

Concepts in Calculus III

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

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 978-1-61610-162-6

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 11. Vectors and the Space Geometry	1
71. Rectangular Coordinates in Space	1
72. Vectors in Space	12
73. The Dot Product	25
74. The Cross Product	38
75. The Triple Product	51
76. Planes in Space	65
77. Lines in Space	73
78. Quadric Surfaces	82
Chapter 12. Vector Functions	97
79. Curves in Space and Vector Functions	97
80. Differentiation of Vector Functions	111
81. Integration of Vector Functions	120
82. Arc Length of a Curve	128
83. Curvature of a Space Curve	136
84. Applications to Mechanics and Geometry	147
Chapter 13. Differentiation of Multivariable Functions	163
85. Functions of Several Variables	163
86. Limits and Continuity	173
87. A General Strategy for Studying Limits	183
88. Partial Derivatives	196
89. Higher-Order Partial Derivatives	202
90. Linearization of Multivariable Functions	211
91. Chain Rules and Implicit Differentiation	221
92. The Differential and Taylor Polynomials	231
93. Directional Derivative and the Gradient	245
94. Maximum and Minimum Values	257
95. Maximum and Minimum Values (Continued)	268
96. Lagrange Multipliers	278

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

Chapter 14. Multiple Integrals	293
97. Double Integrals	293
98. Properties of the Double Integral	301
99. Iterated Integrals	310
100. Double Integrals Over General Regions	315
101. Double Integrals in Polar Coordinates	330
102. Change of Variables in Double Integrals	341
103. Triple Integrals	356
104. Triple Integrals in Cylindrical and Spherical Coordinates	369
105. Change of Variables in Triple Integrals	382
106. Improper Multiple Integrals	392
107. Line Integrals	403
108. Surface Integrals	408
109. Moments of Inertia and Center of Mass	423
Chapter 15. Vector Calculus	437
110. Line Integrals of a Vector Field	437
111. Fundamental Theorem for Line Integrals	446
112. Green's Theorem	458
113. Flux of a Vector Field	470
114. Stokes' Theorem	481
115. Gauss-Ostrogradsky (Divergence) Theorem	490
Acknowledgments	501

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 denoted by positive real numbers, while distances to points on the other half of the line are denoted 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 rectangular box 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 rectangular box. It is uniquely determined by its three sides adjacent at the origin. Let the number x denote 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 denoted as x , y , and z axes. Thus, to find a point in space with rectangular coordinates $(1, 2, -3)$, one has to construct a rectangular box 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 $z_0 < 0$.

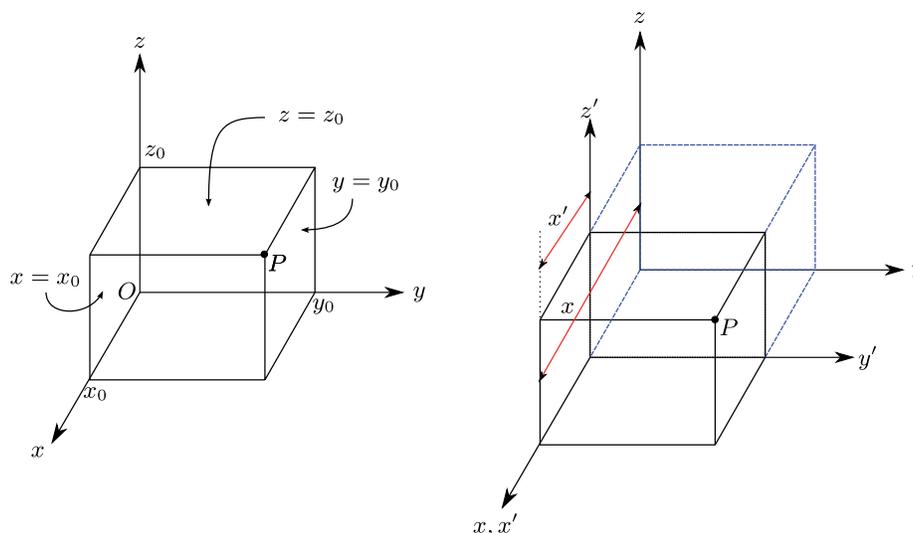


FIGURE 11.1. **Left:** Any point P in space can be viewed as the intersection of three coordinate planes $x = x_0$, $y = y_0$, and $z = z_0$; hence, P can be given an algebraic description as an ordered triple of numbers $P = (x_0, y_0, z_0)$. **Right:** Translation of the coordinate system. The origin 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

$$(11.1) \quad (x, y, z) \longrightarrow (x - x_0, y - y_0, z - z_0).$$

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 can be established. A simple case, when a coordinate system is rotated about one of its axes, is discussed in Study Problem 11.2.

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 rectangular box 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 rectangular box 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 rectangular box 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 rectangular 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 rectangular box 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.

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

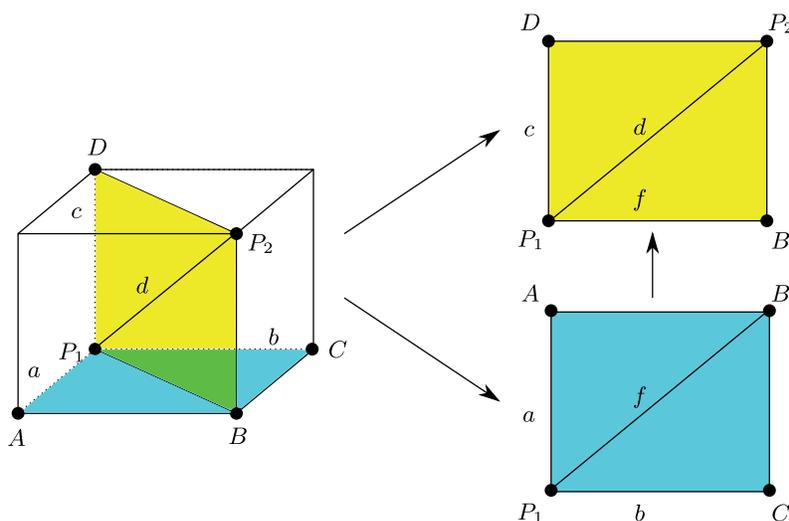


FIGURE 11.2. Distance between two points with coordinates $P_1 = (x_1, y_1, z_1)$ and $P_2 = (x_2, y_2, z_2)$. The line segment P_1P_2 is viewed as the largest diagonal of the rectangular box whose faces are the coordinate planes corresponding to the coordinates of the points. Therefore, the distances 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).

found:

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

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.

EXAMPLE 11.1. *A point moves 3 units of length parallel to a line, then it moves 6 units parallel to the second line that is perpendicular to the first line, and then it moves 6 units parallel to the third line that is perpendicular to the first and second lines. Find the distance between the initial and final positions.*

SOLUTION: The distance between points does not depend on the choice of the coordinate system. Let the origin be positioned at the initial point of the motion and let the coordinate axes be directed along the three mutually perpendicular lines parallel to which the point has moved. In this coordinate system, the final point has the coordinates $(3, 6, 6)$. The distance between this point and the origin $(0, 0, 0)$ is

$$D = \sqrt{3^2 + 6^2 + 6^2} = \sqrt{9(1 + 4 + 4)} = 9.$$

□

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, $x \rightarrow -x$ (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. An algebraic description of rotations in a plane and in space is given in Study Problems 11.2 and 11.20.

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. One of the simplest examples 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 *center* of the sphere. The distance from the sphere center to any point of the sphere is called the *radius* of the sphere.

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 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 radius of the sphere, 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

$$(11.3) \quad (x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 = R^2.$$

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 $x_0 = y_0 = z_0 = 0$.

EXAMPLE 11.2. Find the center and the radius of the sphere $x^2 + y^2 + z^2 - 2x + 4y - 6z + 5 = 0$.

SOLUTION: In order to find the coordinates of the center and the radius of the sphere, the equation must be transformed to the standard form (11.3) by completing the squares: $x^2 - 2x = (x - 1)^2 - 1$, $y^2 + 4y = (y + 2)^2 - 4$, and $z^2 - 6z = (z - 3)^2 - 9$. Then the equation of the sphere becomes

$$\begin{aligned} (x - 1)^2 - 1 + (y + 2)^2 - 4 + (z - 3)^2 - 9 + 5 &= 0, \\ (x - 1)^2 + (y + 2)^2 + (z - 3)^2 &= 9. \end{aligned}$$

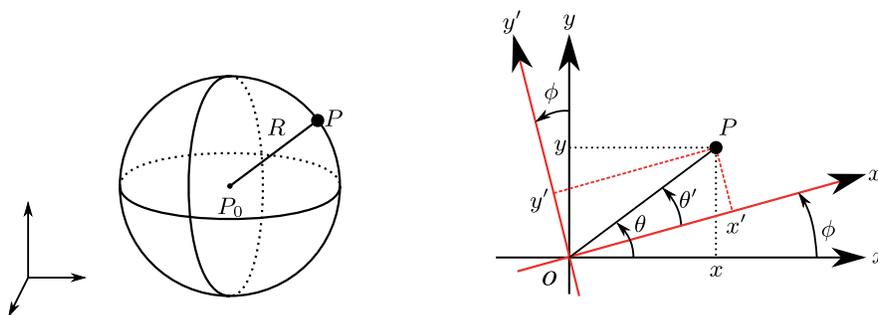


FIGURE 11.3. **Left:** A sphere is defined as a point set in space. Each point P of the set has a fixed distance R from a fixed point P_0 . The point P_0 is the center of the sphere, and R is the radius of the sphere. **Right:** An illustration to Study Problem 11.2. Transformation of coordinates under a rotation of the coordinate system in a plane.

A comparison with (11.3) shows that the center is at $(x_0, y_0, z_0) = (1, -2, 3)$ and the radius is $R = \sqrt{9} = 3$. \square

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

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

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:

$$\mathcal{P} = \left\{ (x, y, z) \mid z = 0 \right\}.$$

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

$$\mathcal{O}_1 = \left\{ (x, y, z) \mid x \geq 0, y \geq 0, z \geq 0 \right\}.$$

The spatial region

$$\mathcal{B} = \left\{ (x, y, z) \mid x > 0, y > 0, z > 0, x^2 + y^2 + z^2 < 4 \right\}$$

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 point with coordinates equal to half-sums of the corresponding coordinates of P_1 and P_2 . One has to prove that $|MP_1| = |MP_2| = \frac{1}{2}|P_1P_2|$. These two conditions define M as the midpoint. Consider a rectangular box B_1 whose largest diagonal is P_1M . The length of its side parallel to the x axis is $|(x_1 + x_2)/2 - x_1| = |x_2 - x_1|/2$. Similarly, its sides parallel to the y and z axes have the lengths $|y_2 - y_1|/2$ and $|z_2 - z_1|/2$, respectively. Consider a rectangular box B_2 whose largest diagonal is the segment MP_2 . Then its side parallel to the x axis has the length $|x_2 - (x_1 + x_2)/2| = |x_2 - x_1|/2$. Similarly, the sides parallel to the y and z axes have lengths $|y_2 - y_1|/2$ and $|z_2 - z_1|/2$, respectively. Thus, the sides of B_1 and B_2 parallel to each coordinate axis have the same length. By the distance formula (11.2), the diagonals of B_1 and B_2 must have the same length $|P_2M| = |MP_1|$. The lengths of the sides of a rectangular box whose largest diagonal is P_1P_2 are $|x_2 - x_1|$, $|y_2 - y_1|$, and $|z_2 - z_1|$. They are twice as long as the corresponding sides of B_1 and B_2 . If the length of each side of a rectangular box is scaled by a positive factor q , then the length of its diagonal is also scaled by q . In particular, this implies that $|MP_2| = \frac{1}{2}|P_1P_2|$. \square

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,

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

These equations define the transformation $(x, y) \rightarrow (x', y')$ of the old coordinates to the new ones. The inverse transformation $(x', y') \rightarrow (x, y)$ can be found by solving the equations for (x, y) . A simpler way is to note that if (x', y') are viewed as “old” coordinates and (x, y) as “new” coordinates, then the transformation that relates them is the rotation through the angle $-\phi$ (a clockwise rotation). Therefore, the inverse relations can be obtained by swapping the “old” and “new” coordinates and replacing ϕ by $-\phi$ in the direct relations. This yields

$$x = x' \cos \phi - y' \sin \phi, \quad y = y' \cos \phi + x' \sin \phi.$$

□

Problem 11.3. Give a geometrical description of the set

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

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

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

SOLUTION: As in the previous problem, the condition can be written via the sum of perfect squares $(x - 1)^2 + (y - 2)^2 \leq 9$ by means of the relations $x^2 - 2x = (x - 1)^2 - 1$ and $y^2 - 4y = (y - 2)^2 - 4$. In the xy plane, the inequality describes a disk of radius 3 whose center

is the point $(1, 2, 0)$. As the algebraic condition imposes no restriction on the z coordinate of points in the set, in any plane $z = z_0$ parallel to the xy plane, the x and y coordinates satisfy the same inequality, and hence the corresponding points also form a disk of radius 3 with the center $(1, 2, z_0)$. Thus, the set is a solid cylinder of radius 3 whose axis is parallel to the z axis and passes through the point $(1, 2, 0)$. \square

Problem 11.5. Give a geometrical description of the set

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

SOLUTION: The condition is satisfied if either $z = 0$ or $y = x$. The former equation describes the xy plane. 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. \square

71.8. Exercises.

(1) Find the distance between the following specified points:

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

(2) Find the distance from the point $(1, 2, 3)$ to each of the coordinate planes and to each of the coordinate axes.

(3) Find the length of the medians of the triangle with vertices $A(1, 2, 3)$, $B(-3, 2, -1)$, and $C(-1, -4, 1)$.

(4) 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 $(3, 2, 1)$. Sketch the set \mathcal{S} .

(5) 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) $(x - a)(y - b)(z - c) = 0$
- (x) $|x| \leq 1, |y| \leq 2, |z| \leq 3$

(6) 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) Three spheres centered at $(1, 2, 3)$ that touch the xy , yz , and xz planes, respectively.
- (iii) Three spheres centered at $(1, -2, 3)$ that touch the x , y , and z coordinate axes, respectively.
- (iv) 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. Compare the cases when the boundaries of the solid are included in the set or excluded from it.
- (v) 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. Compare the cases when the boundaries of the solid are included in the set or excluded from it.
- (vi) The part of a ball of radius R that lies between two parallel planes each of which is at a distance of $a < R$ from the center of the ball. Choose a convenient coordinate system. Compare the cases when the boundaries of the solid are included in the set or excluded from it.

(7) Consider the points P such that the distance from P to the point $(-3, 6, 9)$ is twice the distance from P to the origin. Show that the set of all such points is a sphere and find its center and radius.

(8) Find the volume of the solid bounded by the spheres $x^2 + y^2 + z^2 - 6z = 0$ and $x^2 + y^2 - 2y + z^2 - 6z = -9$.

(9) The solid region is described by the inequalities $|x - a| \leq a$, $|y - b| \leq b$, $|z - c| \leq c$, and $(y - b)^2 + (z - c)^2 \geq R^2$. If $R \leq \min(b, c)$, sketch the solid and find its volume.

(10) Sketch the set of all points in the xy plane that are equidistant from two given points A and B . Let A and B be $(1, 2)$ and $(-2, -1)$, respectively. Give an algebraic description of the set. Sketch the set of all points in space that are equidistant from two given points A and B . Let A and B be $(1, 2, 3)$ and $(-3, -2, -1)$, respectively. Give an algebraic description of the set.

72. Vectors in Space

72.1. Oriented Segments and Vectors. Suppose there is a pointlike 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* $\overrightarrow{P_1P_2}$ 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*. 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*.

Vectors will be denoted by boldface letters. *Two oriented segments* \overrightarrow{AB} and \overrightarrow{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 $\mathbf{a} = \overrightarrow{AB}$ or $\mathbf{a} = \overrightarrow{CD}$. The fact that the oriented segments \overrightarrow{AB} and \overrightarrow{CD} represent the same vector is denoted by the equality $\overrightarrow{AB} = \overrightarrow{CD}$.

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

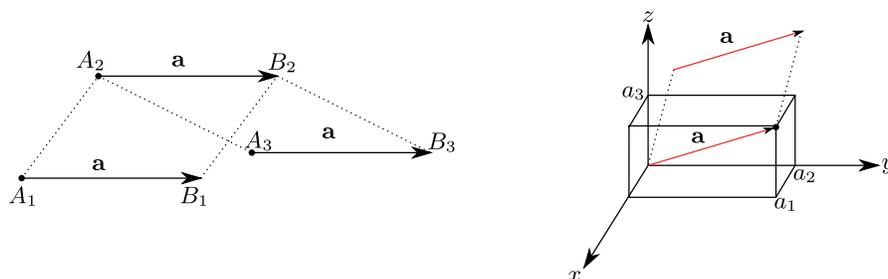


FIGURE 11.4. **Left:** Oriented segments obtained from one another by parallel transport. They all represent the same vector. **Right:** A vector as an ordered triple of numbers. An oriented segment is transported parallel so that its initial point coincides with the origin of a rectangular coordinate 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.

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} .

Consider a point A with coordinates (a_1, a_2, a_3) in some rectangular coordinate system. The vector $\mathbf{a} = \overrightarrow{OA} = \langle a_1, a_2, a_3 \rangle$ is called the *position vector* of A relative to the given coordinate system. This establishes a one-to-one correspondence between vectors and points in space. In particular, if the coordinate system is changed by a rotation of its axes about the origin, the components of a vector \mathbf{a} are transformed in the same way as the coordinates of a point whose position vector is \mathbf{a} .

DEFINITION 11.2. (Equality of Two Vectors).

Two vectors \mathbf{a} and \mathbf{b} are equal or coincide if their corresponding components are equal:

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

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. By virtue of the correspondence between vectors and points in space, this definition reflects the fact that two same points should have the same position vectors.

EXAMPLE 11.3. Find the components of a vector $\overrightarrow{P_1P_2}$ 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 rectangular box 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 $\overrightarrow{P_1P_2}$. Alternatively, the origin of the coordinate system can be moved to the point P_1 , keeping the directions of the coordinate axes. Therefore,

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

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

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 11.3 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 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 rectangular box with one vertex at the origin and the opposite vertex at the point (a_1, a_2, a_3) . The adjacent sides of the rectangular box have lengths given by the corresponding components of \mathbf{a} (modulo the signs if the components happen to be negative). When the lengths of the sides are scaled by a factor $s > 0$, a new rectangular box is obtained whose largest diagonal is parallel to \mathbf{a} . 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

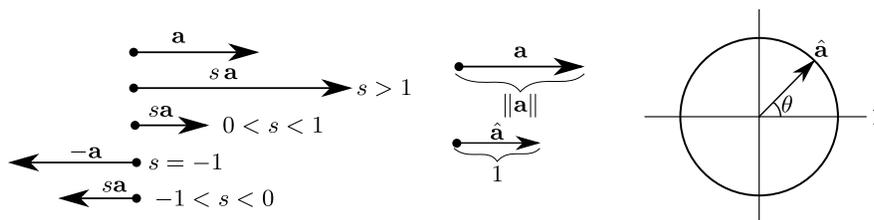


FIGURE 11.5. Left: Multiplication of a vector \mathbf{a} by a number s . If $s > 0$, the result of the multiplication is a vector parallel to \mathbf{a} whose length is scaled by the factor s . If $s < 0$, then $s\mathbf{a}$ is a vector whose direction is the opposite to that of \mathbf{a} and whose 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 $\|\mathbf{a}\|$, that is, $\hat{\mathbf{a}} = s\mathbf{a}$, 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 $\hat{\mathbf{a}} = \langle \cos \theta, \sin \theta, 0 \rangle$.

opposite to \mathbf{a} . The magnitude of $s\mathbf{a}$ is

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

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

$$\mathbf{a} \parallel \mathbf{b} \iff \mathbf{a} = s\mathbf{b}$$

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

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

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 $a_2/b_2 = a_3/b_3$. Owing to the geometrical interpretation of $s\mathbf{b}$, all points in space whose position vectors are parallel to a given nonzero vector \mathbf{b} form a line (through the origin) that is parallel to \mathbf{b} .

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 in the direction of \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 (see the right panel of Figure 11.5). Note that, in many practical applications, the components of a vector often have dimensions. For instance, the components of a position 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 \mathbf{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 \mathbf{a} .

EXAMPLE 11.4. Let $A = (1, 2, 3)$ and $B = (3, 1, 1)$. Find $\mathbf{a} = \overrightarrow{AB}$, $\mathbf{b} = \overrightarrow{BA}$, the unit vectors $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$, and the vector $\mathbf{c} = -2\overrightarrow{AB}$ and its norm.

SOLUTION: By Example 11.3, $\mathbf{a} = \langle 3-1, 2-1, 1-3 \rangle = \langle 2, -1, -2 \rangle$. The norm of \mathbf{a} is $\|\mathbf{a}\| = \sqrt{2^2 + (-1)^2 + (-2)^2} = \sqrt{9} = 3$. The unit vector in the direction of \mathbf{a} is $\hat{\mathbf{a}} = (1/3)\mathbf{a} = \langle 2/3, -1/3, -2/3 \rangle$. Using the rule of multiplication of a vector by a number, $\mathbf{c} = -2\mathbf{a} = -2\langle 2, -1, -2 \rangle = \langle -4, 2, 4 \rangle$ and $\|\mathbf{c}\| = \|(-2)\mathbf{a}\| = |-2|\|\mathbf{a}\| = 2\|\mathbf{a}\| = 6$. The direction of \overrightarrow{BA} is the opposite to \overrightarrow{AB} , and both vectors have the same length. Therefore, $\mathbf{b} = \langle -2, 1, 2 \rangle$, $\|\mathbf{b}\| = 3$, and $\hat{\mathbf{b}} = -\hat{\mathbf{a}} = \langle -2/3, 1/3, 2/3 \rangle$. \square

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 point B of the deck. The velocity vector relative to the deck is $\mathbf{v} = \overrightarrow{AB}$ 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 $\mathbf{u} = \overrightarrow{CD}$ 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 \mathbf{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 $\overrightarrow{AB'}$. Apparently, the displacement vector $\overrightarrow{BB'}$ coincides with the ship's velocity \mathbf{u} because B travels the distance u parallel to \mathbf{u} . This suggests a simple geometrical rule for finding $\overrightarrow{AB'}$ as shown in Figure 11.6. Take the vector $\overrightarrow{AB} = \mathbf{v}$, place the vector \mathbf{u} so that its initial point coincides with B , and make the oriented segment with the initial point of \mathbf{v} and the final point of \mathbf{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 \mathbf{a} and \mathbf{b} , extend from the vertex of the parallelogram. The sum of the vectors \mathbf{a} and \mathbf{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

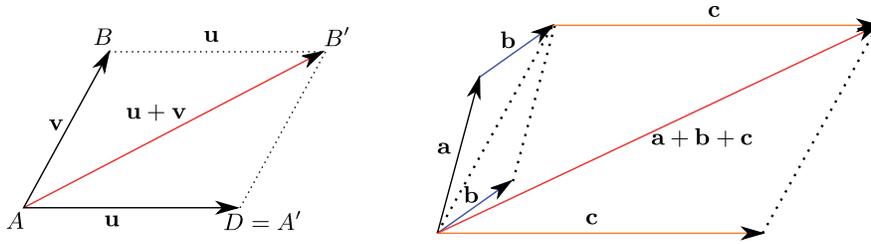


FIGURE 11.6. **Left:** Parallelogram rule for adding two vectors. If two vectors form adjacent sides of a parallelogram at a vertex A , then the sum of the vectors is a vector that coincides with the diagonal of the parallelogram and originates at the vertex A . **Right:** Adding several vectors by using 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 the addition of vectors is *commutative*:

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

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 terminal point of the first vector. The third vector is moved parallel so that its initial point coincides with the terminal 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 terminal point coincides with the terminal 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.

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 $\mathbf{a} = \overrightarrow{OA}$, 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 $\mathbf{a} + \mathbf{b} = \overrightarrow{OC} = \langle c_1, c_2, c_3 \rangle$, where C has coordinates (c_1, c_2, c_3) . By the parallelogram rule, $\mathbf{b} = \overrightarrow{AC} = \langle c_1 - a_1, c_2 - a_2, c_3 - a_3 \rangle$, where the relation between the components of a vector and the coordinates of its endpoints has been used (see Example 11.3). 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.

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:

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

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).

EXAMPLE 11.5. *An object travels 3 seconds with velocity $\mathbf{v} = \langle 1, 2, 4 \rangle$, where the components are given in meters per second, and then 2 seconds with velocity $\mathbf{u} = \langle 2, 4, 1 \rangle$. Find the distance between the initial and terminal points of the motion.*

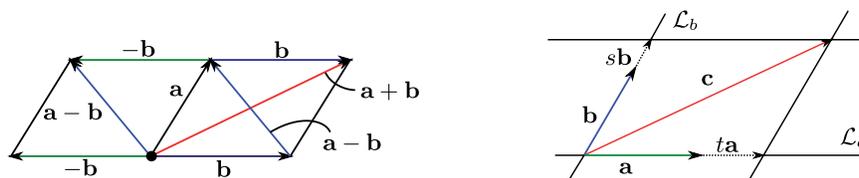


FIGURE 11.7. **Left:** Subtraction of two vectors. The difference $\mathbf{a} - \mathbf{b}$ is viewed as the sum of \mathbf{a} and $-\mathbf{b}$, the vector that has the direction opposite to \mathbf{b} and the same length as \mathbf{b} . The parallelogram rule for adding \mathbf{a} and $-\mathbf{b}$ 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.

SOLUTION: Let the initial and terminal points be A and B , respectively. Let C be the point at which the velocity was changed. Then $\overrightarrow{AC} = 3\mathbf{v}$ and $\overrightarrow{CB} = 2\mathbf{u}$. Therefore,

$$\begin{aligned}\overrightarrow{AB} &= \overrightarrow{AC} + \overrightarrow{CB} = 3\mathbf{v} + 2\mathbf{u} = 3\langle 1, 2, 4 \rangle + 2\langle 2, 4, 1 \rangle \\ &= \langle 3, 6, 12 \rangle + \langle 4, 8, 2 \rangle = \langle 7, 14, 14 \rangle = 7\langle 1, 2, 2 \rangle.\end{aligned}$$

The distance $|AB|$ is the length (or the norm) of the vector \overrightarrow{AB} . So $|AB| = \|\overrightarrow{AB}\| = 7\|\langle 1, 2, 2 \rangle\| = 7\sqrt{1 + 4 + 4} = 21$ meters. \square

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 two lines through the terminal 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 points of intersection of these lines with \mathcal{L}_a and \mathcal{L}_b and the initial and terminal 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, $\mathbf{c} = t\mathbf{a} + s\mathbf{b}$. \square

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 $\mathbf{a} = \overrightarrow{OA} = \langle 1, -1, 2 \rangle$. 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, $\mathbf{b} = \langle 1, -1, 2 \rangle + 2\langle 2, 1, -2 \rangle = \langle 5, 1, -2 \rangle$. \square

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

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, $\mathbf{c} = 6\mathbf{a} + 3\mathbf{b}$. \square

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, $\mathbf{u} = \langle u_1, u_2, u_3 \rangle \rightarrow \langle u_1, -u_2, u_3 \rangle \rightarrow \langle -u_1, -u_2, u_3 \rangle \rightarrow \langle -u_1, -u_2, -u_3 \rangle = -\mathbf{u}$. \square

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 that has endpoints $A(1, 2, 3)$ and $B(-1, 5, 1)$ and is directed from A to B
- (ii) The vector that has endpoints $A(1, 2, 3)$ and $B(-1, 5, 1)$ and is directed from B to A
- (iii) The vector that 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 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) Let \mathbf{a} and \mathbf{b} be two vectors that are neither parallel nor perpendicular. Sketch each of the following vectors: $\mathbf{a} + 2\mathbf{b}$, $\mathbf{b} - 2\mathbf{a}$, $\mathbf{a} - \frac{1}{2}\mathbf{b}$, and $2\mathbf{a} + 3\mathbf{b}$.

