefore, for any such curve C, one can always find a surface S in E such that $\partial S = C$ (e.g., C can be shrunk to a point along such S). If $\operatorname{curl}\mathbf{F} = \mathbf{0}$ throughout E, then, by Stokes' theorem,

$$\oint_C \mathbf{F}\cdot d\mathbf{r} = \iint_S \operatorname{curl}\mathbf{F}\cdot\hat{\mathbf{n}}\,dS = 0$$

for any simple closed curve C in E. By the path independence property, $\mathbf{F}$ is conservative. The assumption that E is simply connected is crucial. For example, if E is the entire space with the z axis removed (see Study Problem 15.1), then the z axis always pierces through any surface S bounded by a closed simple curve encircling the z axis, and one cannot claim that the curl vanishes everywhere on S.

115. Gauss-Ostrogradsky (Divergence) Theorem

115.1. Divergence of a Vector Field.

DEFINITION 15.12. (Divergence of a Vector Field).
Suppose that a vector field $\mathbf{F} = (F_1, F_2, F_3)$ *is differentiable. Then the scalar function*

$$\operatorname{div}\mathbf{F} = \nabla\cdot\mathbf{F} = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}$$

is called the divergence of a vector field.

Example 15.11. *Find the divergence of the vector field* $\mathbf{F} = (x^3 + \cos(yz),\ y + \sin(x^2 z),\ xyz)$.

Solution: One has

$$\operatorname{div}\mathbf{F} = (x^3 + \cos(yz))'_x + (y + \sin(x^2 z))'_y + (xyz)'_z = 3x^2 + 1 + yx.$$

□

Corollary 15.1. *A continuously differentiable rotational vector field is divergence free,* $\operatorname{div}\operatorname{curl}\mathbf{A} = 0$.

Proof. By definition, a rotational vector field has the form $\mathbf{F} = \operatorname{curl}\mathbf{A} = \nabla \times \mathbf{A}$, where $\mathbf{A}$ is twice continuously differentiable because, by the hypothesis, $\mathbf{F}$ is continuously differentiable. Therefore,

$$\operatorname{div}\mathbf{F} = \operatorname{div}\operatorname{curl}\mathbf{A} = \nabla \cdot \operatorname{curl}\mathbf{A} = \nabla \cdot (\nabla \times \mathbf{A}) = 0$$

by the rules of vector algebra (the triple product vanishes if any two vectors in it coincide). These rules are applicable because $\mathbf{A}$ is twice continuously differentiable (Clairaut's theorem holds for its components; see Section 111.4). □

115.2. Another Vector Form of Green's Theorem. Green's theorem relates a line integral along a closed curve of the *tangential* component of a planar vector field to the flux of the curl across the region bounded by the curve. Let us investigate the line integral of the *normal* component. If the vector function $\mathbf{r}(t) = (x(t), y(t))$, $a \le t \le b$, traces out the boundary C of D in the positive (counterclockwise) direction, then

$$\hat{\mathbf{T}}(t) = \frac{1}{\|\mathbf{r}'(t)\|}\Big(x'(t),\ y'(t)\Big), \qquad \hat{\mathbf{n}}(t) = \frac{1}{\|\mathbf{r}'(t)\|}\Big(y'(t),\ -x'(t)\Big),$$
$$\hat{\mathbf{T}} \cdot \hat{\mathbf{n}} = 0$$

are the unit tangent vector and the outward unit normal vector to the curve C, respectively. Consider the line integral $\oint_C \mathbf{F}\cdot\hat{\mathbf{n}}\,ds$ of the normal component of a planar vector field along C. One has $ds = \|\mathbf{r}'(t)\|dt$, and hence

$$\mathbf{F} \cdot \hat{\mathbf{n}}\,ds = F_1 y'\,dt - F_2 x'\,dt = F_1\,dy - F_2\,dx = \mathbf{G} \cdot d\mathbf{r},$$

where $\mathbf{G} = (-F_2, F_1)$. By Green's theorem applied to the line integral of the vector field $\mathbf{G}$,

$$\oint_C \mathbf{F}\cdot\hat{\mathbf{n}}\,ds = \oint_C \mathbf{G}\cdot d\mathbf{r} = \iint_D \Big(\frac{\partial G_2}{\partial x} - \frac{\partial G_1}{\partial y}\Big)\,dA = \iint_D \Big(\frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y}\Big)\,dA.$$

The integrand in the double integral is the divergence of $\mathbf{F}$. Thus, another vector form of Green's theorem has been obtained:

$$\oint_{\partial D} \mathbf{F} \cdot \hat{\mathbf{n}}\, ds = \iint_D \operatorname{div} \mathbf{F}\, dA.$$

For a planar vector field (think of a mass flow on a plane), the line integral on the left side can be viewed as the outward flux of $\mathbf{F}$ across the boundary of a region D. An extension of this form of Green's theorem to three-dimensional vector fields is known as the *divergence or Gauss-Ostrogradsky theorem.*

115.3. The Divergence Theorem.

Let a solid region E be bounded by a closed surface S. If the surface is oriented outward (the normal vector points outside of E), then it is denoted $S = \partial E$.