(3) Let \mathbf{a} , \mathbf{b} , and \mathbf{c} be three vectors in a plane any of which is not parallel to the others. Sketch each of the following vectors: $\mathbf{a} + (\mathbf{b} - \mathbf{c})$, $(\mathbf{a} + \mathbf{b}) - \mathbf{c}$, $2\mathbf{a} - 3(\mathbf{b} + \mathbf{c})$, and $(2\mathbf{a} - 3\mathbf{b}) - 3\mathbf{c}$.

(4) Let $\mathbf{a} = \langle 2, -1, -2 \rangle$ and $\mathbf{b} = \langle -3, 0, 4 \rangle$. Find unit vectors $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$. Express $6\hat{\mathbf{a}} - 15\hat{\mathbf{b}}$ via \mathbf{a} and \mathbf{b} .

(5) Let \mathbf{a} and \mathbf{b} be vectors in the xy plane such that their sum $\mathbf{c} = \mathbf{a} + \mathbf{b}$ makes the angle $\pi/3$ with \mathbf{a} and has length twice the length of \mathbf{a} . Find

\mathbf{b} if \mathbf{a} lies in the first quadrant, makes the angle $\pi/3$ with the positive x axis, and has length 2.

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

(7) Let $\hat{\mathbf{u}}_k$, $k = 1, 2, \dots, n$, be unit vectors in a plane such that the smallest angle between the two vectors $\hat{\mathbf{u}}_k$ and $\hat{\mathbf{u}}_{k+1}$ is $2\pi/n$. What is the sum $\mathbf{v}_n = \hat{\mathbf{u}}_1 + \hat{\mathbf{u}}_2 + \dots + \hat{\mathbf{u}}_n$ for an even n ? Sketch the sum for $n = 1$, $n = 3$, and $n = 5$. Compare the norms $\|\mathbf{v}_n\|$ for $n = 1, 3, 5$. Investigate the limit of \mathbf{v}_n as $n \rightarrow \infty$ by studying the limit of $\|\mathbf{v}_n\|$ as $n \rightarrow \infty$.

(8) Let $\hat{\mathbf{u}}_k$, $k = 1, 2, \dots, n$, be unit vectors as defined in exercise (7). Let $\mathbf{w}_k = \hat{\mathbf{u}}_{k+1} - \hat{\mathbf{u}}_k$ for $k = 1, 2, \dots, n-1$ and $\mathbf{w}_n = \hat{\mathbf{u}}_1 - \hat{\mathbf{u}}_n$. Find the limit of $\|\mathbf{w}_1\| + \|\mathbf{w}_2\| + \dots + \|\mathbf{w}_n\|$ as $n \rightarrow \infty$.

(9) 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?

(10) Use vector algebra to show that the line segment joining the midpoints of two sides of a triangle is parallel to the third side and half its length.

(11) Let \mathbf{a} and \mathbf{b} be position vectors in the xy plane. Describe the set of all points in the plane whose position vectors \mathbf{r} satisfy the condition $\|\mathbf{r} - \mathbf{a}\| + \|\mathbf{r} - \mathbf{b}\| = k$, where $k > \|\mathbf{a} - \mathbf{b}\|$.

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

$$m_1 \overrightarrow{P_0 P_1} + m_2 \overrightarrow{P_0 P_2} + \dots + m_n \overrightarrow{P_0 P_n} = \mathbf{0}.$$

Express the position vector \mathbf{r}_0 of P_0 via the position vectors \mathbf{r}_i of P_i . 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 $C(1, 1, -1)$.

(13) Consider the graph $y = f(x)$ of a differentiable function and the line tangent to it at a point $x = a$. Express components of a vector parallel to the line via $f'(a)$ and find a vector perpendicular to the line. In particular, find such vectors for the graph $y = x^2$ at the point $x = 1$.

(14) Let the vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} have fixed lengths a , b , and c , respectively, while their directions may be changed. Is it always possible to achieve $\mathbf{a} + \mathbf{b} + \mathbf{c} = \mathbf{0}$? If not, formulate the condition under which it is possible.

(15) Let the vectors \mathbf{a} and \mathbf{b} have fixed lengths, while their directions may be changed. Put $c_{\pm} = \|\mathbf{a} \pm \mathbf{b}\|$. Is it always possible to achieve $c_- > c_+$, or $c_- = c_+$, or $c_- < c_+$? If so, give examples of the corresponding relative directions of \mathbf{a} and \mathbf{b} .

(16) A point object travels so that its trajectory is in a plane and consists of straight line segments. The object always makes a turn 90° counterclockwise after traveling a distance d and then travels the distance sd , $0 < s < 1$, before making the next turn. If the object travels a distance a before the first turn, how far can the object get from the initial point if it keeps moving forever? *Hint:* Investigate the components of the position vector of the object in an appropriate coordinate system.

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 \mathbf{a} , \mathbf{b} , and \mathbf{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 \mathbf{a} is scaled first and then multiplied by \mathbf{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*.

EXAMPLE 11.6. Let $\mathbf{a} = \langle 1, 2, 3 \rangle$, $\mathbf{b} = \langle 2, -1, 1 \rangle$, and $\mathbf{c} = \langle 1, 1, -1 \rangle$. Find $\mathbf{a} \cdot (2\mathbf{b} - 5\mathbf{c})$.

SOLUTION: One has $\mathbf{a} \cdot \mathbf{b} = 1 \cdot 2 + 2 \cdot (-1) + 3 \cdot 1 = 2 - 2 + 3 = 3$ and, similarly, $\mathbf{a} \cdot \mathbf{c} = 1 + 2 - 3 = 0$. By the properties of the dot product:

$$\mathbf{a} \cdot (2\mathbf{b} - 5\mathbf{c}) = 2\mathbf{a} \cdot \mathbf{b} - 5\mathbf{a} \cdot \mathbf{c} = 6 - 0 = 6.$$

□

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:

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

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 \mathbf{a} are defined. Next, consider the triangle whose adjacent sides are the vectors \mathbf{a} and \mathbf{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

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

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:

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

This means that the numerical value of the dot product is independent of the choice of a 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 said to be *orthogonal*, and they point in the opposite direction 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,

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

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.

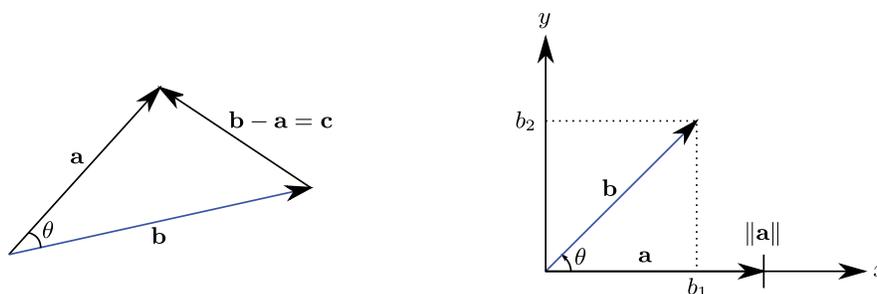


FIGURE 11.8. **Left:** Independence of the dot product from the choice of a coordinate system. The dot product of two vectors that are adjacent sides of a triangle can be expressed via the lengths of the triangle sides as shown in (11.5). **Right:** Geometrical significance of the dot product. It determines 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.

THEOREM 11.1. (Geometrical Significance of the Dot Product).

If θ is the angle between the vectors \mathbf{a} and \mathbf{b} , then $\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta$. In particular, two nonzero vectors are orthogonal if and only if their dot product vanishes:

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

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, $c^2 = a^2 + b^2$ (the Pythagorean theorem).

EXAMPLE 11.7. 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 $\alpha + \beta + \gamma = 180^\circ$. So it is sufficient to find any two angles. To find the angle α , define the vectors $\mathbf{a} = \overrightarrow{AB} = \langle -2, 1, 2 \rangle$ and $\mathbf{b} = \overrightarrow{AC} = \langle 0, 3, -4 \rangle$. 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,

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

To find the angle β , define the vectors $\mathbf{a} = \overrightarrow{BA} = \langle 2, -1, -2 \rangle$ and $\mathbf{b} = \overrightarrow{BC} = \langle 2, 2, -6 \rangle$ 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 $\beta = \cos^{-1}(7/(3\sqrt{11})) \approx 45.3^\circ$. Therefore, $\gamma \approx 180^\circ - 109.5^\circ - 45.3^\circ = 25.2^\circ$. 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. \square

73.2. Further Geometrical Properties of the Dot Product.

COROLLARY 11.1. (Orthogonal Decomposition of a Vector).

Given a nonzero vector \mathbf{a} , any vector \mathbf{b} can be uniquely decomposed into the sum of two orthogonal vectors, one of which is parallel to \mathbf{a} :

$$\mathbf{b} = \mathbf{b}_\perp + \mathbf{b}_\parallel, \quad \mathbf{b}_\perp = \mathbf{b} - s\mathbf{a}, \quad \mathbf{b}_\parallel = s\mathbf{a}, \quad s = \frac{\mathbf{b} \cdot \mathbf{a}}{\|\mathbf{a}\|^2}.$$

Indeed, given \mathbf{a} and \mathbf{b} , put $\mathbf{b}_\perp = \mathbf{b} - s\mathbf{a}$ and assume that \mathbf{b}_\perp is orthogonal to \mathbf{a} , that is, $\mathbf{a} \cdot \mathbf{b}_\perp = 0$. This condition uniquely determines the coefficient s : $\mathbf{a} \cdot \mathbf{b} - s\mathbf{a} \cdot \mathbf{a} = 0$ or $s = \mathbf{b} \cdot \mathbf{a} / \|\mathbf{a}\|^2$. The vectors \mathbf{b}_\perp and \mathbf{b}_\parallel are called the *orthogonal* and *parallel* components of \mathbf{b} relative to the vector \mathbf{a} . The vector \mathbf{b}_\parallel is also called a *vector projection* of \mathbf{b} onto \mathbf{a} . The orthogonal decomposition $\mathbf{b} = \mathbf{b}_\perp + \mathbf{b}_\parallel$ is shown in Figure 11.10 (right panel). If $\hat{\mathbf{a}} = \mathbf{a} / \|\mathbf{a}\|$ is the unit vector along \mathbf{a} , then $\mathbf{b}_\parallel = b_\parallel \hat{\mathbf{a}}$, where the coefficient $b_\parallel = \mathbf{a} \cdot \mathbf{b} / \|\mathbf{a}\| = \|\mathbf{b}\| \cos \theta$ is called a *scalar projection* of \mathbf{b} onto \mathbf{a} . It is also easy to see from the figure that $\|\mathbf{b}_\perp\| = \|\mathbf{b}\| \sin \theta$.

EXAMPLE 11.8. Let $\mathbf{a} = \langle 1, -2, 1 \rangle$ and $\mathbf{b} = \langle 5, 1, 9 \rangle$. Find the orthogonal decomposition $\mathbf{b} = \mathbf{b}_\perp + \mathbf{b}_\parallel$ relative to the vector \mathbf{a} .

SOLUTION: One has $\mathbf{a} \cdot \mathbf{b} = 5 - 2 + 9 = 12$ and $\|\mathbf{a}\|^2 = \mathbf{a} \cdot \mathbf{a} = 1 + (-2)^2 + 1 = 6$. Therefore, $s = 12/6 = 2$, $\mathbf{b}_\parallel = s\mathbf{a} = 2\langle 1, -2, 1 \rangle = \langle 2, -4, 2 \rangle$, and $\mathbf{b}_\perp = \mathbf{b} - \mathbf{b}_\parallel = \langle 5, 1, 9 \rangle - \langle 2, -4, 2 \rangle = \langle 3, 5, 7 \rangle$. The result can also be verified: $\mathbf{a} \cdot \mathbf{b}_\perp = 3 - 10 + 7 = 0$; that is, \mathbf{a} is orthogonal to \mathbf{b}_\perp as required. \square

THEOREM 11.2. (Cauchy-Schwarz Inequality).

For any two vectors \mathbf{a} and \mathbf{b} ,

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

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 $|\cos \theta| \leq 1$. The equality is reached only when $\theta = 0$ or $\theta = \pi$, that is, when \mathbf{a} and \mathbf{b} are parallel or antiparallel.

THEOREM 11.3. (Triangle Inequality).

For any two vectors \mathbf{a} and \mathbf{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. \square

The triangle inequality has a simple geometrical meaning. Consider a triangle with sides \mathbf{a} , \mathbf{b} , and \mathbf{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 \mathbf{a} and \mathbf{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 $|\|\mathbf{a}\| - \|\mathbf{b}\||$. The absolute value is necessary as the length of \mathbf{a} may be less than the length of \mathbf{b} . This observation can be stated in the following algebraic form:

$$(11.7) \quad \left| \|\mathbf{a}\| - \|\mathbf{b}\| \right| \leq \|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|.$$

73.3. 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 rectangular box whose largest diagonal coincides with the vector \mathbf{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 \mathbf{a} . In other words, the angle α is the angle between $\hat{\mathbf{e}}_1$ and \mathbf{a} . Similarly, the angles β and

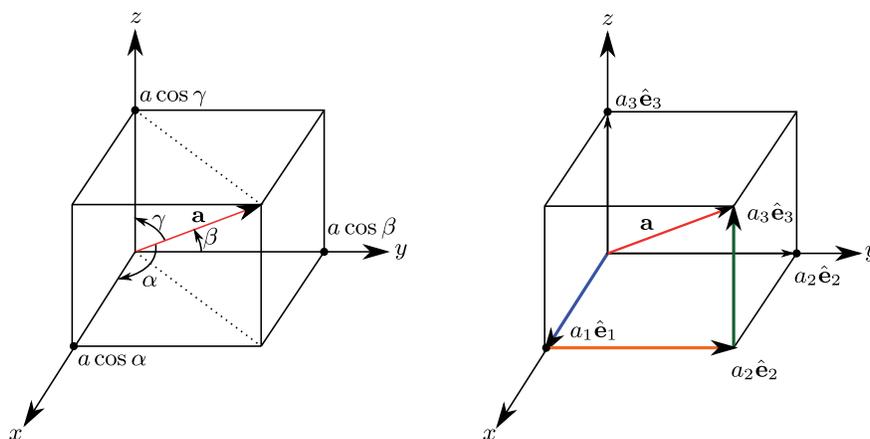


FIGURE 11.9. **Left:** The direction angles of a vector are defined as the angles between the vector and three coordinate axes. Each angle ranges between 0 and π and is counted from the corresponding positive coordinate semiaxis toward the vector. The cosines of the direction angles of a vector are components of the unit vector parallel to that vector. **Right:** The decomposition of a vector \mathbf{a} 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 rectangular box whose edges are formed by the vectors in the sum.

γ are, by definition, the angles between \mathbf{a} and the unit vectors $\hat{\mathbf{e}}_2$ and $\hat{\mathbf{e}}_3$, respectively. Then

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

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 $a_3 = a \cos \gamma$.

73.4. Basis Vectors in Space. A collection of all ordered triples of real numbers $\langle a_1, a_2, a_3 \rangle$ in which the addition, the multiplication by a number, the dot product, and the norm are defined as in vector algebra is also called a three-dimensional *Euclidean space*. Similarly, a collection of ordered doublets of real numbers is a two-dimensional Euclidean space. As noted, any element of a three-dimensional Euclidean space can be uniquely represented as a linear combination of three particular elements $\hat{\mathbf{e}}_1 = \langle 1, 0, 0 \rangle$, $\hat{\mathbf{e}}_2 = \langle 0, 1, 0 \rangle$, and $\hat{\mathbf{e}}_3 = \langle 0, 0, 1 \rangle$. They are called the *standard basis*. There are other triples of vectors with the characteristic property that any vector is a unique linear combination of them. Given any three mutually orthogonal unit vectors $\hat{\mathbf{u}}_i$, $i = 1, 2, 3$, any vector in space can be *uniquely* expanded into the sum $\mathbf{a} = a_1 \hat{\mathbf{u}}_1 + a_2 \hat{\mathbf{u}}_2 + a_3 \hat{\mathbf{u}}_3$, where the numbers a_i are the scalar projections of \mathbf{a} onto $\hat{\mathbf{u}}_i$. Any such triple of vectors is called an *orthonormal basis* in space. So with any orthonormal basis one can associate a rectangular coordinate system in which the coordinates of a point are given by the scalar projections of its position vector onto the basis vectors.

DEFINITION 11.9. (Basis in Space).

A triple of vectors \mathbf{u}_1 , \mathbf{u}_2 , and \mathbf{u}_3 is called a basis in space if any vector \mathbf{a} can be uniquely represented as a linear combination of them: $\mathbf{a} = a_1 \mathbf{u}_1 + a_2 \mathbf{u}_2 + a_3 \mathbf{u}_3$.

A basis may contain vectors that are not necessarily orthogonal or unit. For example, a vector in a plane is a *unique* linear combination of two given nonparallel vectors in the plane (Study Problem 11.6). In this sense, any two nonparallel vectors in a plane define a (nonorthogonal) basis in a plane. Consider three vectors in space. If one of them is a linear combination of the others, then the vectors are in one plane and called *coplanar*. Suppose that none of the vectors is a linear combination of the other two; that is, they are not coplanar. Such vectors are called *linearly independent*. Thus, *three vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} are linearly independent if and only if the vector equation $x\mathbf{a} + y\mathbf{b} + z\mathbf{c} = \mathbf{0}$ has only a trivial solution $x = y = z = 0$ because, otherwise, one of the vectors is a linear combination of the others. For example, if $x \neq 0$, then $\mathbf{a} = -(y/x)\mathbf{b} - (z/x)\mathbf{c}$. It can be proved that *any three linearly independent vectors form a basis in space* (Study Problems 11.11 and 11.12).*

73.5. Applications of the Dot Product.

Static Problems. According to Newton's mechanics, a pointlike object that is 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. A system of objects is at rest if all its elements are at rest. Thus, *for any element of a system at rest, the scalar projection of the total force onto any vector vanishes.* In particular, the components of the total force should vanish in *any orthonormal basis* or, as a point of fact, they vanish in any basis in space (see Study Problem 11.11). This principle is used 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 11.9. *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: The system in question is shown in Figure 11.10 (left panel). The equilibrium condition is

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

Let $\hat{\mathbf{e}}_1$ be a unit vector that is horizontal and directed from left to right and let $\hat{\mathbf{e}}_2$ be a unit vector directed upward. They form an orthonormal basis in the plane. Using the scalar projections, the forces can be expanded in this basis as

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

where T_1 and T_2 are the magnitudes of the tension forces. The scalar projections of the total force onto the horizontal and vertical directions defined by $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$ should vanish:

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

This system is then 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 $T_1 = mg \cos \theta_2 / \sin(\theta_1 + \theta_2)$ is obtained. \square

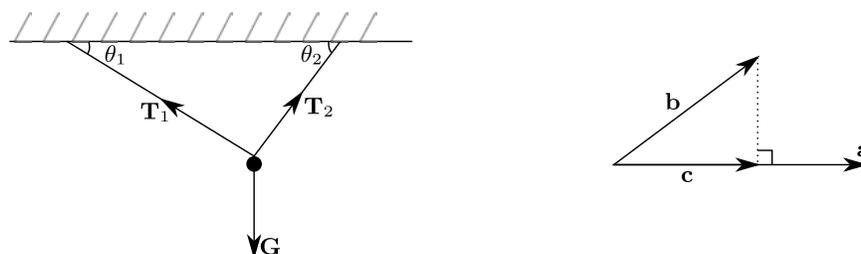


FIGURE 11.10. **Left:** Illustration to Example 11.9. At equilibrium, the vector sum of all forces acting on the ball vanishes. The components of the forces are easy to find in the coordinate system in which the x axis is horizontal and the y axis is vertical. **Right:** The vector \mathbf{c} is the vector projection of a vector \mathbf{b} onto \mathbf{a} . The line through the terminal points of \mathbf{b} and \mathbf{c} is perpendicular to \mathbf{a} . The scalar projection of \mathbf{b} onto \mathbf{a} is $\|\mathbf{b}\| \cos \theta$, where θ is the angle between \mathbf{a} and \mathbf{b} . It is positive if $\theta < \pi/2$, vanishes if $\theta = \pi/2$, or is negative if $\theta > \pi/2$.

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 11.10. *Let an object slide on an inclined plane without friction under the gravitational force. The magnitude of the gravitational force is equal to mg , where m is the mass of the object and g is a universal constant for all objects near the surface of the Earth, $g \approx 9.8 \text{ m/s}^2$. 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 $\overrightarrow{P_1P_2} = \langle L, -h, 0 \rangle$. Since $K_1 = 0$, one has

$$\begin{aligned} \frac{mv^2}{2} &= W = \mathbf{F} \cdot \overrightarrow{P_1P_2} \\ &= \langle 0, -mg, 0 \rangle \cdot \langle L, -h, 0 \rangle = mgh \quad \Rightarrow \quad v = \sqrt{2gh}. \end{aligned}$$

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. \square

73.6. Study Problems.

Problem 11.11. (General Basis in Space).

Let \mathbf{u}_i , $i = 1, 2, 3$, be three linearly independent (non-coplanar) vectors. Show that they form a basis in space; that is, any vector \mathbf{a} can be uniquely expanded into the sum

$$\mathbf{a} = s_1\mathbf{u}_1 + s_2\mathbf{u}_2 + s_3\mathbf{u}_3.$$

The numbers s_i are called components of \mathbf{a} relative to the basis \mathbf{u}_i .

SOLUTION: A solution employs the same approach as in the solution of Study Problem 11.6. Let P_1 be the parallelogram with adjacent sides \mathbf{u}_2 and \mathbf{u}_3 , and P_2 be the parallelogram with sides \mathbf{u}_1 and \mathbf{u}_3 , and P_3 be the parallelogram with sides \mathbf{u}_1 and \mathbf{u}_2 . Consider a box whose faces are the parallelograms P_1 , P_2 , and P_3 . This box is called a *parallelepiped*. By the rules of vector algebra, the largest diagonal of the parallelepiped is the sum $\mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3$. Let the vectors \mathbf{u}_i and a vector \mathbf{a} have common initial point. Consider three planes through the terminal point of \mathbf{a} that are parallel to the parallelograms P_1 , P_2 , P_3 and similar planes through the initial point of \mathbf{a} . These six planes enclose a parallelepiped whose largest diagonal is the vector \mathbf{a} and whose adjacent sides are *parallel* to the vectors \mathbf{u}_i and therefore are proportional to them; that is, the adjacent edges are the vectors $s_1\mathbf{u}_1$, $s_2\mathbf{u}_2$, and $s_3\mathbf{u}_3$, where the numbers s_1 , s_2 , and s_3 are uniquely determined by the proposed construction of the parallelepiped. Hence, $\mathbf{a} = s_1\mathbf{u}_1 + s_2\mathbf{u}_2 + s_3\mathbf{u}_3$. Note that the same geometrical construction has been used to expand a vector in an *orthonormal* basis $\hat{\mathbf{e}}_i$ as shown in Figure 11.9. \square

Problem 11.12. Let $\mathbf{u}_1 = \langle 1, 1, 0 \rangle$, $\mathbf{u}_2 = \langle 1, 0, 1 \rangle$, and $\mathbf{u}_3 = \langle 2, 2, 1 \rangle$. Show that these vectors are linearly independent and hence form a basis in space. Find the components of $\mathbf{a} = \langle 1, 2, 3 \rangle$ in this basis.

SOLUTION: If the vectors \mathbf{u}_i are not linearly independent, then there should exist numbers c_1 , c_2 , and c_3 that do not simultaneously vanish such that

$$c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + c_3\mathbf{u}_3 = \mathbf{0}.$$

Indeed, this algebraic condition means that one of the vectors is a linear combination of the other two whenever c_i do not vanish simultaneously. This vector equation can be written in the components:

$$\begin{cases} c_1 + c_2 + 2c_3 = 0 \\ c_1 + 2c_3 = 0 \\ c_2 + c_3 = 0 \end{cases} \iff \begin{cases} c_1 + c_2 + 3c_3 = 0 \\ c_1 = -2c_3 \\ c_2 = -c_3 \end{cases}.$$

The substitution of the last two equations into the first one yields $-c_3 - 2c_3 + 2c_3 = 0$ or $c_3 = 0$ and hence $c_1 = c_2 = 0$. Thus, the vectors \mathbf{u}_i are linearly independent and form a basis in space. For any vector $\mathbf{a} = s_1\mathbf{u}_1 + s_2\mathbf{u}_2 + s_3\mathbf{u}_3$, the numbers s_i , $i = 1, 2, 3$, are components of \mathbf{a} in the basis \mathbf{u}_i . By writing this vector equation in components relative to the standard basis for $\mathbf{a} = \langle 1, 2, 3 \rangle$, the system of equations is obtained:

$$\begin{cases} s_1 + s_2 + 2s_3 = 1 \\ s_1 + 2s_3 = 2 \\ s_2 + s_3 = 3 \end{cases} \iff \begin{cases} s_1 + s_2 + 2s_3 = 1 \\ s_1 = 2 - 2s_3 \\ s_2 = 3 - s_3 \end{cases}.$$

The substitution of the last two equations into the first one yields $s_3 = 4$ and hence $s_1 = -6$ and $s_2 = -1$ so that $\mathbf{a} = -6\mathbf{u}_1 - \mathbf{u}_2 + 4\mathbf{u}_3$. \square

Problem 11.13. Describe the set of all points in space whose position vectors \mathbf{r} satisfy 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,

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

Hence, the set is a sphere of radius $R = \|\mathbf{R}\|$, and its center is positioned at \mathbf{c} . If \mathbf{a} and \mathbf{b} are the position vectors of points A and B , then, by the parallelogram rule, the center of the sphere is the midpoint of the

straight line segment AB and the segment AB is a diameter of the sphere. \square

73.7. 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 $\mathbf{b} = 3\hat{\mathbf{e}}_1 - 2\hat{\mathbf{e}}_2 + \hat{\mathbf{e}}_3$
 - (iii) $\mathbf{a} = 2\mathbf{c} - 3\mathbf{d}$ and $\mathbf{b} = \mathbf{c} + 2\mathbf{d}$ if \mathbf{c} is the unit vector that makes the angle $\pi/3$ with the vector \mathbf{d} and $\|\mathbf{d}\| = 2$
- (2) For what values of b are the vectors $\langle -6, b, 5 \rangle$ and $\langle b, b, 1 \rangle$ 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) Consider a triangle whose one side is a diameter of a circle and the vertex opposite to this side is on the circle. Use vector algebra to prove that any such triangle is right-angled. *Hint:* Consider position vectors of the vertices relative to the center of the circle.
- (6) Let $\mathbf{a} = s\hat{\mathbf{u}} + \hat{\mathbf{v}}$ and $\mathbf{b} = \hat{\mathbf{u}} + s\hat{\mathbf{v}}$, where the angle between unit vectors $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ is $\pi/3$. Find the values of s for which the dot product $\mathbf{a} \cdot \mathbf{b}$ is maximal, or minimal, or vanishes. If such values exist, graph the vectors \mathbf{a} and \mathbf{b} for these values of s .
- (7) Consider a cube whose edges have length a . Find the angle between its largest diagonal and any edge adjacent to the diagonal.
- (8) Consider a parallelepiped with adjacent sides $\mathbf{a} = \langle 1, -2, 2 \rangle$, $\mathbf{b} = \langle 2, 2, -1 \rangle$, and $\mathbf{c} = \langle 1, 1, 1 \rangle$ (see the definition of a parallelepiped in Study Problem 11.11). Find the angles between its largest diagonal and the adjacent sides.
- (9) Let $\mathbf{a} = \langle 1, 2, 2 \rangle$. For the vector $\mathbf{b} = \langle -2, 3, 1 \rangle$, find the scalar and vector projections of \mathbf{b} onto \mathbf{a} and construct the orthogonal decomposition $\mathbf{b} = \mathbf{b}_\perp + \mathbf{b}_\parallel$ relative to \mathbf{a} .
- (10) Find all vectors that have a given length a and make the angle $\pi/3$ with the positive x axis and the angle $\pi/4$ with the positive z axis.
- (11) 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.
- (12) If $\mathbf{c} = \|\mathbf{a}\|\mathbf{b} + \|\mathbf{b}\|\mathbf{a}$, where \mathbf{a} and \mathbf{b} are nonzero vectors, show that \mathbf{c} bisects the angle between \mathbf{a} and \mathbf{b} .

(13) 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.

(14) 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 \mathbf{a} and \mathbf{b} that are adjacent sides of the parallelogram and express the diagonals via \mathbf{a} and \mathbf{b} . Use the dot product to evaluate $d_1^2 + d_2^2$.

(15) Consider a right-angled triangle whose adjacent sides at the right angle have lengths a and b . Let P be a point in space at a distance c from all three vertices of the triangle ($c \geq a/2$ and $c \geq b/2$). Find the angles between the line segments connecting P with the vertices of the triangle. *Hint:* Consider vectors with the initial point P and terminal points at the vertices of the triangle.

(16) Show that the vectors $\mathbf{u}_1 = \langle 1, 1, 2 \rangle$, $\mathbf{u}_2 = \langle 1, -1, 0 \rangle$, and $\mathbf{u}_3 = \langle 2, 2, -2 \rangle$ are mutually orthogonal. For a vector $\mathbf{a} = \langle 4, 3, 4 \rangle$, find the scalar orthogonal projections of \mathbf{a} onto \mathbf{u}_i , $i = 1, 2, 3$, and the numbers s_i such that $\mathbf{a} = s_1\mathbf{u}_1 + s_2\mathbf{u}_2 + s_3\mathbf{u}_3$.

(17) A point object traveled 3 meters from a point A in a particular direction, then it changed the direction by 60° and traveled 4 meters, and then it changed the direction again so that it was traveling at 60° with each of the previous two directions. If the last stretch was 2 meters long, how far from A is the object?

(18) Two balls of the same mass m 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 and the magnitude of tension forces in the ropes.

(19) A ball of mass m is attached by three ropes of the same length a to a horizontal ceiling so that the attachment points on the ceiling form a triangle with sides of length a . Find the magnitude of the tension force in the ropes.

(20) Four dogs are at the vertices of a square. Each dog starts running toward its neighbor on the right. The dogs run with the same speed v . At every moment of time, each dog keeps running in the direction of its right neighbor (its velocity vector always points to the neighbor). Eventually, the dogs meet in the center of the square. When will this happen if the sides of the square have length a ? What is the distance traveled by each dog? *Hint:* Is there a particular direction relative to which the velocity vector of a dog has the same component at each moment of time?

74. The Cross Product

74.1. Determinant of a Square Matrix.

DEFINITION 11.10. *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 11.11. *The determinant of a 3×3 matrix A is the number obtained by the following rule:*

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

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

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 recursively to $N \times N$ square matrices by letting k and m range over $1, 2, \dots, N$.

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*. For 2×2 matrices, this is easy to see directly from Definition 11.10. In general, if the matrix B is obtained from A by swapping the first and second rows, that is, $b_{1k} = a_{2k}$ and $b_{2k} = a_{1k}$, then the matrices B_{2k} and A_{1k} coincide and so do their determinants. By expanding $\det B$ over its *second* row $b_{2k} = a_{1k}$, one infers that

$$\begin{aligned} \det B &= \sum_{k=1}^3 (-1)^{2+k} b_{2k} \det B_{2k} = \sum_{k=1}^3 (-1)^{2+k} a_{1k} \det A_{1k} \\ &= - \sum_{k=1}^3 (-1)^{1+k} a_{1k} \det A_{1k} = - \det A. \end{aligned}$$

This argument can be applied to any two rows or columns in a square matrix of any dimension.

EXAMPLE 11.11. 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. \square

74.2. The Cross Product of Two Vectors.

DEFINITION 11.12. (Cross Product).

Let $\hat{\mathbf{e}}_1 = \langle 1, 0, 0 \rangle$, $\hat{\mathbf{e}}_2 = \langle 0, 1, 0 \rangle$, and $\hat{\mathbf{e}}_3 = \langle 0, 0, 1 \rangle$. 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:

$$\begin{aligned}
 \mathbf{a} \times \mathbf{b} &= \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} \\
 &= \det \begin{pmatrix} a_2 & a_3 \\ b_2 & b_3 \end{pmatrix} \hat{\mathbf{e}}_1 - \det \begin{pmatrix} a_1 & a_3 \\ b_1 & b_3 \end{pmatrix} \hat{\mathbf{e}}_2 + \det \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix} \hat{\mathbf{e}}_3 \\
 (11.8) \quad &= \langle a_2b_3 - a_3b_2, a_3b_1 - a_1b_3, a_1b_2 - a_2b_1 \rangle.
 \end{aligned}$$

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.

EXAMPLE 11.12. Evaluate the cross product $\mathbf{a} \times \mathbf{b}$ if $\mathbf{a} = \langle 1, 2, 3 \rangle$ and $\mathbf{b} = \langle 2, 0, 1 \rangle$.

SOLUTION: By definition,

$$\begin{aligned}
 \langle 1, 2, 3 \rangle \times \langle 2, 0, 1 \rangle &= \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ 1 & 2 & 3 \\ 2 & 0 & 1 \end{pmatrix} \\
 &= \det \begin{pmatrix} 2 & 3 \\ 0 & 1 \end{pmatrix} \hat{\mathbf{e}}_1 - \det \begin{pmatrix} 1 & 3 \\ 2 & 1 \end{pmatrix} \hat{\mathbf{e}}_2 + \det \begin{pmatrix} 1 & 2 \\ 2 & 0 \end{pmatrix} \hat{\mathbf{e}}_3 \\
 &= (2 - 0)\hat{\mathbf{e}}_1 - (1 - 6)\hat{\mathbf{e}}_2 + (0 - 4)\hat{\mathbf{e}}_3 \\
 &= 2\hat{\mathbf{e}}_1 + 5\hat{\mathbf{e}}_2 - 4\hat{\mathbf{e}}_3 = \langle 2, 5, -4 \rangle.
 \end{aligned}$$

□

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). Alternatively, recall that the determinant of a matrix changes its sign if two rows are swapped in the matrix (the rows \mathbf{a} and \mathbf{b} in Definition 11.12). So the cross product is skew-symmetric; that is, it is *not commutative*, and the order in which the vectors are multiplied is essential. Changing the order leads to the opposite vector. In particular, if $\mathbf{b} = \mathbf{a}$, then $\mathbf{a} \times \mathbf{a} = -\mathbf{a} \times \mathbf{a}$ or $2(\mathbf{a} \times \mathbf{a}) = \mathbf{0}$ or

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

The cross product is *distributive* according to the second property. To prove this 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). The double cross product satisfies the so called *bac – cab* rule

$$(11.9) \quad \mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b}(\mathbf{a} \cdot \mathbf{c}) - \mathbf{c}(\mathbf{a} \cdot \mathbf{b})$$

and the Jacobi identity

$$(11.10) \quad \mathbf{a} \times (\mathbf{b} \times \mathbf{c}) + \mathbf{b} \times (\mathbf{c} \times \mathbf{a}) + \mathbf{c} \times (\mathbf{a} \times \mathbf{b}) = \mathbf{0}.$$

Note that the second and third terms on the left side of (11.10) are obtained from the first by cyclic permutations of the vectors. The proofs of the *bac – cab* rule and the Jacobi identity are given in Study Problems 11.16 and 11.17. The Jacobi identity implies that

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

This means that the multiplication of vectors defined by the cross product is *not associative* in contrast to multiplication of numbers. This observation is further discussed in Study Problem 11.18.

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} :

$$(11.11) \quad \mathbf{a} \cdot (\mathbf{a} \times \mathbf{b}) = \mathbf{b} \cdot (\mathbf{a} \times \mathbf{b}) = 0 \quad \iff \quad \mathbf{a} \times \mathbf{b} \perp \mathbf{a} \text{ and } \mathbf{a} \times \mathbf{b} \perp \mathbf{b}.$$

Let us calculate the length of the cross product. By the definition (11.8),

$$\begin{aligned}\|\mathbf{a} \times \mathbf{b}\|^2 &= (\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{a} \times \mathbf{b}) \\ &= (a_2b_3 - a_3b_2)^2 + (a_3b_1 - a_1b_3)^2 + (a_1b_2 - a_2b_1)^2 \\ &= (a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2) - (a_1b_1 + a_2b_2 + a_3b_3)^2 \\ &= \|\mathbf{a}\|^2\|\mathbf{b}\|^2 - (\mathbf{a} \cdot \mathbf{b})^2,\end{aligned}$$

where the third equality is obtained by computing the squares of the components of the cross product and regrouping terms in the obtained expression. The last equality uses the definitions of the norm and the dot product. Next, recall the geometrical property of the dot product (11.6). If θ is the angle between the vectors \mathbf{a} and \mathbf{b} , then

$$\begin{aligned}\|\mathbf{a} \times \mathbf{b}\|^2 &= \|\mathbf{a}\|^2\|\mathbf{b}\|^2 - \|\mathbf{a}\|^2\|\mathbf{b}\|^2 \cos^2 \theta \\ &= \|\mathbf{a}\|^2\|\mathbf{b}\|^2(1 - \cos^2 \theta) = \|\mathbf{a}\|^2\|\mathbf{b}\|^2 \sin^2 \theta.\end{aligned}$$

Since $0 \leq \theta \leq \pi$, $\sin \theta \geq 0$ and the square root of both sides of this equation can be taken with the result that

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

This relation shows that the length of the cross product defined by (11.8) does not depend on the choice of the coordinate system as it is expressed via the geometrical invariants, the lengths of \mathbf{a} and \mathbf{b} and the angle between them. 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. Therefore, the norm of the cross product, $\|\mathbf{a} \times \mathbf{b}\| = \|\mathbf{a}\|h = A$, is the area of the parallelogram.

Owing to the mutual orientation of the vectors \mathbf{a} , \mathbf{b} , and $\mathbf{a} \times \mathbf{b} \neq \mathbf{0}$ established in (11.11) as well as that their lengths are preserved under rotations of the coordinate system, the coordinate system can be oriented so that \mathbf{a} is along the x axis, \mathbf{b} is in the xy plane, while $\mathbf{a} \times \mathbf{b}$ is parallel to the z axis. In this coordinate system, $\mathbf{a} = \langle \|\mathbf{a}\|, 0, 0 \rangle$ and $\mathbf{b} = \langle b_1, b_2, 0 \rangle$, where $b_1 = \|\mathbf{b}\| \cos \theta$ and $b_2 = \|\mathbf{b}\| \sin \theta$ if \mathbf{b} lies in either the first or second quadrant of the xy plane and $b_2 = -\|\mathbf{b}\| \sin \theta$ if \mathbf{b} lies in either the third or fourth quadrant. In the former case, $\mathbf{a} \times \mathbf{b} = \langle 0, 0, A \rangle$, where A is the area of the parallelogram. In the latter case, the definition (11.8) yields $\mathbf{a} \times \mathbf{b} = \langle 0, 0, -A \rangle$. It turns out that the direction of the cross product in both cases can be described by a simple rule 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 $\mathbf{a} \times \mathbf{b}$.*

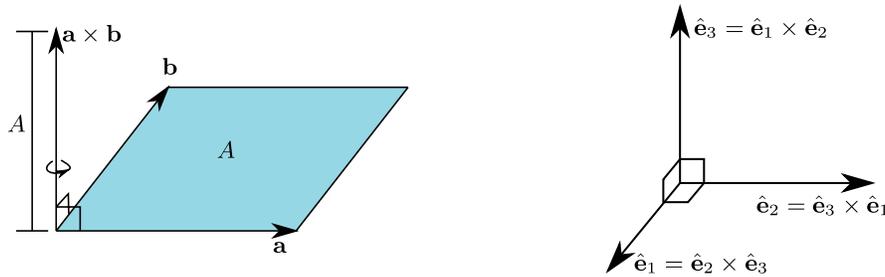


FIGURE 11.11. **Left:** Geometrical interpretation of the cross product of two vectors. The cross product is a vector that is perpendicular to both vectors in the product. Its length equals the area of the parallelogram whose adjacent sides are the vectors in the product. If the fingers of the right hand curl in the direction of a rotation from the first vector to the 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.

In particular, by Definition 11.12, $\hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_2 = \langle 1, 0, 0 \rangle \times \langle 0, 1, 0 \rangle = \langle 0, 0, 1 \rangle = \hat{\mathbf{e}}_3$. If \mathbf{a} is orthogonal to \mathbf{b} , then the relative orientation of the triple of vectors \mathbf{a} , \mathbf{b} , and $\mathbf{a} \times \mathbf{b}$ is the same as that of the standard basis vectors $\hat{\mathbf{e}}_1$, $\hat{\mathbf{e}}_2$, and $\hat{\mathbf{e}}_3$.

The stated geometrical properties are depicted in the left panel of Figure 11.11 and summarized in the following theorem.

THEOREM 11.4. (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.

Two useful consequences can be deduced from this theorem.

COROLLARY 11.2. *Two nonzero vectors are parallel if and only if their cross product vanishes:*

$$\mathbf{a} \times \mathbf{b} = \mathbf{0} \iff \mathbf{a} \parallel \mathbf{b}.$$

If $\mathbf{a} \times \mathbf{b} = \mathbf{0}$, then the area of the corresponding parallelogram vanishes, $\|\mathbf{a} \times \mathbf{b}\| = 0$, which is only possible if the adjacent sides of the parallelogram are parallel. Conversely, for two parallel vectors, there is a number s such that $\mathbf{a} = s\mathbf{b}$. Hence, $\mathbf{a} \times \mathbf{b} = (s\mathbf{b}) \times \mathbf{b} = s(\mathbf{b} \times \mathbf{b}) = \mathbf{0}$.

If in the cross product $\mathbf{a} \times \mathbf{b}$ the vector \mathbf{b} is changed by adding to it any vector parallel to \mathbf{a} , the cross product does not change:

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

Let $\mathbf{b} = \mathbf{b}_\perp + \mathbf{b}_\parallel$ be the orthogonal decomposition of \mathbf{b} relative to a nonzero vector \mathbf{a} . By Corollary 11.1, $\mathbf{a} \times \mathbf{b}_\parallel = \mathbf{0}$ because \mathbf{b}_\parallel is parallel to \mathbf{a} . It is then concluded that *the cross product depends only on the component \mathbf{b}_\perp of \mathbf{b} that is orthogonal to \mathbf{a}* . Thus, $\mathbf{a} \times \mathbf{b} = \mathbf{a} \times \mathbf{b}_\perp$ and $\|\mathbf{a} \times \mathbf{b}\| = \|\mathbf{a}\| \|\mathbf{b}_\perp\|$.

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

COROLLARY 11.3. (Area of a Triangle).

Let vectors \mathbf{a} and \mathbf{b} be two sides of a triangle that have the same initial point at a vertex of a 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 11.13. *Let $A = (1, 1, 1)$, $B = (2, -1, 3)$, and $C = (-1, 3, 1)$. Find the area of the triangle ABC and a vector orthogonal to the plane that contains the triangle.*

SOLUTION: Take two vectors with the initial point at any of the vertices of the triangle that form the adjacent sides of the triangle at that vertex. For example, $\mathbf{a} = \overrightarrow{AB} = \langle 1, -2, 2 \rangle$ and $\mathbf{b} = \overrightarrow{AC} = \langle -2, 2, 0 \rangle$. Then $\mathbf{a} \times \mathbf{b} = \langle -4, -4, -6 \rangle$. 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 11.3. 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). Any vector in the plane that contains the triangle is a linear combination of \mathbf{a} and \mathbf{b} . Therefore, the vector $\mathbf{a} \times \mathbf{b}$ is orthogonal to any such vector and hence to the plane because $\mathbf{a} \times \mathbf{b}$ is orthogonal to both \mathbf{a} and \mathbf{b} . The choice $\mathbf{a} = \overrightarrow{CB}$ and $\mathbf{b} = \overrightarrow{CA}$ or $\mathbf{a} = \overrightarrow{BA}$ and $\mathbf{b} = \overrightarrow{BC}$ would give the same answer (modulo the sign change in the cross product). \square

Applications in Physics: Torque. *Torque*, or *moment of force*, is the tendency of a force to rotate an object about an axis or a pivot. Just as a force is a push or a pull, a torque can be thought of as a twist.

If \mathbf{r} is the vector from a pivot point to the point where a force \mathbf{F} is applied, then the torque is defined as the cross product

$$\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F}.$$

The torque depends only on the component \mathbf{F}_\perp of the force that is orthogonal to \mathbf{r} , that is, $\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F}_\perp$. If θ is the angle between \mathbf{r} and \mathbf{F} , then the magnitude of the torque is $\tau = \|\mathbf{r}\| \|\mathbf{F}_\perp\| = rF \sin \theta$; here $r = \|\mathbf{r}\|$ is the distance from the pivot point to the point where the force of magnitude F is applied. One can think of \mathbf{r} as a lever attached to a pivot point and the force \mathbf{F} is applied to the other end of the lever to rotate it about the pivot point. Naturally, the lever would not rotate if the force is parallel to it ($\theta = 0$ or $\theta = \pi$), whereas the maximal rotational effect is created when the force is applied in the direction perpendicular to the lever ($\theta = \pi/2$). The direction of $\boldsymbol{\tau}$ determines the axis about which the lever rotates. By the property of the cross product, this axis is perpendicular to the plane containing the force and position vectors. According to the right-hand rule, the rotation occurs counterclockwise when viewed from the top of the torque vector. When driving a car, a torque is applied to the steering wheel to change the direction of the car. When a bolt is tightened by applying a force to a wrench, the produced turning effect is the torque.

An extended object is said to be *rigid* if the distance between any two of its points remains constant in time regardless of the external forces exerted on it. Let P be a fixed (pivot) point about which a rigid object can rotate. Suppose that the forces \mathbf{F}_i , $i = 1, 2, \dots, n$, are applied to the object at the points whose position vectors relative to the point P are \mathbf{r}_i . The *principle of moments* states that a rigid object does not rotate about the point P if it was initially at rest and the total torque vanishes:

$$\boldsymbol{\tau} = \boldsymbol{\tau}_1 + \boldsymbol{\tau}_2 + \cdots + \boldsymbol{\tau}_n = \mathbf{r}_1 \times \mathbf{F}_1 + \mathbf{r}_2 \times \mathbf{F}_2 + \cdots + \mathbf{r}_n \times \mathbf{F}_n = \mathbf{0}.$$

If, in addition, the total force vanishes, $\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2 + \cdots + \mathbf{F}_n = \mathbf{0}$, then a rigid object remains at rest and will not rotate about any other pivot point. Indeed, suppose that the torque about P vanishes and let \mathbf{r}_0 be a position vector of P relative to another point P_0 . Then the position vectors of the points at which the forces are applied relative to the new pivot point P_0 are $\mathbf{r}_i + \mathbf{r}_0$. The total torque, or the total moment of the forces, about P_0 also vanishes:

$$\begin{aligned} \boldsymbol{\tau}_0 &= (\mathbf{r}_1 + \mathbf{r}_0) \times \mathbf{F}_1 + \cdots + (\mathbf{r}_n + \mathbf{r}_0) \times \mathbf{F}_n \\ &= \mathbf{r}_1 \times \mathbf{F}_1 + \cdots + \mathbf{r}_n \times \mathbf{F}_n + \mathbf{r}_0 \times (\mathbf{F}_1 + \cdots + \mathbf{F}_n) \\ &= \boldsymbol{\tau} + \mathbf{r}_0 \times \mathbf{F} = \mathbf{0} \end{aligned}$$

because, by the hypothesis, $\boldsymbol{\tau} = \mathbf{0}$ and $\mathbf{F} = \mathbf{0}$. The conditions $\boldsymbol{\tau} = \mathbf{0}$ and $\mathbf{F} = \mathbf{0}$ comprise the fundamental law of statics for rigid objects.

EXAMPLE 11.14. The ends of rigid rods of length L_1 and L_2 are rigidly joined at the angle $\pi/2$. A ball of mass m_1 is attached to the free end of the rod of length L_1 and a ball of mass m_2 is attached to the free end of the rod of length L_2 . The system is hung by the joining point so that the system can rotate freely about it under the gravitational force. Find the equilibrium position of the system if the masses of the rods can be neglected as compared to the masses of the balls.

SOLUTION: The gravitational forces have magnitudes $F_1 = m_1g$ and $F_2 = m_2g$ for the first and second balls, respectively (g is the free fall acceleration). They are directed downward and therefore lie in the plane that contains the position vectors of the balls relative to the pivot point. So the torques of the gravitational forces are orthogonal to this plane, and the equilibrium condition $\boldsymbol{\tau}_1 + \boldsymbol{\tau}_2 = \mathbf{0}$ is equivalent to $\tau_1 - \tau_2 = 0$, where $\tau_{1,2}$ are the magnitudes of the torques. The minus sign follows from the right-hand rule by which the vectors $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ are parallel but have opposite directions. In other words, the gravitational forces applied to the balls generate opposite rotational moments. When the latter are equal in magnitude, the system is at rest. In the plane that contains the system, let θ_1 and θ_2 be the smallest angles between the rods and a horizontal line. Since the rods are perpendicular, $\theta_1 + \theta_2 = \pi/2$. The angle between the position vector of the first ball and the gravitational force acting on it is $\phi_1 = \pi/2 - \theta_1$, and similarly $\phi_2 = \pi/2 - \theta_2$ is the angle between the position vector of the second ball and the gravitational force acting on it. Therefore, $\tau_1 = L_1F_1 \sin \phi_1$ and $\tau_2 = L_2F_2 \sin \phi_2$. Owing to the identity $\sin(\pi/2 - \theta) = \cos \theta$, it follows that

$$\tau_1 = \tau_2 \quad \Leftrightarrow \quad m_1L_1 \cos \theta_1 = m_2L_2 \cos \theta_2 \quad \Leftrightarrow \quad \tan \theta_1 = \frac{m_1L_1}{m_2L_2},$$

where the relation $\theta_2 = \pi/2 - \theta_1$ has been used. \square

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 . \square

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,

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

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 $\{ijk\} \rightarrow \{kij\}$ (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 (11.9).

SOLUTION: If \mathbf{c} and \mathbf{b} are parallel, $\mathbf{b} = s\mathbf{c}$ for some real s , then the relation is true because both its sides vanish. If \mathbf{c} and \mathbf{b} are not parallel, then, by the remark after Corollary 11.2, the double cross product $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$ depends only on the component of \mathbf{a} that is orthogonal to $\mathbf{b} \times \mathbf{c}$. This component lies in the plane containing \mathbf{b} and \mathbf{c} and hence is a linear combination of them (see Study Problem 11.6). So, without loss of generality, $\mathbf{a} = t\mathbf{b} + p\mathbf{c}$. Also, $\mathbf{b} \times \mathbf{c} = \mathbf{b} \times \mathbf{c}_\perp$, where $\mathbf{c}_\perp = \mathbf{c} - s\mathbf{b}$, $s = \mathbf{c} \cdot \mathbf{b} / \|\mathbf{b}\|^2$, is the component of \mathbf{c} orthogonal to \mathbf{b} (note $\mathbf{b} \cdot \mathbf{c}_\perp = 0$). The vectors \mathbf{b} , \mathbf{c}_\perp , and $\mathbf{b} \times \mathbf{c}_\perp$ are mutually orthogonal and oriented according to the right-hand rule. In particular, $\|\mathbf{b} \times \mathbf{c}\| = \|\mathbf{b}\| \|\mathbf{c}_\perp\|$. By applying the right-hand rule twice, it is concluded that $\mathbf{b} \times (\mathbf{b} \times \mathbf{c}_\perp)$ has the direction opposite to \mathbf{c}_\perp . Since \mathbf{b} and $\mathbf{b} \times \mathbf{c}_\perp$ are orthogonal, $\|\mathbf{b} \times (\mathbf{b} \times \mathbf{c}_\perp)\| = \|\mathbf{b}\| \|\mathbf{b} \times \mathbf{c}_\perp\| = \|\mathbf{b}\|^2 \|\mathbf{c}_\perp\|$. Therefore,

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

By swapping the vector \mathbf{b} and \mathbf{c} in this equation, one also obtains

$$\mathbf{c} \times (\mathbf{b} \times \mathbf{c}) = -\mathbf{c} \times (\mathbf{c} \times \mathbf{b}) = \mathbf{b}_\perp \|\mathbf{c}\|^2 = -\mathbf{c}(\mathbf{c} \cdot \mathbf{b}) + \mathbf{b}(\mathbf{c} \cdot \mathbf{c}).$$

It follows from these relations that

$$\begin{aligned} \mathbf{a} \times (\mathbf{b} \times \mathbf{c}) &= t\mathbf{b} \times (\mathbf{b} \times \mathbf{c}) + p\mathbf{c} \times (\mathbf{b} \times \mathbf{c}) \\ &= \mathbf{b}[(t\mathbf{b} + p\mathbf{c}) \cdot \mathbf{c}] - \mathbf{c}[(t\mathbf{b} + p\mathbf{c}) \cdot \mathbf{b}] \\ &= \mathbf{b}(\mathbf{a} \cdot \mathbf{c}) - \mathbf{c}(\mathbf{a} \cdot \mathbf{b}). \end{aligned}$$

□

Problem 11.17. *Prove the Jacobi identity (11.10).*

SOLUTION: By the $bac - cab$ rule (11.9) applied to each term,

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

By adding these equalities, it is easy to see that the coefficients at each of the vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} on the right side are added up to make 0. \square

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 counterclockwise 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, in contrast to the cross product, 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 $(\theta_a + \theta_b) + \theta_c = \theta_a + (\theta_b + \theta_c)$. \square

Remark. The vector product introduced for vectors in a plane is known as the *product of complex numbers*, which can be viewed as two-dimensional vectors. 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 $0 \leq \phi < 2\pi$ 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 $0 \leq \phi < 2\pi$: $\phi = -\tan^{-1}(v_1/v_2)$ and $\phi = -\tan^{-1}(v_1/v_2) + \pi$. Geometrically, these solutions correspond to the case when \mathbf{u} is parallel to the line $v_2y + v_1x = 0$ in the xy plane. \square

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) $\mathbf{a} = \langle 1, -1, 2 \rangle$ and $\mathbf{b} = \langle 3, -2, 1 \rangle$
 - (iii) $\mathbf{a} = \hat{\mathbf{e}}_1 + 3\hat{\mathbf{e}}_2 - \hat{\mathbf{e}}_3$ and $\mathbf{b} = 3\hat{\mathbf{e}}_1 - 2\hat{\mathbf{e}}_2 + \hat{\mathbf{e}}_3$
 - (iv) $\mathbf{a} = 2\mathbf{c} - \mathbf{d}$ and $\mathbf{b} = 3\mathbf{c} + 4\mathbf{d}$, where $\mathbf{c} \times \mathbf{d} = \langle 1, 2, 3 \rangle$
- (2) Let $\mathbf{a} = \langle 3, 2, 1 \rangle$, $\mathbf{b} = \langle -2, 1, -1 \rangle$, and $\mathbf{c} = \langle 1, 0, -1 \rangle$. Find $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$, $\mathbf{b} \times (\mathbf{c} \times \mathbf{a})$, and $\mathbf{c} \times (\mathbf{a} \times \mathbf{b})$.
- (3) Let \mathbf{a} be a unit vector orthogonal to \mathbf{b} and \mathbf{c} . If $\mathbf{c} = \langle 1, 2, 2 \rangle$, find the length of the vector $\mathbf{a} \times [(\mathbf{a} + \mathbf{b}) \times (\mathbf{a} + \mathbf{b} + \mathbf{c})]$.
- (4) Given two nonparallel vectors \mathbf{a} and \mathbf{b} , show that the vectors \mathbf{a} , $\mathbf{a} \times \mathbf{b}$, and $\mathbf{a} \times (\mathbf{a} \times \mathbf{b})$ are mutually orthogonal.
- (5) 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 $\pi/2$.
- (6) 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.
- (7) 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 orthogonal to the plane containing the triangle.
- (8) Use the cross product to show that the area of the triangle whose vertices are midpoints of the sides of a triangle with area A is $A/4$.
- (9) Consider a triangle whose vertices are midpoints of any three sides of a parallelogram. If the area of the parallelogram is A , find the area of the triangle.
- (10) 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.