THEOREM 15.9. (Gauss-Ostrogradsky (Divergence) Theorem). *Suppose E is a bounded, closed region in space that has a piecewise-smooth boundary $S = \partial E$ oriented outward. If $\mathbf{F}$ is a continuously differentiable vector field on an open region that contains E, then*

$$\iint_{\partial E} \mathbf{F} \cdot \hat{\mathbf{n}}\, dS = \iiint_E \operatorname{div} \mathbf{F}\, dV.$$

The divergence theorem states that the outward flux of a vector field across a closed surface S is given by the triple integral of the divergence of the vector field over the solid region bounded by S. It provides a convenient technical tool to evaluate the flux of a vector field across a closed surface.

Remark. It should be noted that the boundary ∂E may contain several disjoint pieces. For example, let E be a solid region with a cavity. Then ∂E consists of two pieces, the outer boundary and the cavity boundary. Both pieces are oriented outward in the divergence theorem.

EXAMPLE 15.12. *Evaluate the flux of the vector field $\mathbf{F} = (4xy^2z + e^z\,, 4yx^2z\,, z^4 + \sin(xy))$ across the closed surface oriented outward that is the boundary of the part of the ball $x^2 + y^2 + z^2 \leq R^2$ in the first octant $(x, y, z \geq 0)$.*

SOLUTION: The divergence of the vector field is

$$\operatorname{div} \mathbf{F} = (4xy^2z + e^z)'_x + (4yx^2z)'_y + (z^4 + \sin(xy))'_z = 4z(x^2 + y^2 + z^2).$$

By the divergence theorem,

$$\begin{aligned}\iint_S \mathbf{F}\cdot\hat{\mathbf{n}}\,dS &= \iiint_E 4z(x^2+y^2+z^2)\,dV\\ &= \int_0^{\pi/2}\int_0^{\pi/2}\int_0^R 4\rho^3\cos\phi\,\rho^2\sin\phi\,d\rho\,d\phi\,d\theta = \frac{\pi R^6}{24},\end{aligned}$$

where the triple integral has been converted to spherical coordinates. The reader is advised to evaluate the flux *without* using the divergence theorem to appreciate the power of the latter! □

The divergence theorem can be used to change (simplify) the surface in the flux integral.

COROLLARY 15.2. *Let the boundary ∂E of a solid region E be the union of two surfaces S_1 and S_2. Suppose that all the hypotheses of the divergence theorem hold. Then*

$$\iint_{S_2} \mathbf{F}\cdot\hat{\mathbf{n}}\,dS = \iiint_E \operatorname{div}\mathbf{F}\,dV - \iint_{S_1}\mathbf{F}\cdot\hat{\mathbf{n}}\,dS.$$

This establishes a relation between the flux across S_1 and the flux across S_2 with a common boundary curve. Indeed, since ∂E is the union of two disjoint pieces S_1 and S_2, the surface integral over ∂E is the sum of the integrals over S_1 and S_2. On the other hand, the integral over ∂E can be expressed as a triple integral by the divergence theorem, which establishes the stated relation between the fluxes across S_1 and S_2.

EXAMPLE 15.13. *Evaluate the upward flux of the vector field $\mathbf{F} = (z^2\tan^{-1}(y^2+1),\ z^4\ln(x^2+1),\ z)$ across the part of the paraboloid $z = 2 - x^2 - y^2$ that lies above the plane $z = 1$.*

SOLUTION: Consider a solid E bounded by the paraboloid and the plane $z = 1$. Let S_2 be the part of the paraboloid that bounds E and let S_1 be the part of the plane $z = 1$ that bounds E. If S_2 is oriented upward and S_1 is oriented downward, then the boundary of E is oriented outward, and Corollary 15.2 applies. The surface S_1 is the part of the plane $z = 1$ bounded by the intersection curve of the paraboloid and the plane: $1 = 2 - x^2 - y^2$ or $x^2 + y^2 = 1$. So S_2 is the graph $z = g(x, y) = 1$ over D, which is the disk $x^2 + y^2 \leq 1$. The downward normal vector to S_1 is $\mathbf{n} = (g'_x, g'_y, -1) = (0, 0, -1)$, and hence $F_n = \mathbf{F}\cdot\mathbf{n} = -F_3(x, y, g) = -1$ on S_1 and

$$\iint_{S_1}\mathbf{F}\cdot\hat{\mathbf{n}}\,dS = \iint_D F_n(x,y)\,dA = -\iint_D dA = -A(D) = -\pi.$$