(11) Consider four points in space. Suppose that the coordinates of the points are known. Describe a procedure based on the properties of the cross product 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?

(12) 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.

(13) 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 , $i = 1, 2, 3$.

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

(15) 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} . Express z via \mathbf{r} , \mathbf{a} , and \mathbf{b} . In particular, put $\mathbf{a} = \langle 1, 1, 1 \rangle$ and $\mathbf{b} = \langle 1, 1, 0 \rangle$. Find the coefficients x , y , and z for $\mathbf{r} = \langle 1, 2, 3 \rangle$. *Hint:* See Study Problems 11.6 and 11.14.

(16) 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 $\mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3 + \mathbf{v}_4 = \mathbf{0}$.

(17) 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 $\mathbf{b} = \mathbf{c}$?

(18) Given two nonparallel vectors \mathbf{a} and \mathbf{b} , construct three mutually orthogonal unit vectors $\hat{\mathbf{u}}_i$, $i = 1, 2, 3$, one of which is parallel to \mathbf{a} . Are such unit vectors unique? In particular, put $\mathbf{a} = \langle 1, 2, 2 \rangle$ and $\mathbf{b} = \langle 1, 0, 2 \rangle$ and find $\hat{\mathbf{u}}_i$.

(19) Let $\hat{\mathbf{u}}_i$, $i = 1, 2, 3$, be an orthonormal basis in space with the property that $\hat{\mathbf{u}}_3 = \hat{\mathbf{u}}_1 \times \hat{\mathbf{u}}_2$. If a_1 , a_2 , and a_3 are the components of vector \mathbf{a} relative to this basis and b_1 , b_2 , and b_3 are the components of \mathbf{b} , show that the components of the cross product $\mathbf{a} \times \mathbf{b}$ can also be

computed by the determinant rule given in Definition 11.12 where $\hat{\mathbf{e}}_i$ are replaced by $\hat{\mathbf{u}}_i$. *Hint:* Use the $bac - cab$ rule to find all pairwise cross products of the basis vectors $\hat{\mathbf{u}}_i$.

(20) Let the angle between the rigid rods in Example 11.14 be $0 < \varphi < \pi$. Find the equilibrium position of the system.

(21) Two rigid rods of the same length are rigidly attached to a ball of mass m so that the angle between the rods is $\pi/2$. A ball of mass $2m$ is attached to one of the free ends of the system. The remaining free end is used to hang the system. Find the angle between the rod connecting the pivot point and the ball of mass m and the vertical axis along which the gravitational force is acting. Assume that the masses of the rods can be neglected as compared to m .

(22) Three rigid rods of the same length are rigidly joined by one end so that the rods lie in a plane and the other end of each rod is free. Let three balls of masses m_1 , m_2 , and m_3 be attached to the free ends of the rods. The system is hung by the joining point and can rotate freely about it. Assume that the masses of the rods can be neglected as compared to the masses of the balls. Find the angles between the rods at which the balls remain in a horizontal plane under gravitational forces acting vertically. Do such angles exist for any masses of the balls?

75. The Triple Product

DEFINITION 11.13. *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. Also, the value of the triple product is invariant under cyclic permutations of vectors in it: $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})$.

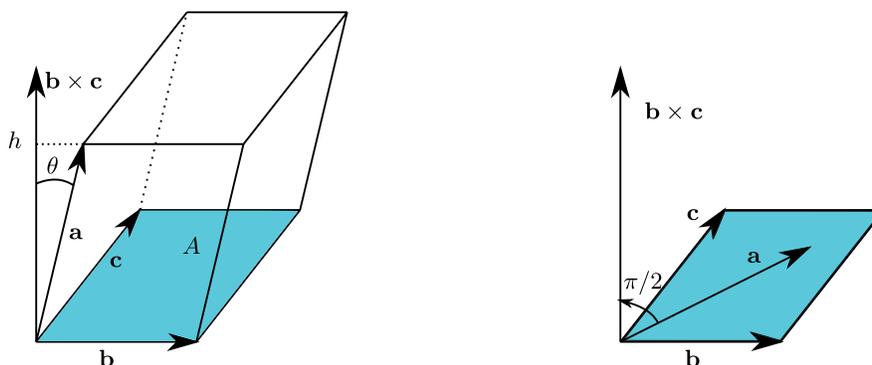


FIGURE 11.12. **Left:** Geometrical interpretation of the triple product as the volume of the parallelepiped whose adjacent sides are the vectors in the product: $h = \|\mathbf{a}\| \cos \theta$, $A = \|\mathbf{b} \times \mathbf{c}\|$, $V = hA = \|\mathbf{a}\| \|\mathbf{b} \times \mathbf{c}\| \cos \theta = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$. **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 parallelograms whose adjacent sides are pairs of the vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} . They enclose a nonrectangular box whose edges are the vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} . A box with parallelogram faces is called a *parallelepiped* with adjacent sides \mathbf{a} , \mathbf{b} , and \mathbf{c} . The cross product $\mathbf{b} \times \mathbf{c}$ is orthogonal 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 11.5. (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. Note also that the vectors can be taken in any order in

the triple product to compute the volume because the triple product only changes its sign when two vectors are swapped in it.

EXAMPLE 11.15. Find the volume of a parallelepiped with adjacent sides $\mathbf{a} = \langle 1, 2, 3 \rangle$, $\mathbf{b} = \langle -2, 0, 1 \rangle$, and $\mathbf{c} = \langle 2, 1, 2 \rangle$.

SOLUTION: The expansion of the determinant over the first row yields

$$\mathbf{b} \cdot (\mathbf{a} \times \mathbf{c}) = \det \begin{pmatrix} -2 & 0 & 1 \\ 1 & 2 & 3 \\ 2 & 1 & 2 \end{pmatrix} = -2(4 - 3) + 1(1 - 4) = -5.$$

Taking the absolute value of the triple product, the volume is obtained, $V = |-5| = 5$. The components of the vectors must be given in the same units of length (e.g., meters). Then the volume is 5 cubic meters. \square

Any vector in a plane is a linear combination of two particular vectors in the plane. Vectors that lie in a plane are called *coplanar* (see Section 73.4). Clearly, any two vectors are always coplanar. However, three nonzero vectors do not generally lie in one plane. None of three non-coplanar vectors can be expressed as a linear combination of the other two vectors. Such vectors are said to be *linearly independent*. As noted in Section 73.4, any three linearly independent vectors form a basis in space. Simple criteria for three vectors to be either coplanar or linearly independent can be deduced from Theorem 11.5.

COROLLARY 11.4. (Criterion for Three Vectors to Be Coplanar).

Three vectors are coplanar if and only if their triple product vanishes:

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

Consequently, three nonzero vectors are linearly independent if and only if their triple product does not vanish.

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 (orthogonal to $\mathbf{b} \times \mathbf{c}$).

In Section 74.3, it was stated that three linearly independent vectors form a basis in space. The linear independence means that none of the vectors is a linear combination of the other two. Geometrically,

this means that the vectors are not coplanar. Therefore, the following simple criterion holds for three vectors to form a basis in space.

COROLLARY 11.5. (Basis in Space).

Three vectors \mathbf{u}_1 , \mathbf{u}_2 , and \mathbf{u}_3 are linearly independent and hence form a basis in space if and only if their triple product does not vanish.

EXAMPLE 11.16. *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 $\mathbf{a} = \overrightarrow{AB} = \langle 1, -1, 1 \rangle$, $\mathbf{b} = \overrightarrow{AC} = \langle 2, 0, 2 \rangle$, and $\mathbf{c} = \overrightarrow{AD} = \langle -1, 1, 2 \rangle$. 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 $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = 0$. The evaluation of the triple product yields

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

Therefore, the points are not in the same plane. \square

EXAMPLE 11.17. *Let $\mathbf{u}_1 = \langle 1, 2, 3 \rangle$, $\mathbf{u}_2 = \langle 2, 1, -6 \rangle$, and $\mathbf{u}_3 = \langle 1, 1, -1 \rangle$. Can the vector $\mathbf{a} = \langle 1, 1, 1 \rangle$ be represented as a linear combination of the vectors \mathbf{u}_1 , \mathbf{u}_2 , and \mathbf{u}_3 ?*

SOLUTION: Any vector in space is a linear combination of \mathbf{u}_1 , \mathbf{u}_2 , and \mathbf{u}_3 if they form a basis. Let us verify first whether or not they form a basis. By Corollary 11.5,

$$\mathbf{u}_1 \cdot (\mathbf{u}_2 \times \mathbf{u}_3) = \det \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & -6 \\ 1 & 1 & -1 \end{pmatrix} = 1(-1 + 6) - 2(-2 + 6) + 3(2 - 1) = 0.$$

Therefore, these vectors do not form a basis and are coplanar. Note that $\mathbf{u}_3 = \frac{1}{3}(\mathbf{u}_1 + \mathbf{u}_2)$. If the vector \mathbf{a} lies in the same plane as the vectors \mathbf{u}_1 , \mathbf{u}_2 , and \mathbf{u}_3 , then it is a linear combination of any two nonparallel vectors, say, \mathbf{u}_1 and \mathbf{u}_3 . Since the following triple product does not vanish,

$$\mathbf{a} \cdot (\mathbf{u}_1 \times \mathbf{u}_3) = \det \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 1 & -1 \end{pmatrix} = 1(-2 - 3) - 1(-1 - 3) + 1(1 - 2) = -2,$$

the vector \mathbf{a} is not in the plane in which the vectors \mathbf{u}_1 , \mathbf{u}_2 , and \mathbf{u}_3 lie, and therefore \mathbf{a} is not a linear combination of them. \square

75.2. Right- and Left-Handed Coordinate Systems Consider a rectangular box whose sides are parallel to three given unit vectors $\hat{\mathbf{e}}_i$, $i = 1, 2, 3$. Any vector in space can be viewed as the diagonal of one such box and therefore is uniquely expanded into the sum $\mathbf{r} = x\hat{\mathbf{e}}_1 + y\hat{\mathbf{e}}_2 + z\hat{\mathbf{e}}_3$, where the ordered triple of numbers (x, y, z) is determined by scalar projections of \mathbf{r} onto $\hat{\mathbf{e}}_i$. Thus, *with any triple of mutually orthogonal unit vectors one can associate a rectangular coordinate system*. The vector $\hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_2$ must be parallel to $\hat{\mathbf{e}}_3$ because the latter is orthogonal to both $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$. Furthermore, owing to the orthogonality of $\hat{\mathbf{e}}_1$, and $\hat{\mathbf{e}}_2$, $\|\hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_2\| = \|\hat{\mathbf{e}}_1\| \|\hat{\mathbf{e}}_2\| = 1$ and hence $\hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_2 = \pm\hat{\mathbf{e}}_3$. Consequently, $\hat{\mathbf{e}}_3 \cdot (\hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_2) = \hat{\mathbf{e}}_1 \cdot (\hat{\mathbf{e}}_2 \times \hat{\mathbf{e}}_3) = \pm 1$ or, owing to the mutual orthogonality of the vectors, $\hat{\mathbf{e}}_2 \times \hat{\mathbf{e}}_3 = \pm\hat{\mathbf{e}}_1$. A coordinate system is called *right-handed* if $\hat{\mathbf{e}}_1 \cdot (\hat{\mathbf{e}}_2 \times \hat{\mathbf{e}}_3) = 1$, and a coordinate system for which $\hat{\mathbf{e}}_1 \cdot (\hat{\mathbf{e}}_2 \times \hat{\mathbf{e}}_3) = -1$ is called *left-handed*. A right-handed system can be visualized as follows. With the thumb, index, and middle fingers of the *right hand* at right angles to each other (with the index finger pointed straight), the middle finger points in the direction of $\hat{\mathbf{e}}_1 = \hat{\mathbf{e}}_2 \times \hat{\mathbf{e}}_3$ when the thumb represents $\hat{\mathbf{e}}_2$ and the index finger represents $\hat{\mathbf{e}}_3$. A left-handed system is obtained by the reflection $\hat{\mathbf{e}}_1 \rightarrow -\hat{\mathbf{e}}_1$ and therefore is visualized by the fingers of the *left hand* in the same way. Since the dot product cannot be changed by rotations and translations in space, the handedness of the coordinate system does not change under simultaneous rotations and translations of the triple of vectors $\hat{\mathbf{e}}_i$ (three mutually orthogonal fingers of the left hand cannot be made pointing in the same direction as the corresponding fingers of the right hand by any rotation of the hand). The reflection of all three vectors $\hat{\mathbf{e}}_i \rightarrow -\hat{\mathbf{e}}_i$ or just one of them turns a right-handed system into a left-handed one and vice versa. A mirror reflection of a right-handed system is the left-handed one. The coordinate system associated with the standard basis $\hat{\mathbf{e}}_1 = \langle 1, 0, 0 \rangle$, $\hat{\mathbf{e}}_2 = \langle 0, 1, 0 \rangle$, and $\hat{\mathbf{e}}_3 = \langle 0, 0, 1 \rangle$ is *right-handed* because $\hat{\mathbf{e}}_1 \cdot (\hat{\mathbf{e}}_2 \times \hat{\mathbf{e}}_3) = 1$.

75.3. 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 (Theorems 11.4 and 11.5). 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.

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 belong 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 $|A_1 A_2|$ 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 $|A_1 A_2|$. Such a number is called the *infimum* of the set of numbers $|A_1 A_2|$ and denoted $\inf |A_1 A_2|$. Thus,

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

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).

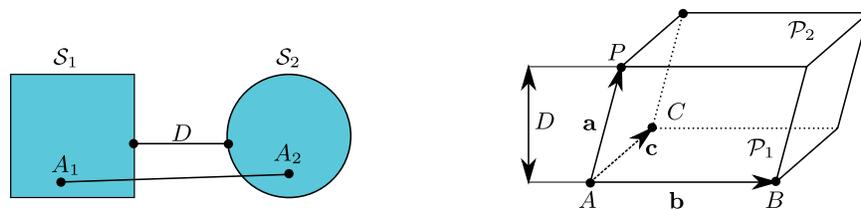


FIGURE 11.13. **Left:** Distance between two point sets \mathcal{S}_1 and \mathcal{S}_2 defined as the largest number that is less than or equal to all distances $|A_1 A_2|$, where A_1 ranges over all points in \mathcal{S}_1 and A_2 ranges over all points in \mathcal{S}_2 . **Right:** Distance between two parallel planes (Corollary 11.6). 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.

COROLLARY 11.6. (Distance Between Parallel Planes).

The distance between parallel planes \mathcal{P}_1 and \mathcal{P}_2 is given by

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

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 $\mathbf{b} = \overrightarrow{AB}$ and $\mathbf{c} = \overrightarrow{AC}$ 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 $\mathbf{a} = \overrightarrow{AP}$, \mathbf{b} , and \mathbf{c} . Two of its faces, the parallelograms with adjacent sides \mathbf{b} and \mathbf{c} , lie in the parallel planes, one in \mathcal{P}_1 and the other in \mathcal{P}_2 . The distance between the planes is, by construction, the height of the parallelepiped, which is equal to V/A_p , where A_p is the area of the face parallel to \mathbf{b} and \mathbf{c} and V is the volume of the parallelepiped. The conclusion follows from the geometrical properties of the triple and cross products: $V = |\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})|$ and $A_p = \|\mathbf{b} \times \mathbf{c}\|$. \square

Similarly, the distance between two parallel lines \mathcal{L}_1 and \mathcal{L}_2 can be determined. Two 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 $\mathbf{a} = \overrightarrow{AB}$ and $\mathbf{b} = \overrightarrow{AC}$ as depicted in Figure 11.14 (left panel). The distance between the lines is the height of this parallelogram, which is $D = A_p/\|\mathbf{a}\|$, where $A_p = \|\mathbf{a} \times \mathbf{b}\|$ is the area of the parallelogram and $\|\mathbf{a}\|$ is the length of its base.

COROLLARY 11.7. (Distance Between Parallel Lines).

The distance between two parallel lines \mathcal{L}_1 and \mathcal{L}_2 is

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

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 .

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 .

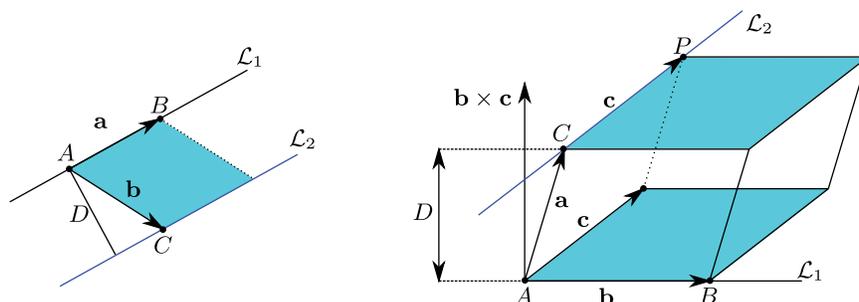


FIGURE 11.14. **Left:** Distance between two parallel lines. Consider a parallelogram whose two parallel sides lie in the lines. Then the distance between the lines is the height of the parallelogram (Corollary 11.7). **Right:** Distance between skew lines. Consider a parallelepiped whose two nonparallel 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 11.8).

Define the vectors $\mathbf{b} = \overrightarrow{AB}$ and $\mathbf{c} = \overrightarrow{CP}$ 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 $\mathbf{a} = \overrightarrow{AC}$, \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 is the distance between the parallel planes containing them. By Corollary 11.6, this distance is $D = V/A_p$, where V and $A_p = \|\mathbf{b} \times \mathbf{c}\|$ are, respectively, the volume of the parallelepiped and the area of its base.

COROLLARY 11.8. (Distance Between Skew Lines).

The distance between two skew lines \mathcal{L}_1 and \mathcal{L}_2 is

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

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 .

As a consequence of the obtained distance formulas, the following criterion for mutual orientation of two lines in space holds.

COROLLARY 11.9. *Let \mathcal{L}_1 be a line through A and $B \neq A$, and let \mathcal{L}_2 be a line through C and $P \neq C$. Then*

- (1) \mathcal{L}_1 and \mathcal{L}_2 are parallel if $\overrightarrow{AB} \times \overrightarrow{CP} = \mathbf{0}$
- (2) \mathcal{L}_1 and \mathcal{L}_2 are skew if $\overrightarrow{AC} \cdot (\overrightarrow{AB} \times \overrightarrow{CP}) \neq 0$
- (3) \mathcal{L}_1 and \mathcal{L}_2 intersect if $\overrightarrow{AC} \cdot (\overrightarrow{AB} \times \overrightarrow{CP}) = 0$ and $\overrightarrow{AB} \times \overrightarrow{CP} \neq \mathbf{0}$
- (4) \mathcal{L}_1 and \mathcal{L}_2 coincide if $\overrightarrow{AC} \cdot (\overrightarrow{AB} \times \overrightarrow{CP}) = 0$ and $\overrightarrow{AB} \times \overrightarrow{CP} = \mathbf{0}$

For parallel lines \mathcal{L}_1 and \mathcal{L}_2 , the vectors \overrightarrow{AB} and \overrightarrow{CP} are parallel, and hence their cross product vanishes. If $\overrightarrow{AC} \cdot (\overrightarrow{AB} \times \overrightarrow{CP}) \neq 0$, then the lines are not parallel and lie in the parallel faces of a parallelepiped that has nonzero volume. Such lines must be skew. If $\overrightarrow{AB} \times \overrightarrow{CP} \neq \mathbf{0}$, the lines are not parallel. The additional condition $\overrightarrow{AC} \cdot (\overrightarrow{AB} \times \overrightarrow{CP}) = 0$ implies that the distance between them vanishes, and hence the lines must intersect at a point. The conditions $\overrightarrow{AC} \cdot (\overrightarrow{AB} \times \overrightarrow{CP}) = 0$ and $\overrightarrow{AB} \times \overrightarrow{CP} = \mathbf{0}$ imply that the lines are parallel and intersecting. So they must coincide.

EXAMPLE 11.18. *Find the distance between the line through the points $A = (1, 1, 2)$ and $B = (1, 2, 3)$ and the line through $C = (1, 0, -1)$ and $P = (-1, 1, 2)$.*

SOLUTION: Let $\mathbf{a} = \overrightarrow{AB} = \langle 0, 1, 1 \rangle$ and $\mathbf{b} = \overrightarrow{CP} = \langle -2, 1, 3 \rangle$. Then $\mathbf{a} \times \mathbf{b} = \langle 3 - 1, -(0 + 2), 0 + 2 \rangle = \langle 2, -2, 2 \rangle \neq \mathbf{0}$. So the lines are not parallel by property (1) in Corollary 11.9. Put $\mathbf{c} = \overrightarrow{AC} = \langle 0, -1, -3 \rangle$. Then

$$\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) = \langle 0, -1, -2 \rangle \cdot \langle 2, -2, 2 \rangle = 0 + 2 - 4 = -2 \neq 0$$

By property (2) in Corollary 11.9, the lines are skew. Next, $\|\mathbf{a} \times \mathbf{b}\| = \|\langle 2, -2, 2 \rangle\| = \|2\langle 1, -1, 1 \rangle\| = 2\|\langle 1, -1, 1 \rangle\| = 2\sqrt{3}$. By Corollary 11.8, the distance between the lines is

$$D = \frac{|\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})|}{\|\mathbf{a} \times \mathbf{b}\|} = \frac{|-2|}{2\sqrt{3}} = \frac{1}{\sqrt{3}}.$$

□

75.4. Study Problems.

Problem 11.20. (Rotations in Space).

Let $\mathbf{a} = \langle a_1, a_2, a_3 \rangle$ and $\mathbf{a}' = \langle a'_1, a'_2, a'_3 \rangle$ be position vectors of a point relative two coordinate systems related to one another by a rotation. As

noted in Section 74.1, the coordinates a'_i and a_i are related by a linear transformation that preserves the length,

$$a'_i = v_{i1}a_1 + v_{i2}a_2 + v_{i3}a_3, \quad i = 1, 2, 3, \quad \|\mathbf{a}\| = \|\mathbf{a}'\|,$$

and excludes the reflections of all coordinate axes or just one of them. So a rotation is described by a 3×3 matrix V with elements v_{ij} . The vectors $\mathbf{v}_i = \langle v_{i1}, v_{i2}, v_{i3} \rangle$ and $\mathbf{w}_i = \langle v_{1i}, v_{2i}, v_{3i} \rangle$, $i = 1, 2, 3$, are the rows and columns of V , respectively. Show that the rows of V are mutually orthonormal, the columns of V are also mutually orthonormal, and the determinant of V is unit:

$$(11.12) \quad \mathbf{v}_i \cdot \mathbf{v}_j = \mathbf{w}_i \cdot \mathbf{w}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad \text{and} \quad \det V = 1.$$

In particular, show that the direct and inverse transformations of the coordinates under a rotation in space are:

$$(11.13) \quad a'_i = \mathbf{v}_i \cdot \mathbf{a} \quad \text{and} \quad a_i = \mathbf{w}_i \cdot \mathbf{a}'.$$

How many independent parameters can the matrix V have for a generic rotation in space?

Hint: If $\hat{\mathbf{u}}_i$ and $\hat{\mathbf{e}}_i$, $i = 1, 2, 3$, are orthonormal unit vectors (bases) associated with the rotated and original coordinate systems, show that the rows of V determine the components of $\hat{\mathbf{u}}_i$ relative to the basis $\hat{\mathbf{e}}_i$, whereas the columns of V determine the components of $\hat{\mathbf{e}}_i$ relative to the basis $\hat{\mathbf{u}}_i$. Use this observation to show that $\hat{\mathbf{u}}_i \cdot \hat{\mathbf{u}}_j = \mathbf{v}_i \cdot \mathbf{v}_j$, $\hat{\mathbf{e}}_i \cdot \hat{\mathbf{e}}_j = \mathbf{w}_i \cdot \mathbf{w}_j$, and

$$(11.14) \quad \hat{\mathbf{u}}_1 \cdot (\hat{\mathbf{u}}_2 \times \hat{\mathbf{u}}_3) = \det V \hat{\mathbf{e}}_1 \cdot (\hat{\mathbf{e}}_2 \times \hat{\mathbf{e}}_3).$$

SOLUTION: A vector \mathbf{a} can be expanded into the sum of mutually orthogonal vectors in each of the coordinate systems

$$\mathbf{a} = a_1\hat{\mathbf{e}}_1 + a_2\hat{\mathbf{e}}_2 + a_3\hat{\mathbf{e}}_3 = a'_1\hat{\mathbf{u}}_1 + a'_2\hat{\mathbf{u}}_2 + a'_3\hat{\mathbf{u}}_3.$$

Note that $\|\mathbf{a}\| = \|\mathbf{a}'\|$ by the orthonormality of the bases $\hat{\mathbf{e}}_i$ and $\hat{\mathbf{u}}_i$. Let us multiply both sides of this equality by $\hat{\mathbf{u}}_i$. Put $v_{ij} = \hat{\mathbf{u}}_i \cdot \hat{\mathbf{e}}_j$. Then

$$\begin{aligned} a'_i &= \hat{\mathbf{u}}_i \cdot \mathbf{a} = \hat{\mathbf{u}}_i \cdot \hat{\mathbf{e}}_1 a_1 + \hat{\mathbf{u}}_i \cdot \hat{\mathbf{e}}_2 a_2 + \hat{\mathbf{u}}_i \cdot \hat{\mathbf{e}}_3 a_3 \\ &= v_{i1}a_1 + v_{i2}a_2 + v_{i3}a_3 = \mathbf{v}_i \cdot \mathbf{a}. \end{aligned}$$

Thus, the components of the rotation matrix V are the dot products $v_{ij} = \hat{\mathbf{u}}_i \cdot \hat{\mathbf{e}}_j$. For a fixed i , the numbers v_{ij} are scalar projections of $\hat{\mathbf{u}}_i$ onto $\hat{\mathbf{e}}_j$, $j = 1, 2, 3$, and hence are components of $\hat{\mathbf{u}}_i$ relative to the basis $\hat{\mathbf{e}}_j$, that is, $\hat{\mathbf{u}}_i = v_{i1}\hat{\mathbf{e}}_1 + v_{i2}\hat{\mathbf{e}}_2 + v_{i3}\hat{\mathbf{e}}_3$. It is then concluded that

the i th row of V coincides with the components of $\hat{\mathbf{u}}_i$ relative to the basis $\hat{\mathbf{e}}_j$. Similarly,

$$\begin{aligned} a_j &= \hat{\mathbf{e}}_j \cdot \mathbf{a} = \hat{\mathbf{e}}_j \cdot \hat{\mathbf{u}}_1 a'_1 + \hat{\mathbf{e}}_j \cdot \hat{\mathbf{u}}_2 a'_2 + \hat{\mathbf{e}}_j \cdot \hat{\mathbf{u}}_3 a'_3 \\ &= v_{1j} a'_1 + v_{2j} a'_2 + v_{3j} a'_3 = \mathbf{w}_j \cdot \mathbf{a}'. \end{aligned}$$

For a fixed j , the numbers v_{ij} are scalar projections of $\hat{\mathbf{e}}_j$ onto $\hat{\mathbf{u}}_i$, $i = 1, 2, 3$, and hence are components of $\hat{\mathbf{e}}_j$ relative to the basis $\hat{\mathbf{u}}_i$, that is, $\hat{\mathbf{e}}_j = v_{1j}\hat{\mathbf{u}}_1 + v_{2j}\hat{\mathbf{u}}_2 + v_{3j}\hat{\mathbf{u}}_3$. Thus, the j th column of V coincides with the components of $\hat{\mathbf{e}}_j$ relative to the basis $\hat{\mathbf{u}}_i$. This proves (11.13). Making use of the expansion of $\hat{\mathbf{u}}_i$ in the *orthonormal* basis $\hat{\mathbf{e}}_j$ and of the expansion of $\hat{\mathbf{e}}_i$ in the basis $\hat{\mathbf{u}}_i$, one obtains

$$\begin{aligned} \hat{\mathbf{u}}_i \cdot \hat{\mathbf{u}}_j &= v_{i1}v_{j1} + v_{i2}v_{j2} + v_{i3}v_{j3} = \mathbf{v}_i \cdot \mathbf{v}_j, \\ \hat{\mathbf{e}}_i \cdot \hat{\mathbf{e}}_j &= v_{1i}v_{1j} + v_{2i}v_{2j} + v_{3i}v_{3j} = \mathbf{w}_i \cdot \mathbf{w}_j. \end{aligned}$$

The first relation in (11.12) follows from the orthonormality of the basis vectors $\hat{\mathbf{u}}_i$ and the basis vectors $\hat{\mathbf{e}}_i$. Next, consider the cross product

$$\begin{aligned} \hat{\mathbf{u}}_2 \times \hat{\mathbf{u}}_3 &= (v_{21}\hat{\mathbf{e}}_1 + v_{22}\hat{\mathbf{e}}_2 + v_{23}\hat{\mathbf{e}}_3) \times (v_{31}\hat{\mathbf{e}}_1 + v_{32}\hat{\mathbf{e}}_2 + v_{33}\hat{\mathbf{e}}_3) \\ &= \det V_{11}(\hat{\mathbf{e}}_2 \times \hat{\mathbf{e}}_3) - \det V_{12}(\hat{\mathbf{e}}_3 \times \hat{\mathbf{e}}_1) + \det V_{13}(\hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_2), \\ V_{11} &= \begin{pmatrix} v_{22} & v_{23} \\ v_{23} & v_{33} \end{pmatrix}, \quad V_{12} = \begin{pmatrix} v_{21} & v_{23} \\ v_{31} & v_{33} \end{pmatrix}, \quad V_{13} = \begin{pmatrix} v_{22} & v_{22} \\ v_{31} & v_{32} \end{pmatrix}, \end{aligned}$$

where the skew symmetry of the cross product $\hat{\mathbf{e}}_i \times \hat{\mathbf{e}}_j = -\hat{\mathbf{e}}_j \times \hat{\mathbf{e}}_i$ and the definition of the determinant of a 2×2 matrix have been used; the matrices V_{1i} , $i = 1, 2, 3$, are obtained by removing from V the row and column that contain v_{1i} . Using the symmetry of the triple product under cyclic permutations of the vectors, one has

$$\hat{\mathbf{u}}_1 \cdot (\hat{\mathbf{u}}_2 \times \hat{\mathbf{u}}_3) = (v_{11} \det V_{11} - v_{12} \det V_{12} + v_{13} \det V_{13}) \hat{\mathbf{e}}_1 \cdot (\hat{\mathbf{e}}_2 \times \hat{\mathbf{e}}_3).$$

Equation (11.14) follows from this relation and the definition of the determinant of a 3×3 matrix. Now recall that the handedness of a coordinate system is preserved under rotations, $\hat{\mathbf{u}}_1 \cdot (\hat{\mathbf{u}}_2 \times \hat{\mathbf{u}}_3) = \hat{\mathbf{e}}_1 \cdot (\hat{\mathbf{e}}_2 \times \hat{\mathbf{e}}_3) = \pm 1$, and therefore $\det V = 1$. It is also worth noting that a combination of rotations and reflections is described by matrices V whose rows and columns are orthonormal, but $\det V = \pm 1$. The handedness of a coordinate system is changed if $\det V = -1$.

The vector $\hat{\mathbf{u}}_3$ is determined by its three directional angles in the original coordinate system. Only two of these angles are independent. A rotation about the axis containing the vector $\hat{\mathbf{u}}_3$ does not affect $\hat{\mathbf{u}}_3$ and can be specified by a rotation angle in a plane perpendicular to the axis. This angle determines the vectors $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$ relative to the

original basis. So a general rotation matrix V has three independent parameters.

In particular, the matrix V for counterclockwise rotations about the z axis through an angle ϕ (see Study Problem 11.2) is

$$V = \begin{pmatrix} \cos \phi & \sin \phi & 0 \\ -\sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Relations (11.12) for V and its rows and columns

$$\begin{aligned} \mathbf{v}_1 &= \langle \cos \phi, \sin \phi, 0 \rangle, & \mathbf{w}_1 &= \langle \cos \phi, -\sin \phi, 0 \rangle, \\ \mathbf{v}_2 &= \langle -\sin \phi, \cos \phi, 0 \rangle, & \mathbf{w}_2 &= \langle \sin \phi, \cos \phi, 0 \rangle, \\ \mathbf{v}_3 &= \langle 0, 0, 1 \rangle, & \mathbf{w}_3 &= \langle 0, 0, 1 \rangle \end{aligned}$$

are easy to verify. The result of Study Problem 11.2 can be stated in the form (11.13), where $\mathbf{a} = \langle x, y, z \rangle$ and $\mathbf{a}' = \langle x', y', z' \rangle$. \square

Problem 11.21. 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). \square

Problem 11.22. (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 $\mathbf{b} = \overrightarrow{BC}$, $\mathbf{c} = \overrightarrow{BD}$, and $\mathbf{a} = \overrightarrow{AP}$. 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. \square

Problem 11.23. (Systems of Linear Equations).

Consider a system of linear equations for the variables x , y , and z :

$$\begin{cases} a_1x + b_1y + c_1z = d_1 \\ a_2x + b_2y + c_2z = d_2 \\ a_3x + b_3y + c_3z = d_3 \end{cases} .$$

Define vectors $\mathbf{a} = \langle a_1, a_2, a_3 \rangle$, $\mathbf{b} = \langle b_1, b_2, b_3 \rangle$, $\mathbf{c} = \langle c_1, c_2, c_3 \rangle$, and $\mathbf{d} = \langle d_1, d_2, d_3 \rangle$. Show that the system has a unique solution for any \mathbf{d} if $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) \neq 0$. If $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = 0$, formulate conditions on \mathbf{d} under which the system has a solution.

SOLUTION: The system of linear equations can be cast in the vector form

$$x\mathbf{a} + y\mathbf{b} + z\mathbf{c} = \mathbf{d}.$$

This equation states that a given vector \mathbf{d} is a linear combination of three given vectors. In Study Problem 11.11, it was demonstrated that any vector in space can be uniquely represented as a linear combination of three non-coplanar vectors. So, by Corollary 11.4, the numbers x , y , and z exist and are unique if $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) \neq 0$. When $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = 0$, the vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} lie in one plane. If \mathbf{d} is not in this plane, the system cannot have a solution because \mathbf{d} cannot be represented as a linear combination of vectors in this plane. Suppose that two of the vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} are not parallel. Then their cross product is orthogonal to the plane, and \mathbf{d} must be orthogonal to the cross product in order to be in the plane. If, say, $\mathbf{a} \times \mathbf{b} \neq \mathbf{0}$, then the system has a solution if $\mathbf{d} \cdot (\mathbf{a} \times \mathbf{b}) = 0$. Finally, it is possible that all the vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} are parallel; that is, all pairwise cross products vanish. Then \mathbf{d} must be parallel to them. If, say, $\mathbf{a} \neq \mathbf{0}$, then the system has a solution if $\mathbf{d} \times \mathbf{a} = \mathbf{0}$.

□

75.5. Exercises.

- (1) Find the triple products $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$, $\mathbf{b} \cdot (\mathbf{a} \times \mathbf{c})$, and $\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})$ if
 - (i) $\mathbf{a} = \langle 1, -1, 2 \rangle$, $\mathbf{b} = \langle 2, 1, 2 \rangle$, and $\mathbf{c} = \langle 2, 1, 3 \rangle$
 - (ii) $\mathbf{a} = \mathbf{u}_1 + 2\mathbf{u}_2$, $\mathbf{b} = \mathbf{u}_1 - \mathbf{u}_2 + 2\mathbf{u}_3$, and $\mathbf{c} = \mathbf{u}_2 - 3\mathbf{u}_3$ if $\mathbf{u}_1 \cdot (\mathbf{u}_2 \times \mathbf{u}_3) = 2$
- (2) 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.
- (3) Find the value of s , if any, for which the vectors $\mathbf{a} = \langle 1, 2, 3 \rangle$, $\mathbf{b} = \langle -1, 0, 1 \rangle$, and $\mathbf{c} = \langle s, 1, 2s \rangle$ are coplanar.
- (4) Let $\mathbf{a} = \langle 1, 2, 3 \rangle$, $\mathbf{b} = \langle 2, 1, 0 \rangle$, and $\mathbf{c} = \langle 3, 0, 1 \rangle$. Find the volume of the parallelepiped with adjacent sides $s\mathbf{a} + \mathbf{b}$, $\mathbf{c} - t\mathbf{b}$, and $\mathbf{a} - p\mathbf{c}$,

where s , t , and p are numbers such that $stp = 1$.

(5) 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} ?

(6) 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 .

(7) 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 edges AB , AC , and AD is 9 units

(8) 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 cyclic permutations of vectors in it and the $bac - cab$ rule (11.9).

(9) Let P be a parallelepiped of volume V . Find:

- (i) The volumes of all parallelepipeds whose adjacent edges are diagonals of the adjacent faces of P
- (ii) The volumes of all parallelepipeds whose two adjacent edges are diagonals of two nonparallel faces of P , while the third adjacent edge is the diagonal of P

(10) Given two nonparallel vectors \mathbf{a} and \mathbf{b} , find the most general vector \mathbf{r} if

- (i) $\mathbf{a} \cdot (\mathbf{r} \times \mathbf{b}) = 0$
- (ii) $\mathbf{a} \cdot (\mathbf{r} \times \mathbf{b}) = 0$ and $\mathbf{b} \cdot \mathbf{r} = 0$

(11) 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 ?

(12) 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 $\mathbf{a} = \langle -1, 2, 2 \rangle$.

(13) 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.

(14) 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?

(15) 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.

(16) Consider two parallel straight line segments in space. Formulate an algorithm to compute the distance between them if the coordinates of their endpoints are given. In particular, find the distance between AB and CD if

(i) $A = (1, 1, 1)$, $B = (4, 1, 5)$, $C = (2, 3, 3)$, and $D = (5, 3, 7)$

(ii) $A = (1, 1, 1)$, $B = (4, 1, 5)$, $C = (3, 5, 5)$, and $D = (6, 5, 9)$

Note that this distance does not generally coincide with the distance between the parallel lines containing AB and CD .

(17) Consider the parallelepiped with adjacent edges AB , AC , and AD , where $A = (3, 0, 1)$, $B = (-1, 2, 5)$, $C = (5, 1, -1)$, and $D = (0, 4, 2)$. Find the distances

- (i) Between the edge AB and all other edges parallel to it
- (ii) Between the edge AC and all other edges parallel to it
- (iii) Between the edge AD and all other edges parallel to it
- (iv) Between all parallel planes containing the faces of the parallelepiped

76. Planes in Space

76.1. A Geometrical Description of a Plane in Space. Consider the coordinate plane $z = 0$. It contains the origin and all vectors that are orthogonal to the z axis (all vectors that are orthogonal to $\hat{\mathbf{e}}_3$). Since the coordinate system can be arbitrarily chosen by translating the origin and rotating the coordinate axes, a *plane in space is defined as a set of points whose position vectors relative to a particular point in the set are orthogonal to a given nonzero vector \mathbf{n}* . The vector \mathbf{n} is called a *normal* of the plane. 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. Then the position vector of P relative to P_0 is $\overrightarrow{P_0P} = \mathbf{r} - \mathbf{r}_0 = \langle x - x_0, y - y_0, z - z_0 \rangle$. This vector lies in the plane \mathcal{P} if 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, $\mathbf{n} \perp \overrightarrow{P_0P}$, reads $\mathbf{n} \cdot \overrightarrow{P_0P} = 0$. Thus, the following theorem has just been proved.

THEOREM 11.6. (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

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

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

Given a nonzero vector \mathbf{n} and a number d , it is always possible to find a particular vector \mathbf{r}_0 such that $\mathbf{n} \cdot \mathbf{r}_0 = d$. Since at least one component of \mathbf{n} does not vanish, say, $n_1 \neq 0$, then $\mathbf{r}_0 = \langle d/n_1, 0, 0 \rangle$. Therefore, a general solution of the linear equation $\mathbf{n} \cdot \mathbf{r} = d$ is a set of position vectors of all points of a plane that is orthogonal to \mathbf{n} . The number d determines the position of the plane in space in the following way. Suppose that every point of the plane is displaced by a vector \mathbf{a} , that is, $\mathbf{r} \rightarrow \mathbf{r} + \mathbf{a}$. The equation of the displaced plane is $\mathbf{n} \cdot (\mathbf{r} + \mathbf{a}) = d$ or $\mathbf{n} \cdot \mathbf{r} = d - \mathbf{n} \cdot \mathbf{a}$. If $\mathbf{n} \cdot \mathbf{a} = 0$, each point of the plane is translated within the plane because \mathbf{a} is orthogonal to \mathbf{n} . The plane as a point set does not change and neither does the number d . If the displacement vector \mathbf{a} is not orthogonal to \mathbf{n} , then d changes by the amount $-\mathbf{n} \cdot \mathbf{a} \neq 0$. Since every point of the original plane is translated by the same vector, the result of this transformation is a parallel plane. Variations of d correspond to shifts of the plane parallel to itself along its normal (see Figure 11.15, right panel). Thus, the equations $\mathbf{n} \cdot \mathbf{r} = d_1$ and $\mathbf{n} \cdot \mathbf{r} = d_2$ describe two *parallel* planes. The planes coincide if and only if $d_1 = d_2$. Consequently, two planes $\mathbf{n}_1 \cdot \mathbf{r} = d_1$ and $\mathbf{n}_2 \cdot \mathbf{r} = d_2$ are parallel if and only if their normals are proportional or if and only if their normals are parallel:

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

for some real $s \neq 0$. For example, the planes $x - 2y - z = 5$ and $-2x + 4y + 2z = 1$ are parallel because their normals, $\mathbf{n}_1 = \langle 1, -2, -1 \rangle$ and $\mathbf{n}_2 = \langle -2, 4, 2 \rangle$, are proportional $\mathbf{n}_2 = -2\mathbf{n}_1$.

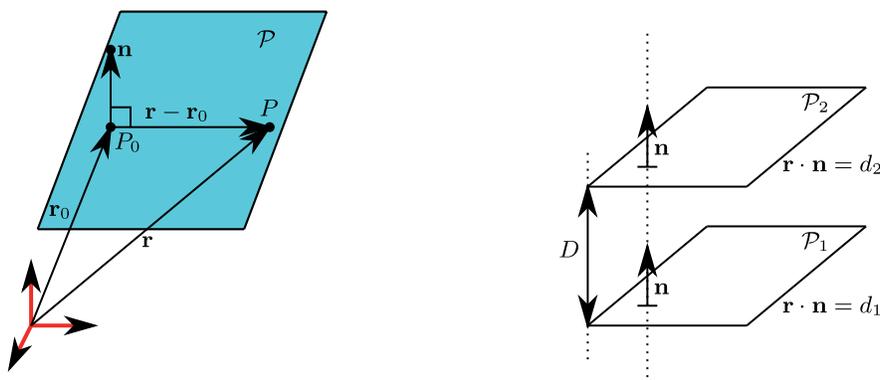


FIGURE 11.15. **Left:** Algebraic description of a plane. If \mathbf{r}_0 is a position vector of a particular point in the plane and \mathbf{r} is the position vector of a generic point in the plane, then the vector $\mathbf{r} - \mathbf{r}_0$ lies in the plane and is orthogonal to its normal, that 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 (11.17).

A normal to a given plane can always be obtained by taking the cross product of any two nonparallel vectors in the plane. Indeed, any vector in a plane is a linear combination of two nonparallel vectors \mathbf{a} and \mathbf{b} (Study Problem 11.6). The vector $\mathbf{n} = \mathbf{a} \times \mathbf{b}$ is orthogonal to both \mathbf{a} and \mathbf{b} and hence to any linear combination of them.

EXAMPLE 11.19. 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 in 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 $\mathbf{a} = \overrightarrow{AB} = \langle 1, 2, -1 \rangle$ and $\mathbf{b} = \overrightarrow{AC} = \langle -2, -1, 2 \rangle$. 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 orthogonal to \mathbf{n} and hence parallel to the plane. \square

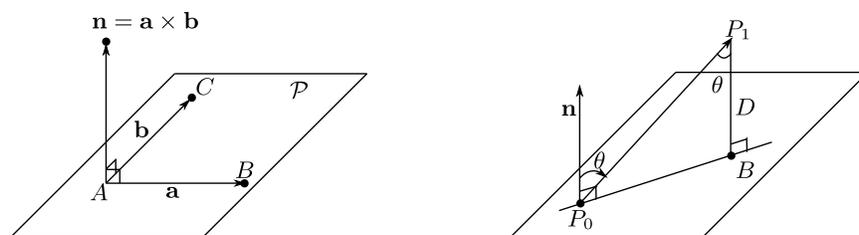


FIGURE 11.16. **Left:** Illustration to Example 11.19. The cross product of two nonparallel vectors in a plane is a normal of the plane. **Right:** Distance between a point P_1 and a plane. An illustration to the derivation of the distance formula (11.15). The segment P_1B is parallel to the normal \mathbf{n} so that the triangle P_0P_1B is right-angled. Therefore, $D = |P_1B| = |P_0P_1| \cos \theta$.

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 point set in space is not changed if the direction of its normal is reversed (i.e., $\mathbf{n} \rightarrow -\mathbf{n}$). 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 $\mathbf{n}_1 \rightarrow -\mathbf{n}_1$ or $\mathbf{n}_2 \rightarrow -\mathbf{n}_2$ so that $\cos \theta \rightarrow -\cos \theta$.

The angle between the planes is useful for determining their relative orientation. Two planes intersect if the angle between them is not 0. Two planes are parallel if the angle between them vanishes. The planes are perpendicular if their normals are orthogonal. 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 orthogonal: $\mathbf{n}_1 \cdot \mathbf{n}_2 = 1 + 2 - 3 = 0$ (i.e., $\mathbf{n}_1 \perp \mathbf{n}_2$).

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 $\overrightarrow{P_0P_1}$ be θ (see Figure 11.16, right panel).

Then the distance in question is $D = \|\overrightarrow{P_0P_1}\| \cos \theta$ if $\theta \leq \pi/2$ (the length of the straight line segment connecting P_1 and the plane along the normal \mathbf{n}). For $\theta > \pi/2$, $\cos \theta$ must be replaced by $-\cos \theta$ because $D \geq 0$. So

$$(11.15) \quad D = \|\overrightarrow{P_0P_1}\| |\cos \theta| = \frac{\|\mathbf{n}\| \|\overrightarrow{P_0P_1}\| |\cos \theta|}{\|\mathbf{n}\|} = \frac{|\mathbf{n} \cdot \overrightarrow{P_0P_1}|}{\|\mathbf{n}\|}.$$

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

$$(11.16) \quad D = \frac{|\mathbf{n} \cdot (\mathbf{r}_1 - \mathbf{r}_0)|}{\|\mathbf{n}\|} = \frac{|\mathbf{n} \cdot \mathbf{r}_1 - d|}{\|\mathbf{n}\|},$$

which is a bit more convenient than (11.15) if the plane is defined by an equation $\mathbf{n} \cdot \mathbf{r} = d$.

Distance Between Parallel Planes. Equation (11.16) 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):

$$(11.17) \quad D = \frac{|d_2 - d_1|}{\|\mathbf{n}\|}.$$

Indeed, the distance between two parallel planes is the distance between the first plane and a point \mathbf{r}_2 in the second plane. By (11.16), this distance is $D = |\mathbf{n} \cdot \mathbf{r}_2 - d_1|/\|\mathbf{n}\| = |d_2 - d_1|/\|\mathbf{n}\|$ because $\mathbf{n} \cdot \mathbf{r}_2 = d_2$ for any point in the second plane.

EXAMPLE 11.20. 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 (11.17) 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 $d = \pm 9 + 2$. \square

76.4. Study Problems.

Problem 11.24. 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, $\mathbf{n} = \overrightarrow{AB} = \langle -2, 2, 4 \rangle$. 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 $-x + y + 2z = 8$. \square

Problem 11.25. 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 orthogonal 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 orthogonal 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 $3x - y - 2z = -5$. \square

Problem 11.26. 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 $\mathbf{n}_2 = \langle 2, 4, 4 \rangle = 2\langle 1, 2, 2 \rangle$. They are not proportional. Hence, the planes are not parallel. Since $\|\mathbf{n}_1\| = 3$, $\|\mathbf{n}_2\| = 6$, and $\mathbf{n}_1 \cdot \mathbf{n}_2 = 2$, the angle is determined by $\cos \theta = 2/18 = 1/9$ or $\theta = \cos^{-1}(1/9)$. \square

Problem 11.27. Find a family of all planes that contains the straight line segment AB if $A = (1, 2, -1)$ and $B = (2, 4, 1)$.

SOLUTION: All the planes in question contain the point A . So it can be chosen as a particular point in every plane. Since the segment AB lies in every plane of the family, the question amounts to describing all vectors orthogonal to $\mathbf{a} = \overrightarrow{AB} = \langle 1, 2, 2 \rangle$ that determine the normals of the planes in the family. It is easy to find a particular vector orthogonal to \mathbf{a} . For example, $\mathbf{b} = \langle 0, 1, -1 \rangle$ is orthogonal to \mathbf{a} because $\mathbf{a} \cdot \mathbf{b} = 0$. Next, the vector $\mathbf{a} \times \mathbf{b} = \langle -4, 1, 1 \rangle$ is orthogonal to both \mathbf{a} and \mathbf{b} . Any vector orthogonal to \mathbf{a} lies in a plane orthogonal to \mathbf{a} and hence must be a linear combination of any two nonparallel vectors in this plane. So the sought-after normals are all linear combinations of \mathbf{b} and $\mathbf{c} = \mathbf{a} \times \mathbf{b}$. Since the length of each normal is irrelevant, the family of the planes is described by all unit vectors orthogonal to \mathbf{a} . Recall that any unit vector in a plane can be written in the form $\hat{\mathbf{n}}_\theta = \cos \theta \hat{\mathbf{u}}_1 + \sin \theta \hat{\mathbf{u}}_2$, where $\hat{\mathbf{u}}_{1,2}$ are two unit orthogonal vectors in the plane and $0 \leq \theta < 2\pi$ (see Figure 11.5, right panel). So put $\hat{\mathbf{u}}_1 = \hat{\mathbf{b}} = \mathbf{b}/\|\mathbf{b}\|$ and $\hat{\mathbf{u}}_2 = \hat{\mathbf{c}} = \mathbf{c}/\|\mathbf{c}\|$, where $\|\mathbf{b}\| = \sqrt{2}$ and $\|\mathbf{c}\| = 3\sqrt{2}$. The family of the planes is described by equations $\hat{\mathbf{n}}_\theta \cdot (\mathbf{r} - \mathbf{r}_0) = 0$, where $0 \leq \theta < 2\pi$ and $\mathbf{r}_0 = \langle 1, 2, -1 \rangle$ (the position vector of A). \square

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) Determine whether the planes $2x + y - z = 3$ and $x + y + z = 1$ are intersecting. If they are, find the angle between them.
- (4) Consider a parallelepiped with one vertex at the origin O at which the adjacent sides are the vectors $\mathbf{a} = \langle 1, 2, 3 \rangle$, $\mathbf{b} = \langle 2, 1, 1 \rangle$, and $\mathbf{c} = \langle -1, 0, 1 \rangle$. Let OP be the largest diagonal of the parallelepiped. Find an equation of the planes that contain:
 - (i) The faces of the parallelepiped
 - (ii) The largest diagonal of the parallelepiped and the diagonal of each of three of its faces adjacent at P
 - (iii) Parallel diagonals in the opposite faces of the parallelepiped
- (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 equations of the planes that are perpendicular to the line through $(1, -1, 1)$ and $(2, 0, 1)$ and that are at the distance 2 from the point $(1, 2, 3)$.

- (7) 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.
- (8) 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 $(-1, 1, 0)$.
- (9) To which of the planes $x + y + z = 1$ and $x + 2y - z = 2$ is the point $(1, 2, 3)$ the closest?
- (10) 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.