Next, the divergence of $\mathbf{F}$ is

$$\operatorname{div}\mathbf{F} = (z^2\tan^{-1}(y^2+1))'_x + (z^4\ln(x^2+1))'_y + (z)'_z = 0+0+1 = 1.$$

Hence,

$$\begin{aligned}\iiint_E \operatorname{div}\mathbf{F}\,dV &= \iiint_E dV = \int_0^{2\pi}\int_0^1\int_1^{2-r^2} r\,dz\,dr\,d\theta \\ &= 2\pi\int_0^1(1-r^2)r\,dr = \frac{\pi}{2},\end{aligned}$$

where the triple integral has been transformed into cylindrical coordinates for $E = \{(x,y,z)|z_{\mathrm{bot}} = 1 \le z \le 2-x^2-y^2 = z_{\mathrm{top}}\,, (x,y)\in D\}$. The upward flux of $\mathbf{F}$ across the paraboloid is now easy to find by Corollary 15.2:

$$\iint_{S_2}\mathbf{F}\cdot\hat{\mathbf{n}}\,dS = \iiint_E \operatorname{div}\mathbf{F}\,dV - \iint_{S_1}\mathbf{F}\cdot\hat{\mathbf{n}}\,dS = \frac{\pi}{2}+\pi = \frac{3\pi}{2}.$$

□

The reader is again advised to try to evaluate this directly to appreciate the power of the divergence theorem!

Corollary 15.3. *The flux of a continuously differentiable rotational vector field across an orientable, closed, piecewise-smooth surface S vanishes:*

$$\iint_S \operatorname{curl}\mathbf{A}\cdot\hat{\mathbf{n}}\,dS = 0.$$

Proof. The hypotheses of the divergence theorem are satisfied. Therefore,

$$\iint_S \operatorname{curl}\mathbf{A}\cdot\hat{\mathbf{n}}\,dS = \iiint_E \operatorname{div}\operatorname{curl}\mathbf{A}\,dV = 0$$

by Corollary 15.1. □

By Helmholtz's theorem, a vector field can always be decomposed into the sum of conservative and rotational vector fields. It follows then that only the conservative component of the vector field contributes to the flux across a closed surface. This observation is further elucidated with the help of the concept of vector field sources.

115.4. Sources of a Vector Field. Consider a simple region E_a of volume ΔV. Let a be the radius of the smallest ball that contains E_a and is centered at a point $\mathbf{r}_0$. Let us calculate the outward flux *per unit volume* of a continuously differentiable vector field $\mathbf{F}$ across the boundary ∂E_a, which is defined by

$$\lim_{a\to 0}\frac{1}{\Delta V}\iint_{\partial E_a}\mathbf{F}\cdot\hat{\mathbf{n}}\,dS=\lim_{a\to 0}\frac{1}{\Delta V}\iiint_{E_a}\operatorname{div}\mathbf{F}\,dV=\operatorname{div}\mathbf{F}(\mathbf{r}_0).$$

The latter equality follows from the integral mean value theorem. By the continuity of $\operatorname{div}\mathbf{F}$, and the integral mean value theorem, there is a point $\mathbf{r}_a\in E_a$ such that the triple integral equals $\Delta V\operatorname{div}\mathbf{F}(\mathbf{r}_a)$. In the limit $a\to 0$, $\mathbf{r}_a\to\mathbf{r}_0$. Thus, if the divergence is positive $\operatorname{div}\mathbf{F}(\mathbf{r}_0)>0$, the flux of the vector field across any small surface around $\mathbf{r}_0$ is positive. This, in turn, means that the flow lines of $\mathbf{F}$ are outgoing from $\mathbf{r}_0$ as if there is a *source* creating a flow at $\mathbf{r}_0$. Following the analogy with water flow, such a source is called a *faucet*. If $\operatorname{div}\mathbf{F}(\mathbf{r}_0)<0$, the flow lines disappear at $\mathbf{r}_0$ (the inward flow is positive). Such a source is called a *sink*. Thus, the divergence of a vector field determines the density of the sources of a vector field. For example, flow lines of a static electric field originate from positive electric charges and end on negative electric charges. So the divergence of the electric field determines the electric charge density in space.

The divergence theorem states that the outward flux of a vector field across a closed surface is determined by the total source of the vector field in the region bounded by the surface.