- (11) Find values of c for which the plane $x + y + cz = 1$ is closest to the point $P(1, 2, 1)$ and farthest from P .
- (12) 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?
- (13) Find equations of all the planes that are perpendicular to the plane $x + y + z = 1$, have the angle $\pi/3$ with the plane $x + y = 1$, and pass through the point $(1, 1, 1)$.
- (14) Let $\mathbf{a} = \langle 1, 2, 3 \rangle$ and $\mathbf{b} = \langle 1, 0, -1 \rangle$. Find an equation of the plane that contains the point $(1, 2, -1)$, the vector \mathbf{a} , and a vector orthogonal to both \mathbf{a} and \mathbf{b} .
- (15) Consider the plane \mathcal{P} through three points $A(1, 1, 1)$, $B(2, 0, 1)$, and $C(-1, 3, 2)$. Find all the planes that contain the segment AB and have the angle $\pi/3$ with the plane \mathcal{P} . *Hint:* See Study Problem 11.27.
- (16) Find an equation of the plane that contains the line through $(1, 2, 3)$ and $(2, 1, 1)$ and cuts the sphere $x^2 + y^2 + z^2 - 2x + 4y - 6z = 0$ into two hemispheres.
- (17) Find an equation of the plane that is tangent to the sphere $x^2 + y^2 + z^2 - 2x - 4y - 6z + 11 = 0$ at the point $(2, 1, 2)$. *Hint:* What is the angle between a line tangent to a circle at a point P and the segment OP where O is the center of the circle? Extend this observation to a plane tangent to a sphere to determine a normal of the tangent plane.
- (18) Consider a sphere of radius R centered at the origin and two points P_1 and P_2 whose position vectors are \mathbf{r}_1 and \mathbf{r}_2 . Suppose that $\|\mathbf{r}_1\| > R$ and $\|\mathbf{r}_2\| > R$ (the points are outside the sphere). Find the equation $\mathbf{n} \cdot \mathbf{r} = d$ of the plane through P_1 and P_2 whose distance from the sphere is maximal. What is the distance? *Hint:* Show first that a normal of

the plane can always be written in the form $\mathbf{n} = \mathbf{r}_1 + c(\mathbf{r}_2 - \mathbf{r}_1)$. Then find a condition to determine the constant c .

77. Lines in Space

77.1. A Geometrical Description of a Line in Space. Consider the line that coincides with a coordinate axis of a rectangular system, say, the x axis. Any point on it has the characteristic property that its position vector is proportional to the position vector of a particular point (e.g., to $\hat{\mathbf{e}}_1$). Since the coordinate system can be arbitrarily chosen by translating the origin and rotating the coordinate axes, *a line in space is defined as a set of points whose position vectors relative to a particular point in the set are parallel to a given nonzero vector \mathbf{v}* . Thus, the geometrical description of a line \mathcal{L} in space entails specifying a point P_0 that belongs to \mathcal{L} and a vector \mathbf{v} that is parallel to \mathcal{L} .

Remark. 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*. A line in space can also be defined as *a set of points in space such that the shortest path connecting any pair of points of the set belongs to it*. This definition of the line 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 segments of circles and hence are “bent” or “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 are straight lines and 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 systems, 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 . Then the vector $\mathbf{r} - \mathbf{r}_0$ is the position vector of P relative to P_0 . By the geometrical description of the line, it must be parallel to \mathbf{v} . Since any two parallel vectors are proportional, a point (x, y, z) belongs to \mathcal{L} if and only if $\mathbf{r} - \mathbf{r}_0 = t\mathbf{v}$ for some real t (see Figure 11.17, left panel).



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 11.7. (Equations of a Line).

The coordinates of the points of the line \mathcal{L} through a point $P_0(x_0, y_0, z_0)$ and parallel to a vector $\mathbf{v} = \langle v_1, v_2, v_3 \rangle$ satisfy the vector equation

$$(11.18) \quad \mathbf{r} = \mathbf{r}_0 + t\mathbf{v}, \quad -\infty < t < \infty,$$

or the parametric equations

$$(11.19) \quad x = x_0 + tv_1, \quad y = y_0 + tv_2, \quad z = z_0 + tv_3, \quad -\infty < t < \infty.$$

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

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

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 (11.19) 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 11.21. Find the vector, parametric, and symmetric equations of the line through the points $A(1, 1, 1)$ and $B(1, 2, 3)$.

SOLUTION: Take $\mathbf{v} = \overrightarrow{AB} = \langle 0, 1, 2 \rangle$ 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. \square

The line in this example can also be viewed as the set of points of intersection of the planes $x = 1$ and $2y - z = 1$ as follows from the symmetric equations. Clearly, a line can always be described as the set of points of intersection of two nonparallel planes. Since the line lies in each plane, it must be orthogonal to the normals of these planes. Therefore, a vector parallel to the line can always be chosen as the cross product of the normals.

EXAMPLE 11.22. Find the line that is the intersection of the planes $x + y + z = 1$ and $2x - y + z = 2$.

SOLUTION: The normals of the planes are $\mathbf{n}_1 = \langle 1, 1, 1 \rangle$ and $\mathbf{n}_2 = \langle 2, -1, 1 \rangle$. So the vector $\mathbf{v} = \mathbf{n}_1 \times \mathbf{n}_2 = \langle 2, 1, -3 \rangle$ is parallel to the line. To find a particular point of the line, note that its *three* coordinates (x_0, y_0, z_0) should satisfy *two* equations of the planes. So one can choose one of the coordinates at will and find the other two from the equations of the planes. It follows from the parametric equations (11.19) that if, for example, $v_3 \neq 0$, then there is a value of t at which z vanishes, meaning that the line always contains a point with the vanishing z coordinate. Since $v_3 \neq 0$ for the line in question, put $z_0 = 0$. Then

$x_0 + y_0 = 1$ and $2x_0 - y_0 = 2$. By adding these equations, $x_0 = 1$ and hence $y_0 = 0$. The parametric equations of the line of intersection are $x = 1 + 2t$, $y = t$, and $z = -3t$. \square

Distance Between a Point and a Line. Let \mathcal{L} be a line through P_0 and 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 $\overrightarrow{P_0P_1}$ as depicted in Figure 11.17 (right panel). The distance in question is the height of this parallelogram, which is 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 $\overrightarrow{P_0P_1} = \mathbf{r}_1 - \mathbf{r}_0$ and hence

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

77.3. Relative Positions of Lines in Space. Two lines in space can be intersecting, parallel, or skew. The criterion for relative positions of the lines in space is stated in Corollary 11.9. Given an algebraic description of the lines established here, it can now be restated as follows.

COROLLARY 11.10. (Relative Positions of Lines in Space).

Let \mathcal{L}_1 be a line through P_1 and parallel to a vector \mathbf{v}_1 and let \mathcal{L}_2 be a line through P_2 and parallel to a vector \mathbf{v}_2 . Put $\mathbf{r}_{12} = \overrightarrow{P_1P_2}$. Then

- (1) \mathcal{L}_1 and \mathcal{L}_2 are parallel if $\mathbf{v}_1 \times \mathbf{v}_2 = \mathbf{0}$ or $\mathbf{v}_1 = s\mathbf{v}_2$.
- (2) \mathcal{L}_1 and \mathcal{L}_2 are skew if $\mathbf{r}_{12} \cdot (\mathbf{v}_1 \times \mathbf{v}_2) \neq 0$.
- (3) \mathcal{L}_1 and \mathcal{L}_2 intersect if $\mathbf{r}_{12} \cdot (\mathbf{v}_1 \times \mathbf{v}_2) = 0$ and $\mathbf{v}_1 \times \mathbf{v}_2 \neq \mathbf{0}$.
- (4) \mathcal{L}_1 and \mathcal{L}_2 coincide if $\mathbf{r}_{12} \cdot (\mathbf{v}_1 \times \mathbf{v}_2) = 0$ and $\mathbf{v}_1 \times \mathbf{v}_2 = \mathbf{0}$.

Indeed, the vector \mathbf{v}_1 can always be viewed as a vector with initial and terminal points on the line \mathcal{L}_1 . The same observation is true for the vector \mathbf{v}_2 and the line \mathcal{L}_2 .

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 components, the following system of equations is

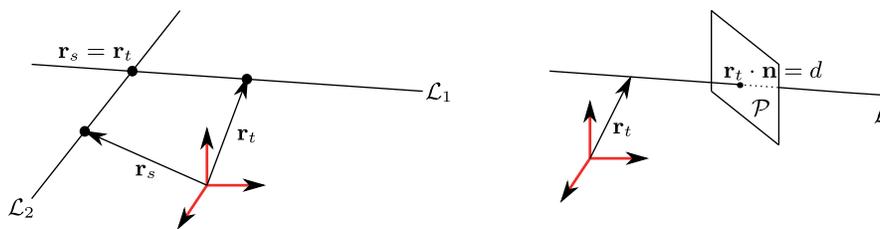


FIGURE 11.18. **Left:** Intersection point of two lines \mathcal{L}_1 and \mathcal{L}_2 . The terminal point of the vector \mathbf{r}_t traverses \mathcal{L}_1 as t ranges over all real numbers, while the terminal point of the vector \mathbf{r}_s traverses \mathcal{L}_2 as s ranges over all real numbers 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: $\mathbf{r}_t \cdot \mathbf{n} = d$.

obtained:

$$\begin{aligned}x_1 + ta_1 &= x_2 + sa_2, \\y_1 + tb_1 &= y_2 + sb_2, \\z_1 + tc_1 &= z_2 + sc_2.\end{aligned}$$

This system of equations is solved in a conventional manner, for example, by expressing t via s from the first equation, substituting it into the second and third ones, and verifying that the resulting two equations have the *same* solution for s . Note that the system has three equations for only two variables. It is an *overdetermined* system, which may or may not have a solution. If the conditions of part (1) or (2) of Corollary 11.10 are satisfied, then the system has no solution (the lines are parallel or skew). If the conditions of part (3) of Corollary 11.10 are fulfilled, then there is a unique solution. Naturally, if the lines coincide, there will be infinitely many solutions. Let $(t, s) = (t_0, s_0)$ be a unique 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 \mathcal{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 orthogonal to the normal of the plane.

COROLLARY 11.11. (Criterion for a Line and a Plane to Be Parallel).
A line \mathcal{L} is parallel to a plane \mathcal{P} if any nonzero vector \mathbf{v} parallel to the line is orthogonal to a normal \mathbf{n} of \mathcal{P} :

$$\mathcal{L} \parallel \mathcal{P} \iff \mathbf{v} \perp \mathbf{n} \iff \mathbf{v} \cdot \mathbf{n} = 0.$$

If a line and a plane are not parallel, they must intersect. In this case, 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 an equation of the plane $\mathbf{r} \cdot \mathbf{n} = d$ (see Figure 11.18, right panel). The value of the parameter t that corresponds to the point of intersection is determined by the equation

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

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 11.23. *A point object is traveling along the line $x - 1 = y/2 = (z + 1)/2$ with a constant speed $v = 6$ meters per second. If all coordinates are measured in meters and the initial position vector of the object is $\mathbf{r}_0 = \langle 1, 0, -1 \rangle$, when does it reach the plane $2x + y + z = 13$? What is the distance traveled by the object?*

SOLUTION: Parametric equations of the line are $x = 1 + s$, $y = 2s$, $z = -1 + 2s$. The value of the parameter s at which the line intersects and the plane is determined by the substitution of these equations into the equation of the plane:

$$2(1 + s) + 2s + (-1 + 2s) = 13 \iff 6s = 12 \iff s = 2.$$

So the position vector of the point of intersection is $\mathbf{r} = \langle 3, 4, 3 \rangle$. The distance between it and the initial point is $D = \|\mathbf{r} - \mathbf{r}_0\| = \|\langle 2, 4, 4 \rangle\| = \|2\langle 1, 2, 2 \rangle\| = 6$ meters and the travel time is $T = D/v = 1$ sec. \square

Remark. In this example, the parameter s does not coincide with the physical time. If an object travels with a constant speed v along the line through \mathbf{r}_0 and parallel to a *unit* vector $\hat{\mathbf{v}}$, then its velocity vector is $\mathbf{v} = v\hat{\mathbf{v}}$ and its position vector is $\mathbf{r} = \mathbf{r}_0 + \mathbf{v}t$, where t is the physical time. Indeed, the vector $\mathbf{r} - \mathbf{r}_0$ is the displacement vector of the object along its trajectory, and hence its length determines the distance traveled by the object: $\|\mathbf{r} - \mathbf{r}_0\| = \|\mathbf{v}t\| = vt$, which shows that the parameter $t > 0$ is the travel time.

EXAMPLE 11.24. 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 orthogonal 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 $\mathbf{n} = \mathbf{n}_1 \times \mathbf{v} = \langle 3, -2, 1 \rangle$. The line lies in \mathcal{P} , and therefore any of its points can be taken as a particular point of \mathcal{P} , for example, $P_0(1, 0, -1)$. An equation of \mathcal{P} reads $3(x - 1) - 2y + (z + 1) = 0$ or $3x - 2y + z = 2$. \square

EXAMPLE 11.25. Find the planes that are perpendicular to the line $x = y/2 = -z/2$ and have the distance 3 from the point $(-1, -2, 2)$ on the line.

SOLUTION: The line is parallel to the vector $\mathbf{v} = \langle 1, 2, -2 \rangle$. So the planes have the same normal $\mathbf{n} = \mathbf{v}$. Particular points in the planes are the points of intersection of the line with the planes. These points are at the distance 3 from $\mathbf{r}_0 = \langle -1, -2, 2 \rangle$, and their position vectors \mathbf{r} should satisfy the condition $\|\mathbf{r} - \mathbf{r}_0\| = 3$. On the other hand, by the vector equation of the line, $\mathbf{r} = \mathbf{r}_0 + t\mathbf{v}$ and hence $\|t\mathbf{v}\| = 3$ or $3|t| = 3$ or $t = \pm 1$. So the position vectors of particular points in the planes are $\mathbf{r} = \mathbf{r}_0 \pm \mathbf{v}$ or $\mathbf{r} = \langle 0, 0, 0 \rangle$ and $\mathbf{r} = \langle -2, -4, 4 \rangle$. Equations of the planes are $x + 2y - 2z = 0$ and $(x + 2) + 2(y + 4) - 2(z - 4) = 0$ or $x + 2y - 2z = -18$. \square

77.5. Study Problems.

Problem 11.28. 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 $\mathbf{r}_{12} = \overrightarrow{P_1P_2} = \langle 3, -1, -3 \rangle$ and $\mathbf{v}_1 \times \mathbf{v}_2 = \langle 1, -2, -3 \rangle$. Therefore, $\mathbf{r}_{12} \cdot (\mathbf{v}_1 \times \mathbf{v}_2) = 14 \neq 0$, and the lines are skew by Corollary 11.10. 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 shown in Figure 11.19 (left panel). The line \mathcal{L} is orthogonal to both vectors \mathbf{v}_1 and \mathbf{v}_2 . As it intersects the lines \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 orthogonal to \mathbf{v}_1

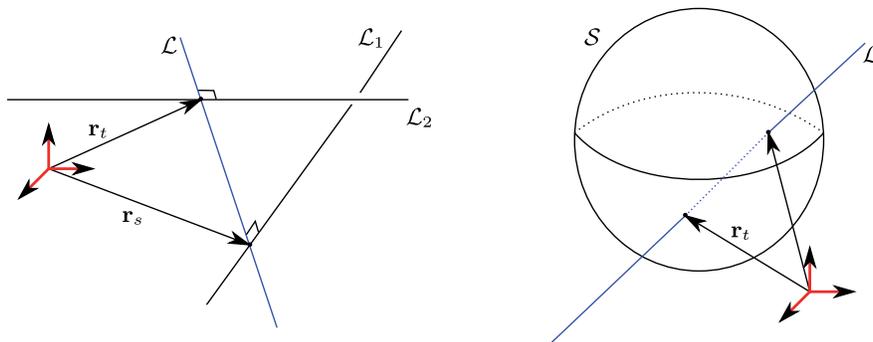


FIGURE 11.19. **Left:** Illustration to Study Problem 11.28. The vectors \mathbf{r}_s and \mathbf{r}_t trace out two given skewed lines \mathcal{L}_1 and \mathcal{L}_2 , respectively. There are particular values of t and s at which the distance $\|\mathbf{r}_t - \mathbf{r}_s\|$ becomes minimal. Therefore, 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 11.29. 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, or one multiple real root, or no real roots. These three cases correspond to two, one, or no points of intersection. The existence of just one point of intersection means that the line is tangent to the sphere.

and \mathbf{v}_2 . The corresponding algebraic conditions are

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

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 (whose position vector is \mathbf{r}_1), the parametric equations read $x = 1 + t$, $y = 1 - 3t$, and $z = 1 - t$. \square

Problem 11.29. Consider a line through the origin that is parallel to the vector $\mathbf{v} = \langle 1, 1, 1 \rangle$. Find the part 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 and

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 $x = t$, $y = t$, and $z = t$, where $-1 \leq t \leq 3$. \square

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 $x = 1 + 2t$, $y = 3t$, and $z = 2 - t$ 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 $P = (1, 1, 2)$ in it. Find parametric equations of the lines through the origin that lie in the plane and are at a distance of 1 unit from P . *Hint:* A vector parallel to these lines can be taken in the form $\mathbf{v} = \langle 1, c, 1 + c \rangle$, where c is to be determined. Explain why!
- (8) Find parametric, symmetric, and vector equations of the line through $(0, 1, 2)$ that is perpendicular to the line $x = 1 + t$, $y = -1 + t$, $z = 2 - 2t$ 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 part 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 .

(11) Find parametric equations of the line through $(1, 2, 5)$ that is perpendicular to the line $x - 1 = 1 - y = z$ and intersects this line.

(12) Find parametric equations of the line that bisects the angle of the triangle ABC at the vertex A if $A = (1, 1, 1)$, $B = (2, -1, 3)$, and $C = (1, 4, -3)$. *Hint:* See exercise 12 in Section 73.7.

(13) Find the distance between the lines $x = y = z$ and $x + 1 = y/2 = z/3$.

(14) A small meteor moves with speed v in the direction of a unit vector $\hat{\mathbf{u}}$. If the meteor passed the point \mathbf{r}_0 , find the condition on $\hat{\mathbf{u}}$ such that the meteor hits an asteroid of the shape of a sphere of radius R centered at the point \mathbf{r}_1 . Determine the position vector of the impact point.

(15) A projectile is fired in the direction $\mathbf{v} = \langle 1, 2, 3 \rangle$ from the point $(1, 1, 1)$. Let the target be a disk of radius R centered at $(2, 3, 6)$ in the plane $2x - 3y + 4z = 19$. If the trajectory of the projectile is a straight line, determine whether it hits a target in two cases $R = 2$ and $R = 3$.

(16) Consider a triangle ABC where $A = (1, 1, 1)$, $B = (3, 1, -1)$, and $C = (1, 3, 1)$. Find the area of a polygon $DPQB$ where the vertices D and Q are the midpoints of CB and AB , respectively, and the vertex P is the intersection of the segments CQ and AD .

78. Quadric Surfaces

DEFINITION 11.17. (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 $A, B, C, p, q, v, \alpha, \beta, \gamma$, and D are real numbers, is called a quadric surface.

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*. 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. If $B = C = 1$, $\alpha = -1$, while the other constants vanish, the quadratic equation $x = y^2 + z^2$ defines a circular paraboloid whose symmetry axis

is the x axis. On the other hand, if $A = B = 1$, $\gamma = -1$, while the other constants vanish, the equation $z = x^2 + y^2$ also defines a paraboloid that can be obtained from the former one by a rotation about the y axis through the angle $\pi/2$ under which $(x, y, z) \rightarrow (z, y, -x)$ so that $x = y^2 + z^2 \rightarrow z = y^2 + x^2$. Thus, there are quadric surfaces of the same shape described by different equations. 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 translations and rotations of the coordinate system. 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} ,

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

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.

THEOREM 11.8. (Classification of Quadric Cylinders).

A general equation for quadric cylinders

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

can be brought to one of the standard forms $A'x^2 + B'y^2 + D' = 0$ or $A'x^2 + \beta'y = 0$ by rotation and translation of the coordinate system, provided A , B , and p do not vanish simultaneously. In particular, these forms define the quadric cylinders:

$$\begin{aligned} y - ax^2 &= 0 && \text{(parabolic cylinder), } \beta' \neq 0, \\ \frac{x^2}{a^2} + \frac{y^2}{b^2} &= 1 && \text{(elliptic cylinder), } \frac{A'}{D'} < 0, \frac{B'}{D'} < 0, D' \neq 0, \\ \frac{x^2}{a^2} - \frac{y^2}{b^2} &= 1 && \text{(hyperbolic cylinder), } A'B' < 0, D' \neq 0. \end{aligned}$$

The shapes of quadric cylinders are shown in Figure 11.20. Other than quadric cylinders, the standard equations may define planes or a

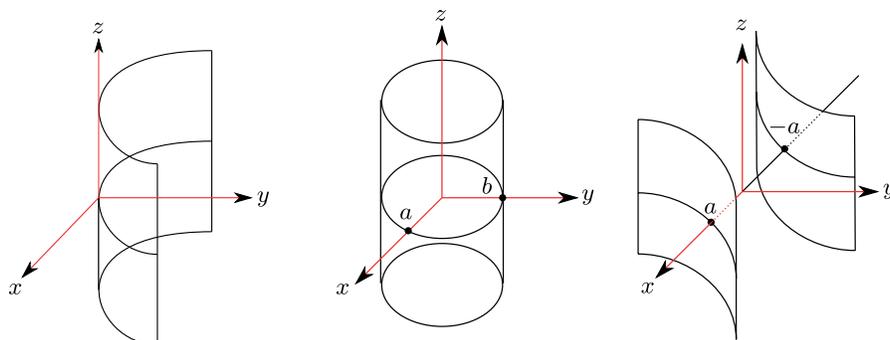


FIGURE 11.20. **Left:** Parabolic cylinder. The cross section by any horizontal plane $z = \text{const}$ is a parabola $y = ax^2$. **Middle:** An elliptic cylinder. The cross section by any horizontal plane $z = \text{const}$ is an ellipse $x^2/a^2 + y^2/b^2 = 1$. **Right:** A hyperbolic cylinder. The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

line for some particular values of the constants A' , B' , D' , and β' . For example, for $A' = -B' = 1$ and $D' = 0$, the equation $x^2 = y^2$ defines two planes $x \pm y = 0$. For $A' = B' = 1$ and $D' = 0$, the equation $x^2 + y^2 = 0$ defines the line $x = y = 0$ (the z axis).

Proof of Theorem 11.8. Let (x, y) be coordinates in the coordinate system obtained by a rotation through an angle ϕ . The equation of \mathcal{S} in the new coordinate system is obtained by the transformation:

$$(x, y) \rightarrow (x \cos \phi - y \sin \phi, y \cos \phi + x \sin \phi)$$

according to Study Problem 11.2. The angle ϕ can be chosen so that the equation for \mathcal{S} does not contain the “mixed” term xy . Indeed, consider the transformation of quadratic terms in the equation for \mathcal{S} :

$$\begin{aligned} x^2 &\rightarrow \cos^2 \phi x^2 + \sin^2 \phi y^2 - 2 \sin \phi \cos \phi xy \\ &= \frac{1}{2}(1 + \cos(2\phi))x^2 + \frac{1}{2}(1 - \cos(2\phi))y^2 - \sin(2\phi)xy, \\ y^2 &\rightarrow \sin^2 \phi x^2 + \cos^2 \phi y^2 + 2 \sin \phi \cos \phi xy \\ &= \frac{1}{2}(1 - \cos(2\phi))x^2 + \frac{1}{2}(1 + \cos(2\phi))y^2 + \sin(2\phi)xy, \\ xy &\rightarrow \sin \phi \cos \phi(x^2 - y^2) + (\cos^2 \phi - \sin^2 \phi)xy \\ &= \frac{1}{2} \sin(2\phi)(x^2 - y^2) + \cos(2\phi)xy. \end{aligned}$$

After the transformation, the coefficient at xy becomes:

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

The angle ϕ is set so that $p' = 0$ or

$$(11.20) \quad \tan(2\phi) = \frac{p}{A-B} \quad \text{and} \quad \phi = \frac{\pi}{4} \quad \text{if} \quad A = B.$$

Similarly, the coefficients A and B (the factors at x^2 and y^2) and α and β (the factors at x and y) become

$$\begin{aligned} A &\rightarrow A' = \frac{1}{2}[A + B + (A - B)\cos(2\phi) + p\sin(2\phi)], \\ B &\rightarrow B' = \frac{1}{2}[A + B - (A - B)\cos(2\phi) - p\sin(2\phi)], \\ \alpha &\rightarrow \alpha' = \alpha\cos\phi + \beta\sin\phi, \quad \beta \rightarrow \beta' = \beta\cos\phi - \alpha\sin\phi, \end{aligned}$$

where ϕ satisfies (11.20). Depending on the values of A , B , and p , the following three cases can occur.

First, $A' = B' = 0$, which is only possible if $A = B = p = 0$. Note that the combination $Ax^2 + By^2 + pxy$ becomes $A'x^2 + B'y^2 + p'xy$ in a rotated coordinate system. If $A' = B' = p' = 0$ for a particular ϕ (chosen to make $p' = 0$), then this combination should be identically 0 in any other coordinate system obtained by rotation. 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. For establishing the shape, it is irrelevant how the horizontal and vertical coordinates in the plane are called. So, without loss of generality, put $B' = 0$. 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,

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

After the *translation* of the coordinate system $x \rightarrow x + x_0$ and $y \rightarrow y + y_0$, the equation is reduced to $A'x^2 + \beta'y = 0$. If $\beta' \neq 0$, it defines a parabola $y - ax^2 = 0$, where $a = -A'/\beta'$.

Third, both A' and B' do not vanish. Then, after the completion of squares, the equation $A'x^2 + B'y^2 + \alpha'x + \beta'y + D = 0$ has the form

$$A'(x - x_0)^2 + B'(y - y_0)^2 + D' = 0,$$

where $x_0 = -\frac{\alpha'}{2A'}$, $y_0 = -\frac{\beta'}{2B'}$, and $D' = -D + \frac{1}{2}(A'x_0^2 + B'y_0^2)$. Finally, after the translation of the origin to the point (x_0, y_0) , the equation becomes

$$A'x^2 + B'y^2 + D' = 0.$$

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 the solution $x = y = 0$ (a line)). 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 $A'/D' < 0$. Note that the rotation of the coordinate system through the angle $\pi/2$ swaps the axes, $(x, y) \rightarrow (y, -x)$, 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.

78.2. Classification of General Quadric Surfaces. The classification of general quadric surfaces can be carried out in the same way. The general quadratic equation can be written in the new coordinate system that is obtained by a translation (11.1) and a rotation (11.13). The rotational freedom (three parameters) can be used to eliminate the “mixed” terms: $p \rightarrow p' = 0$, $q \rightarrow q' = 0$, and $v \rightarrow v' = 0$. After this rotation, the linear terms are eliminated by a suitable translation, provided A' , B' , and C' do not vanish. The corresponding technicalities can be carried out best by linear algebra methods. So the final result is given without a proof.

THEOREM 11.9. (Classification of Quadric Surfaces).

By rotation and translation of a coordinate system, a general equation for quadric surfaces can be brought into one of the standard forms:

$$A'x^2 + B'y^2 + C'z^2 + D' = 0 \quad \text{or} \quad A'x^2 + B'y^2 + \gamma'z = 0.$$

In particular, the standard forms describe quadric cylinders and the following six surfaces:

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

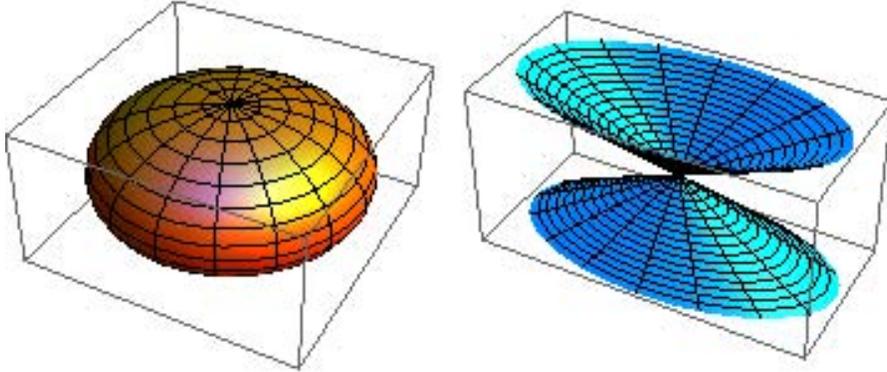


FIGURE 11.21. **Left:** An ellipsoid. A cross section by any coordinate plane is an ellipse. **Right:** An elliptic double cone. A cross section by a horizontal plane $z = \text{const}$ is an ellipse. A cross section by any vertical plane through the z axis is two lines through the origin.

The six shapes are the counterparts in three dimensions of the conic sections in the plane discussed in Calculus II. Other than quadric cylinders or the above six shapes, a quadratic equation may also define planes and lines for particular values of its parameters.

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 (see Figure 11.21, left panel). Traces of an ellipsoid in the planes $x = x_0$, $|x_0| < a$, are ellipses

$$\frac{y^2}{b^2} + \frac{z^2}{c^2} = k \quad \text{or} \quad \frac{y^2}{(b\sqrt{k})^2} + \frac{z^2}{(c\sqrt{k})^2} = 1, \quad k = 1 - \frac{x_0^2}{a^2} > 0.$$

As the plane $x = x_0$ gets closer to $x = a$ or $x = -a$, k becomes smaller and so does the ellipse because its major axes $b\sqrt{k}$ and $c\sqrt{k}$ decrease. 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

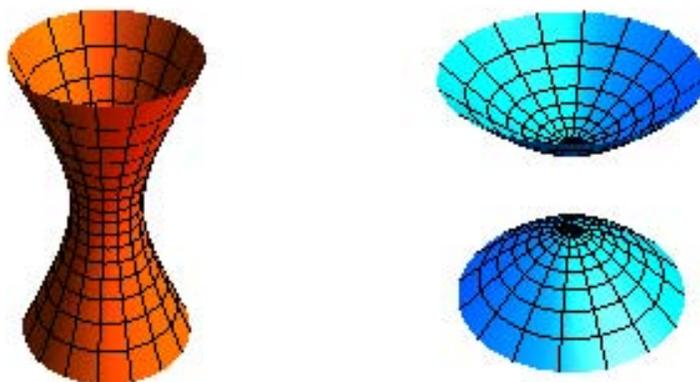


FIGURE 11.22. **Left:** A hyperboloid of one sheet. A cross section by a horizontal plane $z = \text{const}$ is an ellipse. A cross section by a vertical plane $x = \text{const}$ or $y = \text{const}$ is a hyperbola. **Right:** A hyperboloid of two sheets. A nonempty cross section by a horizontal plane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

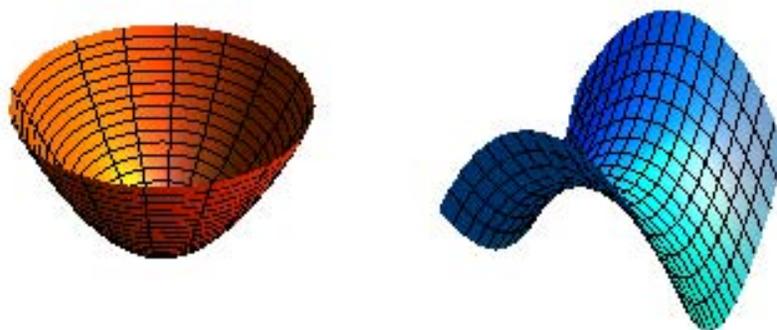


FIGURE 11.23. **Left:** An elliptic paraboloid. A nonempty cross section by a horizontal plane is an ellipse. A cross section by a vertical plane is a parabola. **Right:** A hyperbolic paraboloid (a “saddle”). A cross section by a horizontal plane is a hyperbola. A cross section by a vertical plane is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

$|y_0| \leq b$ 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. Horizontal traces (in the planes $z = z_0$) of the paraboloid are ellipses:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = k \quad \text{or} \quad \frac{x^2}{(a\sqrt{k})^2} + \frac{y^2}{(b\sqrt{k})^2} = 1, \quad k = \frac{z_0}{c}, \quad c > 0.$$

The ellipses become wider as z_0 gets larger because their major axes $a\sqrt{k}$ and $b\sqrt{k}$ grow with increasing k . Vertical traces (traces in the planes $x = x_0$ and $y = y_0$) are parabolas:

$$z - kc = \frac{c}{b^2} y^2, \quad k = \frac{x_0^2}{a^2} \quad \text{and} \quad z - kc = \frac{c}{a^2} x^2, \quad k = \frac{y_0^2}{b^2}.$$

Similarly, a paraboloid with $c < 0$ lies below the xy plane. So *the characteristic geometrical property of a paraboloid is that its horizontal traces are ellipses, while its vertical ones are parabolas* (see Figure 11.23, left panel). If $a = b$, the paraboloid is also called a *circular paraboloid* because its horizontal traces are circles.

A Double Cone. The horizontal traces are ellipses:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = k^2 \quad \text{or} \quad \frac{x^2}{(ak)^2} + \frac{y^2}{(bk)^2} = 1, \quad k = \frac{z_0}{c}.$$

So as $|z_0|$ grows, that is, as the horizontal plane moves away from the xy plane ($z = 0$), the ellipses become wider. In the xy plane, the cone has a trace that consists of a single point (the origin). The vertical traces in the planes $x = 0$ and $y = 0$ are a pair of lines

$$z = \pm(c/b)y \quad \text{and} \quad z = \pm(c/a)x.$$

Furthermore, the trace in any plane that contains the z axis is also a pair of straight lines. Indeed, take parametric equations of a line in the xy plane through the origin, $x = v_1t$, $y = v_2t$. Then the z coordinate of any point of the trace of the cone in the plane that contains the z axis and this line satisfies the equation $z^2/c^2 = [(v_1/a)^2 + (v_2/b)^2]t^2$ or $z = \pm v_3t$, where $v_3 = c\sqrt{(v_1/a)^2 + (v_2/b)^2}$. So the points of intersection, $x = v_1t$, $y = v_2t$, $z = \pm v_3t$ for all real t , form two straight lines through the origin. Given an ellipse in a plane, consider a line through the center of the ellipse that is perpendicular to the plane. Fix a point P on this line that does not coincide with the point of intersection of the line and the plane. Then a double cone is the surface that contains

all lines through P and points of the ellipse. The point P is called the *vertex* of the cone. So *the characteristic geometrical property of a cone is that horizontal traces are ellipses; its vertical traces in planes through the axis of the cone are straight lines* (see Figure 11.21, right panel).

Vertical traces in the planes $x = x_0 \neq 0$ and $y = y_0 \neq 0$ are hyperbolas $y^2/b^2 - z^2/c^2 = k$, where $k = -x_0^2/a^2$, and $x^2/a^2 - z^2/c^2 = k$, where $k = -y_0^2/b^2$. Recall in this regard *conic sections* studied in Calculus II.

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 that is determined by the angle ϕ between the axis of the cone and any of these lines: $c/b = c/a = \cot \phi$. The equation of a circular double 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. The horizontal traces are hyperbolas:

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = k, \quad k = \frac{z_0}{c}.$$

Suppose $c > 0$. If $z_0 > 0$ (horizontal planes below the xy plane), then $k > 0$. In this case, 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) because $x^2/a^2 = y^2/b^2 + k > 0$ (x does not vanish for any y). If $z_0 < 0$, then $k < 0$. In this case, 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) because $y^2/b^2 = x^2/a^2 - k > 0$ (y cannot vanish for any x). Vertical traces in the planes $x = x_0$ and $y = y_0$ are *upward* and *downward* parabolas, respectively:

$$z - z_0 = -\frac{c}{b^2}y^2, \quad z_0 = \frac{cx_0^2}{a^2} \quad \text{and} \quad z - z_0 = \frac{c}{a^2}x^2, \quad z_0 = -\frac{cy_0^2}{b^2}.$$

Take the parabolic trace in the zx plane $z = (c/a^2)x^2$ (i.e., in the plane $y = y_0 = 0$). The traces in the perpendicular planes $x = x_0$ are parabolas whose vertices are $(x_0, 0, z_0)$, where $z_0 = (c/a^2)x_0^2$, and hence lie on the parabola $z = (c/a^2)x^2$ in the zx plane. This observation suggests that the hyperbolic paraboloid is swept by the parabola in the zy plane, $z = -(c/b^2)y^2$, when the latter is moved parallel so that its vertex remains on the parabola $z = (c/a^2)x^2$ in the perpendicular

plane. The obtained surface has the characteristic shape of a “saddle” (see Figure 11.23, right panel).

A Hyperboloid of One Sheet. Traces in horizontal planes $z = z_0$ are ellipses:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = k^2 \quad \text{or} \quad \frac{x^2}{(ka)^2} + \frac{y^2}{(kb)^2} = 1, \quad k = \sqrt{1 + z_0^2/c^2} \geq 1.$$

The ellipse is the smallest in the xy plane ($z_0 = 0$ and $k = 1$). The major axes of the ellipse, ka and kb , grow as the horizontal plane gets away from the xy plane because k increases. The surface looks like a tube with ever-expanding elliptic cross section. The vertical cross section of the “tube” by the planes $x = 0$ and $y = 0$ are hyperbolas:

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

So *the characteristic geometrical property of a hyperboloid of one sheet is that its horizontal traces are ellipses and its vertical traces are hyperbolas* (see Figure 11.22, left panel).

A Hyperboloid of Two Sheets. A distinctive feature of this surface is that it consists of two sheets (see Figure 11.22, right panel). Indeed, the trace in the plane $z = z_0$ satisfies the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{z_0^2}{c^2} - 1,$$

which has no solution if $z_0^2/c^2 - 1 < 0$ or $-c < z_0 < c$. 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 whose major axes increase with increasing $|z_0|$. 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 $(0, 0, -c)$. These points are called *vertices* of a hyperboloid of two sheets. Vertical traces in the planes $x = 0$ and $y = 0$ are hyperbolas:

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

Thus, 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 $x^2/a^2 + y^2/b^2 = \pm 1 + z^2/c^2$ 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. Shifted Quadric Surfaces. If the origin of a coordinated system is shifted to a point (x_0, y_0, z_0) without any rotation of the coordinate axes, then the coordinates of a point in space are translated $(x, y, z) \rightarrow (x - x_0, y - y_0, z - z_0)$. Therefore, any equation of the form $f(x, y, z) = 0$ becomes $f(x - x_0, y - y_0, z - z_0) = 0$ in the new coordinate system. If the equation $f(x, y, z) = 0$ defines a surface in space, then the equation $f(x - x_0, y - y_0, z - z_0) = 0$ defines the very same surface that has been translated as the whole (each point of the surface is shifted by the same vector $\langle x_0, y_0, z_0 \rangle$). For example, the equation

$$\frac{(x - x_0)^2}{a^2} + \frac{(y - y_0)^2}{b^2} = \frac{(z - z_0)^2}{c^2}$$

describes a double elliptic cone whose axis is parallel to the z axis and whose vertex is at (x_0, y_0, z_0) . Equations of shifted quadric surfaces can be reduced to the standard form by completing the squares.

EXAMPLE 11.26. *Classify the quadric surface*

$$9x^2 + 36y^2 + 4z^2 - 18x + 72y + 16z + 25 = 0.$$

SOLUTION: Let us complete the squares for each of the variables:

$$\begin{aligned} 9x^2 - 18x &= 9(x^2 - 2x) = 9[(x - 1)^2 - 1] = 9(x - 1)^2 - 9, \\ 36y^2 + 72y &= 36(y^2 + 2y) = 36[(y + 1)^2 - 1] = 36(y + 1)^2 - 36, \\ 4z^2 + 16z &= 4(z^2 + 4z) = 4[(z + 2)^2 - 4] = 4(z + 2)^2 - 16. \end{aligned}$$

The equation becomes $9(x - 1)^2 + 36(y + 1)^2 + 4(z + 2)^2 = 36$, and, by dividing it by 36, the standard form is obtained

$$\frac{(x - 1)^2}{16} + (y + 1)^2 + \frac{(z + 2)^2}{9} = 1.$$

This equation describes an ellipsoid with the center at $(1, -1, -2)$ and major axes $a = 4$, $b = 1$, and $c = 3$. \square

EXAMPLE 11.27. *Classify the surface $x^2 + 2y^2 - 4y - 2z = 0$.*

SOLUTION: By completing the squares

$$2y^2 - 4y = 2(y^2 - 2y) = 2[(y - 1)^2 - 1],$$

the equation can be written in the form

$$x^2 + 2(y - 1)^2 - 2 - 2z = 0 \quad \text{or} \quad z + 1 = \frac{x^2}{2} + (y - 1)^2,$$

which is an elliptic paraboloid with the vertex at $(0, 1, -1)$ because it is obtained from the standard equation $z = x^2/2 + y^2$ by the shift of the coordinate system $(x, y, z) \rightarrow (x, y - 1, z + 1)$. \square

EXAMPLE 11.28. *Classify the surface $x^2 - 4y^2 + z^2 - 2x - 8z + 1 = 0$.*

SOLUTION: By completing the squares, the equation is transformed to

$$\begin{aligned} (x - 1)^2 - 1 - 4y^2 + 4(z + 1)^2 - 4 + 1 &= 0, \\ \frac{(x - 1)^2}{4} + (z + 1)^2 - y^2 &= 1, \end{aligned}$$

which is a hyperboloid of one sheet whose axis is the line through $(1, 0, -1)$ that is parallel to the y axis. \square

EXAMPLE 11.29. *Use an appropriate rotation in the xy plane to reduce the equation $z = 2xy$ to the standard form and classify the surface.*

SOLUTION: Let (x', y') be coordinates in the rotated coordinate system through the angle ϕ as depicted in Figure 11.3 (right panel). In Study Problem 11.2, the old coordinates (x, y) are expressed via the new ones (x', y') :

$$x = x' \cos \phi - y' \sin \phi, \quad y = y' \cos \phi + x' \sin \phi.$$

In the new coordinate system, the equation

$$z = 2xy = 2x'^2 \cos \phi \sin \phi - 2y'^2 \cos \phi \sin \phi + 2x'y'(\cos^2 \phi - \sin^2 \phi)$$

would have the standard form if the coefficient at $x'y'$ vanishes. So put $\phi = \pi/4$. Then $2 \sin \phi \cos \phi = \sin(2\phi) = 1$ and $z = x'^2 - y'^2$, which is the hyperbolic paraboloid. \square

78.5. Study Problems.

Problem 11.30. *Classify the quadric surface $3x^2 + 3z^2 - 2xz = 4$.*

SOLUTION: The equation does not contain one variable (the y coordinate). The surface is a cylinder parallel to the y axis. To determine the type of cylinder, consider a rotation of the coordinate system in the xz plane and choose the rotation angle so that the coefficient at the xz term vanishes in the transformed equation. According to (11.20), $A = B = 3$, $p = -2$, and hence $\phi = \pi/4$. Then $A' = (A + B - p)/2 = 4$

and $B' = (A + B + p)/2 = 2$. So, in the new coordinates, the equation becomes $4x^2 + 2z^2 = 4$ or $x^2 + z^2/2 = 1$, which is an ellipse with semiaxes $a = 1$ and $b = \sqrt{2}$. The surface is an elliptic cylinder. \square

Problem 11.31. *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: $y \rightarrow \cos \phi y + \sin \phi z$, $z \rightarrow \cos \phi z - \sin \phi y$. Under this rotation, $y - z \rightarrow (\cos \phi + \sin \phi)y + (\sin \phi - \cos \phi)z$. 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. \square

Problem 11.32. *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 $(1, -2, -1)$. \square

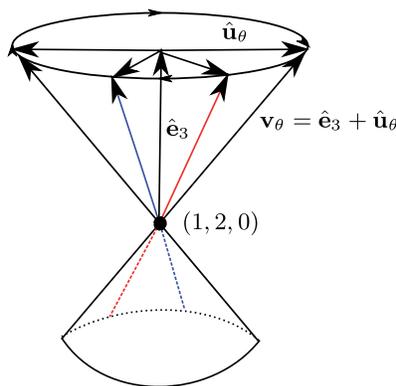


FIGURE 11.24. An illustration to Study Problem 11.33. The vector $\hat{\mathbf{u}}_\theta$ rotates about the vertical line so that the line through $(1, 2, 0)$ and parallel to \mathbf{v}_θ sweeps a double cone with the vertex at $(1, 2, 0)$.

Problem 11.33. *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}_θ rotates about the z axis as θ changes. 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. \square

78.6. 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) $x^2 - y^2 + z^2 = 1$
- (vii) $x^2 + 4y^2 - 9z^2 + 1 = 0$
- (viii) $x^2 + z = 0$
- (ix) $x^2 + 9y^2 + z = 0$
- (x) $y^2 - 4z^2 = 16$

(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) $x^2 + 4y^2 - 6x + z = 0$
- (iv) $y^2 - 4z^2 + 2y - 16z = 0$
- (v) $x^2 - y^2 + z^2 - 2x + 2y = 0$

(3) Use rotations in the appropriate coordinate plane to reduce each of the following equations to one of the standard form and classify the surface:

- (i) $6xy + x^2 + y^2 = 1$
- (ii) $3y^2 + 3z^2 - 2yz = 1$
- (iii) $x - yz = 0$
- (iv) $xy - z^2 = 0$
- (v) $2xz + x^2 - y^2 = 0$

- (4) Find an equation for the surface obtained by rotating the line $y = 2x$ about the y axis. Classify the surface.
- (5) Find an equation for the surface obtained by rotating the curve $y = 1 + z^2$ about the y axis. Classify the surface.
- (6) Find equations for the family of surfaces obtained by rotating the curves $x^2 - 4y^2 = k$ about the y axis where k is real. Classify the surfaces.
- (7) Find an equation for the surface consisting of all points that are equidistant from the point $(1, 1, 1)$ and the plane $z = 2$.
- (8) 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.
- (9) Sketch the solid region bounded by the surfaces $y = 2 - x^2 - z^2$, $y = x^2 + z^2 - 2$, and lies inside the cylinder $x^2 + z^2 = 1$.
- (10) Sketch the solid region bounded by the surfaces $x^2 + y^2 = R^2$ and $x^2 + z^2 = R^2$.
- (11) 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.
- (12) 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.
- (13) Find an equation for the cylinder of radius R whose axis goes through the origin and is parallel to a vector \mathbf{v} .
- (14) 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.
- (15) What are the curves that bound the projections of the ellipsoid $x^2 + y^2 + z^2 - xy = 1$ on the coordinate planes?

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.

The 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,

$$\mathbf{r}(t) = \langle \sqrt{1-t}, \ln(t), t^2 \rangle$$

or

$$x(t) = \sqrt{1-t}, \quad y(t) = \ln(t), \quad z(t) = t^2.$$

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 an interval $\mathcal{D} = I = [a, b]$. Consider all vectors $\mathbf{r}(t)$, as t ranges over I , 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

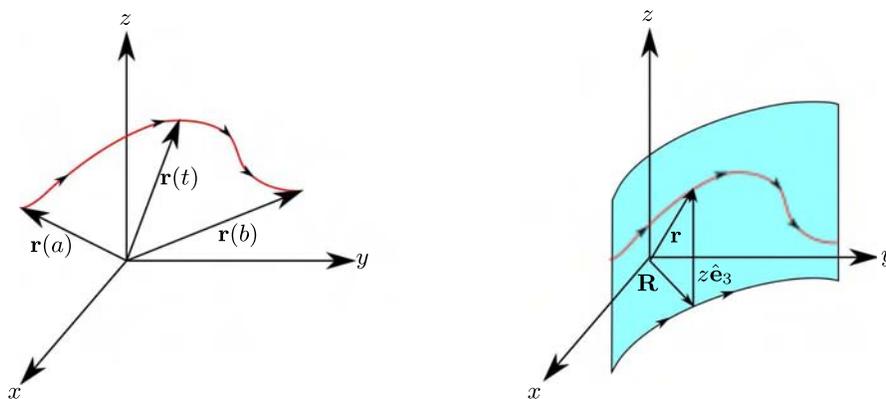


FIGURE 12.1. **Left:** The terminal point of a vector $\mathbf{r}(t)$ whose components are continuous functions of t traces out a curve in space. **Right:** Graphing a space curve. Draw a curve in the xy plane defined by the parametric equations $x = x(t)$, $y = y(t)$. It is traced out by the vector $\mathbf{R}(t) = \langle x(t), y(t), 0 \rangle$. 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.

$\mathbf{r}(t) = \mathbf{r}_0 + t\mathbf{v}$. Thus, the *range* of a vector function defines a curve in space, and a graph of a vector function is a curve in space.

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

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

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 $(x(t_n), y(t_n))$, $n = 1, 2, \dots, N$. 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 $x = x(t)$, $y = y(t)$ while each point t on the t axis coincides with 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 . 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. \square

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 $t \rightarrow t_0$ if

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

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

The left and right limits, $\lim_{t \rightarrow t_0^-} \mathbf{r}(t)$ and $\lim_{t \rightarrow t_0^+} \mathbf{r}(t)$, 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.

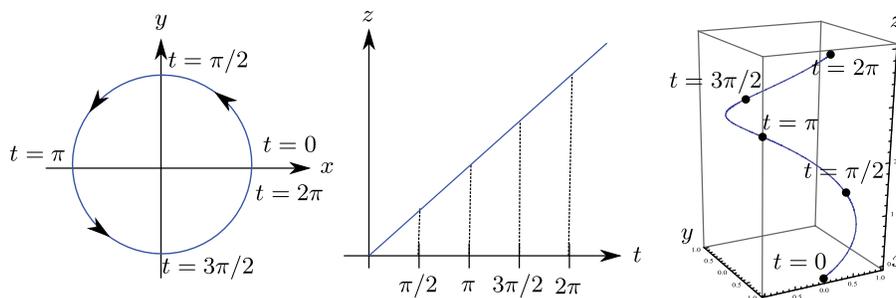


FIGURE 12.2. Graphing a helix. **Left:** The curve $\mathbf{R}(t) = \langle \cos t, \sin t, 0 \rangle$ is a circle of unit radius, traced out counter-clockwise. So the helix lies on the cylinder of unit radius whose symmetry axis is the z axis. **Middle:** The graph $z = z(t) = t$ is a straight line 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.

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 the limit of a vector function exists if and only if the limits of its components exist:

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

This theorem reduces the problem of finding the limit of a vector function to the problem of finding the limits of three ordinary functions.

EXAMPLE 12.2. Let $\mathbf{r}(t) = \langle \sin(t)/t, t \ln t, (e^t - 1 - t)/t^2 \rangle$. Find the limit of $\mathbf{r}(t)$ as $t \rightarrow 0^+$ or show that it does not exist.

SOLUTION: The existence of the limits of the components of the given vector function can be investigated by l'Hospital's rule:

$$\lim_{t \rightarrow 0^+} \frac{\sin t}{t} = \lim_{t \rightarrow 0^+} \frac{(\sin t)'}{(t)'} = \lim_{t \rightarrow 0^+} \frac{\cos t}{1} = 1,$$

$$\lim_{t \rightarrow 0^+} t \ln t = \lim_{t \rightarrow 0^+} \frac{\ln t}{t^{-1}} = \lim_{t \rightarrow 0^+} \frac{(\ln t)'}{(t^{-1})'} = \lim_{t \rightarrow 0^+} \frac{t^{-1}}{-t^{-2}} = - \lim_{t \rightarrow 0^+} t = 0,$$

$$\lim_{t \rightarrow 0^+} \frac{e^t - 1 - t}{t^2} = \lim_{t \rightarrow 0^+} \frac{e^t - 1}{2t} = \lim_{t \rightarrow 0^+} \frac{e^t}{2} = \frac{1}{2},$$

where l'Hospital's rule has been used twice to calculate the last limit. Therefore, $\lim_{t \rightarrow 0^+} \mathbf{r}(t) = \langle 1, 0, 1/2 \rangle$. \square

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

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

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. Let $\mathbf{r}(t) = \langle \sin(2t)/t, t^2, e^t \rangle$ for all $t \neq 0$ and $\mathbf{r}(0) = \langle 1, 0, 1 \rangle$. Determine whether this vector function 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,

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

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

79.3. Space Curves and Continuous Vector Functions. A curve connecting two points in space as a point set can be obtained as a continuous transformation (or a deformation without breaking) of a straight line segment in space. Conversely, every such space curve can be continuously deformed to a straight line segment. So *a curve connecting two points in space is a continuous deformation of a straight line segment, and this deformation has a continuous inverse.*

A straight line segment can be viewed as an interval $a \leq t \leq b$ (a set of real numbers between a and b). Its continuous deformation can be described by a continuous vector function $\mathbf{r}(t)$ on $[a, b]$. So *the range of a continuous vector function defines a curve in space.* Conversely, given a curve C as a point set in space, 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 the same line \mathcal{L} because the vectors $-2\mathbf{v}$ and \mathbf{v} are parallel.

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: $x^2 + y^2 = R^2$ and $y \geq 0$. The following three vector functions trace out the semicircle:

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

This is easy to see by noting that the y components are nonnegative in the specified intervals and the norm of these vector functions is constant for any value of t : $\|\mathbf{r}_i(t)\|^2 = R^2$ or $x_i^2(t) + y_i^2(t) = R^2$, where $i = 1, 2, 3$. The latter means that the endpoints of the vectors $\mathbf{r}_i(t)$ always remain on the circle of radius R . It can therefore be concluded that there are many vector functions whose ranges define the same curve in space.

Another observation is that there are vector functions that trace out the same curve in opposite directions as t increases from its smallest value a to its largest value b . 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 curve. However, this notion of the orientation of a curve should be regarded with caution because a vector function may traverse its range (or a part of it) several times. 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 . 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.

To emphasize the noted differences between space curves as point sets and continuous vector functions, the notion of a *parametric curve* is introduced.

DEFINITION 12.4. (Parametric Curve).

A continuous vector function on an interval is called a parametric curve.

If a continuous vector function $\mathbf{r}(t) = \langle x(t), y(t), z(t) \rangle$, $a \leq t \leq b$, establishes a one-to-one correspondence between an interval $[a, b]$ and a space curve C , then the vector function is also called a *parameterization* of the curve C , the equations $x = x(t)$, $y = y(t)$, and $z = z(t)$ are called *parametric equations* of C , and t is called a *parameter*. As noted, a parameterization of a given space curve is not unique, and there are different parametric equations that describe the very same space

curve. A curve is said to be *simple* if, loosely speaking, it does not intersect itself. To make this notion precise, it is rephrased in terms of parametric curves. A parametric curve $\mathbf{r}(t)$ is called *simple* on a closed interval $[a, b]$ if $\mathbf{r}(t_1) \neq \mathbf{r}(t_2)$ if t_1 and t_2 lie in $[a, b]$ and $t_1 < t_2$, except possibly if both $t_1 = a$ and $t_2 = b$. A *simple parametric curve* is a parametric curve that is simple on every closed interval $[a, b]$ contained in its domain. A point set C is a *simple curve* if there is a simple parametric curve whose range is C . A parametric curve is *closed* if $\mathbf{r}(a) = \mathbf{r}(b)$. A simple parametric curve is always oriented.

EXAMPLE 12.4. Find linear vector functions that orient the straight line segment between $\mathbf{r}_1 = \langle 1, 2, 3 \rangle$ and $\mathbf{r}_2 = \langle 2, 0, 1 \rangle$ from \mathbf{r}_1 to \mathbf{r}_2 and from \mathbf{r}_2 to \mathbf{r}_1 .

SOLUTION: The vector $\mathbf{r}_2 - \mathbf{r}_1 = \langle 1, -2, -2 \rangle$ is parallel to the line segment. So the vector equation $\mathbf{r}(t) = \mathbf{r}_1 + t(\mathbf{r}_2 - \mathbf{r}_1)$ describes the line that contains the segment in question. The vector $\mathbf{r}_2 - \mathbf{r}_1$ is directed from \mathbf{r}_1 to \mathbf{r}_2 . Therefore, when t increases from $t = 0$, the terminal point of $\mathbf{r}(t)$ goes along the line from \mathbf{r}_1 toward \mathbf{r}_2 , reaching the latter at $t = 1$. Thus, the segment is traversed from \mathbf{r}_1 to \mathbf{r}_2 by the vector function

$$\mathbf{r}(t) = \mathbf{r}_1 + t(\mathbf{r}_2 - \mathbf{r}_1) = \langle 1 - t, 2 - 2t, 3 - 2t \rangle, \quad 0 \leq t \leq 1.$$

Swapping the points \mathbf{r}_1 and \mathbf{r}_2 in the above argument, it is concluded that the vector function

$$\mathbf{r}(t) = \mathbf{r}_2 + t(\mathbf{r}_1 - \mathbf{r}_2) = \langle 2 + t, 2t, 1 + 2t \rangle, \quad 0 \leq t \leq 1,$$

traverses the segment from \mathbf{r}_2 to \mathbf{r}_1 . □

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 per one turn. Is such a helix unique?

SOLUTION: Let the helix axis be the z axis. By making the mechanical analogy with the motion of a particle along the helix in question, 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 $\mathbf{r}(t) = \langle R \cos t, R \sin t, ht/(2\pi) \rangle$. Alternatively, one can take parametric equations of the circle in the form $x(t) = R \cos t$, $y(t) = -R \sin t$. In the latter parameterization, the

circle is traced out clockwise, whereas it is traced out counterclockwise in the former parameterization. Consequently, the vector function $\mathbf{r}(t) = \langle R \cos t, -R \sin t, ht/(2\pi) \rangle$ also traces out a helix with the required properties. The two helices are different despite their sharing the same initial and terminal points. One helix winds about the z axis clockwise while the other counterclockwise. \square

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: The vector function $\mathbf{R}(t) = \langle \cos t, \sin t, 0 \rangle$ traverses the circle of unit radius in the xy plane, counterclockwise, starting from the point $(1, 0, 0)$. As t ranges over the specified interval, the circle is traversed 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 $\mathbf{r}(t) = \mathbf{R}(t) + \hat{\mathbf{e}}_3 z(t)$ 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. \square

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 parametric curve $x = x(t)$, $y = y(t)$ 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 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. \square

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 parametric curve $x = 2 \cos(\pi t)$, $z = \sin(\pi t)$ traverses the ellipse $x^2/4 + z^2 = 1$ because the latter equation is satisfied for all real t . Therefore, 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 because the functions

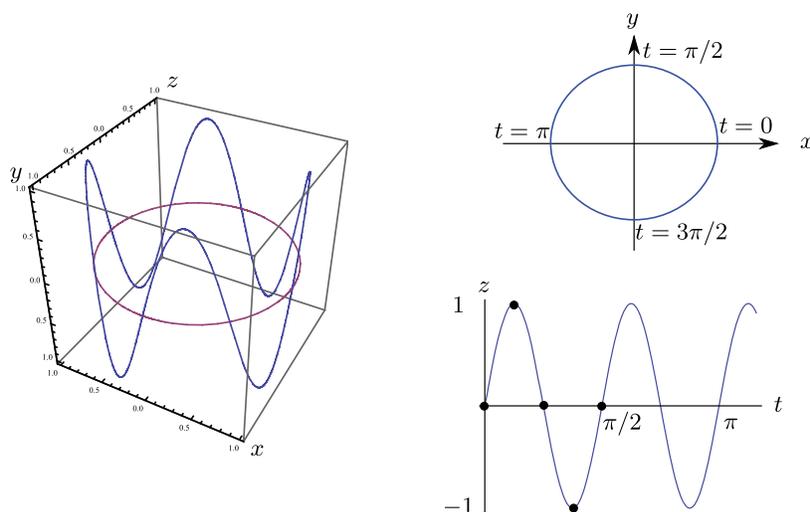


FIGURE 12.3. Illustration to Study Problem 12.2. **Left:** The curve lies on the cylinder of unit radius. It may be viewed as the graph of $z = \sin(4t)$ on the interval $0 \leq t \leq 2\pi$ wrapped around the cylinder. **Top right:** The circle traced out by $\mathbf{R}(t) = \langle \cos t, \sin t, 0 \rangle$. It defines the cylindrical surface on which the curve lies. **Bottom right:** The graph $z = z(t) = \sin(4t)$, which defines the height of points of the curve relative to the circle in the xy plane. The dots on the t axis indicate the points of intersection of the curve with the xy plane in the first quadrant.

$\cos(\pi t)$ and $\sin(\pi t)$ have the period $2\pi/\pi = 2$. The helix rises by 2 along the y axis per turn because $y(t) = t$. 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 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 $-3 \leq t \leq 3$. \square

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 to the

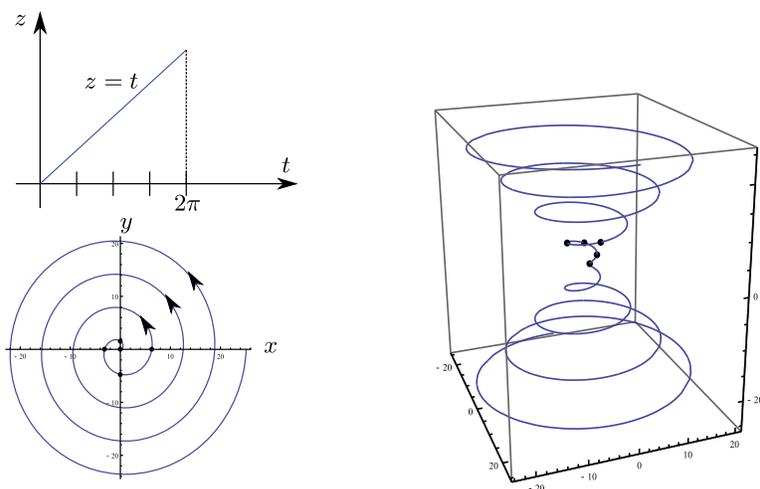


FIGURE 12.4. Illustration to Study Problem 12.3. **Left:** The height of the graph relative to the xy plane (top). The curve $\mathbf{R}(t) = \langle t \cos t, t \sin t, 0 \rangle$. For $t \geq 0$, it looks like an unwinding spiral (bottom). **Right:** For $t > 0$, the curve is traversed by the point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

system of three equations

$$\mathbf{r}_1(t) = \mathbf{r}_2(s) \Leftrightarrow \begin{cases} x_1(t) = x_2(s) \\ y_1(t) = y_2(s) \\ z_1(t) = z_2(s) \end{cases} \Leftrightarrow \begin{cases} t^2 = 8 - 4s \\ t = 2s \\ t^2 + 2t - 8 = s^2 + s - 2 \end{cases}.$$

Substituting the second equation $t = 2s$ 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 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. \square

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 completing the squares, this equation becomes $4 = (x + 1)^2 + (y + 1)^2$, which describes a circle of radius 2 centered at $(-1, -1)$. By construction, this circle is the vertical projection of the curve of intersection onto the xy plane (the plane \mathcal{P}_0 in Figure 12.5). Its parametric equations may be chosen as $x = x(t) = -1 + 2 \cos t$, $y = y(t) = -1 + 2 \sin t$. As t increases from 0 to 2π , the circle is traced out counterclockwise as required (the clockwise orientation can be obtained, e.g., by reversing the sign of $\sin t$). The height along the curve of intersection

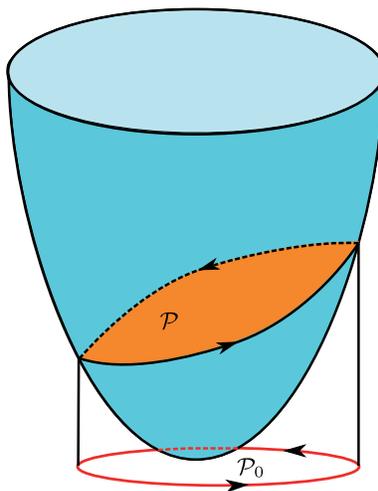


FIGURE 12.5. Illustration to Study Problem 12.6. The curve is an intersection of the paraboloid and the plane \mathcal{P} . It is traversed by the point moving counterclockwise about the circle in the xy plane (indicated by \mathcal{P}_0) and rising so that it remains on the paraboloid.

relative to the xy plane is $z(t) = 2 - 2x(t) - 2y(t)$. Thus, $\mathbf{r}(t) = \langle -1 + 2 \cos t, -1 + 2 \sin t, 6 - 2 \cos t - 2 \sin t \rangle$, where $t \in [0, 2\pi]$. \square

Problem 12.7. Let $\mathbf{v}(t) \rightarrow \mathbf{v}_0$ and $\mathbf{u}(t) \rightarrow \mathbf{u}_0$ as $t \rightarrow t_0$. Prove the limit law for vector functions: $\lim_{t \rightarrow t_0} (\mathbf{v}(t) \cdot \mathbf{u}(t)) = \mathbf{v}_0 \cdot \mathbf{u}_0$ using only Definition 12.2. Then prove this law using Theorem 12.1 and basic limit laws for ordinary functions.

SOLUTION: The idea is similar to the proof of the basic limit laws for ordinary functions given in Calculus I. One has to find an upper bound for $|\mathbf{v} \cdot \mathbf{u} - \mathbf{v}_0 \cdot \mathbf{u}_0|$ in terms of $\|\mathbf{v} - \mathbf{v}_0\|$ and $\|\mathbf{u} - \mathbf{u}_0\|$. By Definition 12.2, the latter quantities converge to 0 as $t \rightarrow t_0$. The conclusion should follow from the squeeze principle. Consider the identities:

$$\begin{aligned} \mathbf{v} \cdot \mathbf{u} - \mathbf{v}_0 \cdot \mathbf{u}_0 &= (\mathbf{v} - \mathbf{v}_0) \cdot \mathbf{u} + \mathbf{v}_0 \cdot \mathbf{u} - \mathbf{v}_0 \cdot \mathbf{u}_0 \\ &= (\mathbf{v} - \mathbf{v}_0) \cdot \mathbf{u} + \mathbf{v}_0 \cdot (\mathbf{u} - \mathbf{u}_0) \\ &= (\mathbf{v} - \mathbf{v}_0) \cdot (\mathbf{u} - \mathbf{u}_0) + (\mathbf{v} - \mathbf{v}_0) \cdot \mathbf{u}_0 + \mathbf{v}_0 \cdot (\mathbf{u} - \mathbf{u}_0). \end{aligned}$$

It follows from the inequality $0 \leq |a + b| \leq |a| + |b|$ and the Cauchy-Schwarz inequality (Theorem 11.2) $|\mathbf{a} \cdot \mathbf{b}| \leq \|\mathbf{a}\| \|\mathbf{b}\|$ that

$$\begin{aligned} 0 &\leq |\mathbf{v} \cdot \mathbf{u} - \mathbf{v}_0 \cdot \mathbf{u}_0| \\ &\leq |(\mathbf{v} - \mathbf{v}_0) \cdot (\mathbf{u} - \mathbf{u}_0)| + |(\mathbf{v} - \mathbf{v}_0) \cdot \mathbf{u}_0| + |\mathbf{v}_0 \cdot (\mathbf{u} - \mathbf{u}_0)| \\ &\leq \|\mathbf{v} - \mathbf{v}_0\| \|\mathbf{u} - \mathbf{u}_0\| + \|\mathbf{v} - \mathbf{v}_0\| \|\mathbf{u}_0\| + \|\mathbf{v}_0\| \|\mathbf{u} - \mathbf{u}_0\|. \end{aligned}$$

By Definition 12.2, $\|\mathbf{v} - \mathbf{v}_0\| \rightarrow 0$ and $\|\mathbf{u} - \mathbf{u}_0\| \rightarrow 0$ as $t \rightarrow t_0$. So the right side of the above inequality converges to 0. By the squeeze principle, it is then concluded that $|\mathbf{v} \cdot \mathbf{u} - \mathbf{v}_0 \cdot \mathbf{u}_0| \rightarrow 0$ as $t \rightarrow t_0$, which proves the assertion. A proof based on Theorem 12.1 is simpler. If $v_i(t)$ and $u_i(t)$, $i = 1, 2, 3$, are components of $\mathbf{v}(t)$ and $\mathbf{u}(t)$, respectively, then, by Theorem 12.1, $v_i(t) \rightarrow v_{0i}$ and $u_i(t) \rightarrow u_{0i}$ as $t \rightarrow t_0$. Hence,

$$\begin{aligned} \lim_{t \rightarrow t_0} \mathbf{v}(t) \cdot \mathbf{u}(t) &= \lim_{t \rightarrow t_0} (v_1(t)u_1(t) + v_2(t)u_2(t) + v_3(t)u_3(t)) \\ &= \lim_{t \rightarrow t_0} v_1(t)u_1(t) + \lim_{t \rightarrow t_0} v_2(t)u_2(t) + \lim_{t \rightarrow t_0} v_3(t)u_3(t) \\ &= v_{01}u_{01} + v_{02}u_{02} + v_{03}u_{03} \\ &= \mathbf{v}_0 \cdot \mathbf{u}_0, \end{aligned}$$

where the basic limit laws for ordinary functions have been used. \square

79.5. Exercises.

(1) Find the domain of each of the following vector functions:

- (i) $\mathbf{r}(t) = \langle t, t^2, e^t \rangle$
- (ii) $\mathbf{r}(t) = \langle \sqrt{t}, t^2, e^t \rangle$
- (iii) $\mathbf{r}(t) = \langle \sqrt{9 - t^2}, \ln t, \cos t \rangle$

- (iv) $\mathbf{r}(t) = \langle \ln(9 - t^2), \ln|t|, (1 + t)/(2 + t) \rangle$
 (v) $\mathbf{r}(t) = \langle \sqrt{t-1}, \ln t, \sqrt{1-t} \rangle$

(2) Find each of the following limits or show that it does not exist:

- (i) $\lim_{t \rightarrow 1} \langle \sqrt{t}, 2 - t - t^2, 1/(t^2 - 2) \rangle$
 (ii) $\lim_{t \rightarrow 1} \langle \sqrt{t}, 2 - t - t^2, 1/(t^2 - 1) \rangle$
 (iii) $\lim_{t \rightarrow 0} \langle e^t, \sin t, t/(1 - t) \rangle$
 (iv) $\lim_{t \rightarrow \infty} \langle e^{-t}, 1/t^2, 4 \rangle$
 (v) $\lim_{t \rightarrow \infty} \langle e^{-t}, (1 - t^2)/t^2, \sqrt[3]{t}/(\sqrt{t} + t) \rangle$
 (vi) $\lim_{t \rightarrow -\infty} \langle 2, t^2, 1/\sqrt[3]{t} \rangle$
 (vii) $\lim_{t \rightarrow 0^+} \langle (e^{2t} - 1)/t, (\sqrt{1+t} - 1)/t, t \ln t \rangle$
 (viii) $\lim_{t \rightarrow 0} \langle \sin^2(2t)/t^2, t^2 + 2, (\cos t - 1)/t^2 \rangle$
 (ix) $\lim_{t \rightarrow 0} \langle (e^{2t} - t)/t, t \cot t, \sqrt{1+t} \rangle$
 (x) $\lim_{t \rightarrow \infty} \langle e^{2t}/\cosh^2 t, t^{2012}e^{-t}, e^{-2t} \sinh^2 t \rangle$

(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) $\mathbf{r}(t) = \langle \sin t, \cos t, \ln t \rangle$
 (v) $\mathbf{r}(t) = \langle t, 1 - t, (t - 1)^2 \rangle$
 (vi) $\mathbf{r}(t) = \langle t^2, t, \sin^2(\pi t) \rangle$
 (vii) $\mathbf{r}(t) = \langle \sin t, \sin t, \sqrt{2} \cos t \rangle$

(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 two vector functions that traverse a given curve C in the opposite directions if C is the curve of intersection of two surfaces:

- (i) $y = x^2$ and $z = 1$
 (ii) $x = \sin y$ and $z = x$
 (iii) $x^2 + y^2 = 9$ and $z = xy$
 (iv) $x^2 + y^2 = z^2$ and $x + y + z = 1$
 (v) $z = x^2 + y^2$ and $y = x^2$
 (vi) $x^2/4 + y^2/9 = 1$ and $x + y + z = 1$
 (vii) $x^2/2 + y^2/2 + z^2/9 = 1$ and $x - y = 0$
 (viii) $x^2 + y^2 - 2x = 0$ and $z = x^2 + y^2$

(6) Specify the parts of the curve $\mathbf{r}(t) = \langle \sin t, \cos t, 4 \sin^2 t \rangle$ that lie above the plane $z = 1$.

(7) 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 $(1, 2, 8)$.

(8) Find the values of a , b , and c , if any, at which each of the following vector functions is continuous: $\mathbf{r}(0) = \langle a, b, c \rangle$ and, for $t \neq 0$,

(i) $\mathbf{r}(t) = \langle t, \cos^2 t, 1 + t + t^2 \rangle$

(ii) $\mathbf{r}(t) = \langle t, \cos^2 t, \sqrt{1 + t^2} \rangle$

(iii) $\mathbf{r}(t) = \langle t, \cos^2 t, \ln |t| \rangle$

(iv) $\mathbf{r}(t) = \langle \sin(2t)/t, \sinh(3t)/t, t \ln |t| \rangle$

(v) $\mathbf{r}(t) = \langle t \cot(2t), t^{1/3} \ln |t|, t^2 + 2 \rangle$

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

$$\lim_{t \rightarrow a} (\mathbf{v}(t) + \mathbf{u}(t)) = \lim_{t \rightarrow a} \mathbf{v}(t) + \lim_{t \rightarrow a} \mathbf{u}(t),$$

$$\lim_{t \rightarrow a} (s\mathbf{v}(t)) = s \lim_{t \rightarrow a} \mathbf{v}(t),$$

$$\lim_{t \rightarrow a} (\mathbf{v}(t) \cdot \mathbf{u}(t)) = \lim_{t \rightarrow a} \mathbf{v}(t) \cdot \lim_{t \rightarrow a} \mathbf{u}(t),$$

$$\lim_{t \rightarrow a} (\mathbf{v}(t) \times \mathbf{u}(t)) = \lim_{t \rightarrow a} \mathbf{v}(t) \times \lim_{t \rightarrow a} \mathbf{u}(t).$$

(10) Prove the last limit law in exercise 9 directly from Definition 12.2, that is, without using Theorem 12.1. *Hint:* See Study Problem 12.7.

(11) Let

$$\mathbf{v}(t) = \langle (e^{2t} - 1)/t, (\sqrt{1+t} - 1)/t, t \ln |t| \rangle,$$

$$\mathbf{u}(t) = \langle \sin^2(2t)/t^2, t^2 + 2, (\cos t - 1)/t^2 \rangle,$$

$$\mathbf{w}(t) = \langle t^{2/3}, 2/(1-t), 1 + t - t^2 + t^3 \rangle.$$

Use the basic laws of limits established in Exercise (9) to find

(i) $\lim_{t \rightarrow 0} (2\mathbf{v}(t) - \mathbf{u}(t) + \mathbf{w}(t))$

(ii) $\lim_{t \rightarrow 0} (\mathbf{v}(t) \cdot \mathbf{u}(t))$

(iii) $\lim_{t \rightarrow 0} (\mathbf{v}(t) \times \mathbf{u}(t))$

(iv) $\lim_{t \rightarrow 0} [\mathbf{w}(t) \cdot (\mathbf{v}(t) \times \mathbf{u}(t))]$

(v) $\lim_{t \rightarrow 0} [\mathbf{w}(t) \times (\mathbf{v}(t) \times \mathbf{u}(t))]$

(vi) $\lim_{t \rightarrow 0} [\mathbf{w}(t) \times (\mathbf{v}(t) \times \mathbf{u}(t)) + \mathbf{v}(t) \times (\mathbf{u}(t) \times \mathbf{w}(t)) + \mathbf{u}(t) \times (\mathbf{w}(t) \times \mathbf{v}(t))]$

(12) 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.

(13) Suppose that the vector functions $\mathbf{v}(t) \times \mathbf{u}(t)$ and $\mathbf{v}(t)$ are continuous. Does this imply that the vector function $\mathbf{u}(t)$ is 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

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

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 Theorem 12.1 that a vector function is differentiable if and only if all its components are differentiable:

$$\begin{aligned} \mathbf{r}'(t) &= \lim_{h \rightarrow 0} \left\langle \frac{x(t+h) - x(t)}{h}, \frac{y(t+h) - y(t)}{h}, \frac{z(t+h) - z(t)}{h} \right\rangle \\ (12.1) \quad &= \langle x'(t), y'(t), z'(t) \rangle. \end{aligned}$$

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 be 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

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

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.

EXAMPLE 12.5. Find the first and second derivatives of the vector function $\mathbf{r}(t) = (\mathbf{a} + t^2\mathbf{b}) \times (\mathbf{c} - t\mathbf{d})$, where \mathbf{a} , \mathbf{b} , \mathbf{c} , and \mathbf{d} are constant vectors.

SOLUTION: By the product rule,

$$\begin{aligned}\mathbf{r}'(t) &= (\mathbf{a} + t^2\mathbf{b})' \times (\mathbf{c} - t\mathbf{d}) + (\mathbf{a} + t^2\mathbf{b}) \times (\mathbf{c} - t\mathbf{d})' \\ &= 2t\mathbf{b} \times (\mathbf{c} - t\mathbf{d}) - (\mathbf{a} + t^2\mathbf{b}) \times \mathbf{d}, \\ \mathbf{r}''(t) &= (2t\mathbf{b})' \times (\mathbf{c} - t\mathbf{d}) + 2t\mathbf{b} \times (\mathbf{c} - t\mathbf{d})' - (\mathbf{a} + t^2\mathbf{b})' \times \mathbf{d} \\ &= 2\mathbf{b} \times (\mathbf{c} - t\mathbf{d}) - 2t\mathbf{b} \times \mathbf{d} - 2t\mathbf{b} \times \mathbf{d} \\ &= 2\mathbf{b} \times \mathbf{c} - 6t\mathbf{b} \times \mathbf{d}.\end{aligned}$$

Alternatively, the cross product can be calculated first and then differentiated:

$$\begin{aligned}\mathbf{r}(t) &= \mathbf{a} \times \mathbf{c} - t\mathbf{a} \times \mathbf{d} + t^2\mathbf{b} \times \mathbf{c} - t^3\mathbf{b} \times \mathbf{d}, \\ \mathbf{r}'(t) &= -\mathbf{a} \times \mathbf{d} + 2t\mathbf{b} \times \mathbf{c} - 3t^2\mathbf{b} \times \mathbf{d}, \\ \mathbf{r}''(t) &= 2\mathbf{b} \times \mathbf{c} - 6t\mathbf{b} \times \mathbf{d}.\end{aligned}$$

□

80.2. Differential of a Vector Function. If $\mathbf{r}(t)$ is differentiable, then

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

where $\mathbf{u}(\Delta t) \rightarrow \mathbf{0}$ as $\Delta t \rightarrow 0$. Indeed, by the definition of the derivative, $\mathbf{u}(\Delta t) = \Delta\mathbf{r}/\Delta t - \mathbf{r}'(t) \rightarrow \mathbf{0}$ as $\Delta t \rightarrow 0$. Therefore, the components of the difference $\Delta\mathbf{r} - \mathbf{r}' \Delta t$ converge to 0 faster than Δt .

Suppose that $\mathbf{r}'(t_0)$ does not vanish. Consider a linear vector function $\mathbf{L}(t)$ with the property $\mathbf{L}(t_0) = \mathbf{r}(t_0)$. Its general form is $\mathbf{L}(t) = \mathbf{r}(t_0) + \mathbf{v}(t - t_0)$, where \mathbf{v} is a constant vector. For t close to t_0 , $\mathbf{L}(t)$ is a *linear approximation* of $\mathbf{r}(t)$ in the sense that the approximation error $\|\mathbf{r}(t) - \mathbf{L}(t)\|$ becomes smaller with decreasing $|t - t_0|$. It follows from (12.2) that

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

By the triangle inequality $\|\|\mathbf{a}\| - \|\mathbf{b}\|\| \leq \|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$, the approximation error is bounded as

$$\left| \|\mathbf{r}'(t_0) - \mathbf{v}\| - \|\mathbf{u}(\Delta t)\| \right| \leq \frac{\|\mathbf{r}(t) - \mathbf{L}(t)\|}{|\Delta t|} \leq \|\mathbf{r}'(t_0) - \mathbf{v}\| + \|\mathbf{u}(\Delta t)\|.$$

If $\|\mathbf{r}'(t_0) - \mathbf{v}\| \neq 0$ or $\mathbf{v} \neq \mathbf{r}'(t_0)$, then $\|\mathbf{u}(\Delta t)\| \ll \|\mathbf{r}'(t_0) - \mathbf{v}\|$ for a sufficiently small Δt because $\|\mathbf{u}(\Delta t)\|$ converges to 0 as $\Delta t \rightarrow 0$ (the sign \ll means “much smaller than”). Therefore, the approximation error decreases linearly with decreasing Δt : $\|\mathbf{r}(t) - \mathbf{L}(t)\| \approx \|\mathbf{r}'(t_0) - \mathbf{v}\| |\Delta t|$. When $\mathbf{v} = \mathbf{r}'(t_0)$, the approximation error decreases faster than Δt :

$$\frac{\|\mathbf{r}(t) - \mathbf{L}(t)\|}{|\Delta t|} = \|\mathbf{u}(\Delta t)\| \rightarrow 0 \quad \text{as } \Delta t \rightarrow 0.$$

Thus, the linear vector function

$$\mathbf{L}(t) = \mathbf{r}(t_0) + \mathbf{r}'(t_0)(t - t_0)$$

is the *best linear approximation* of $\mathbf{r}(t)$ near $t = t_0$. Provided the derivative does not vanish, $\mathbf{r}'(t_0) \neq \mathbf{0}$, 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 $\mathbf{r}(t)$ at the point $\mathbf{r}(t_0)$. The analogy can be made with the tangent line to the graph $y = f(x)$ at a point (x_0, y_0) , where $y_0 = f(x_0)$. The equation of the tangent line is $y = y_0 + f'(x_0)(x - x_0)$ (recall Calculus I). The graph is a curve in the xy plane whose parametric equations are $x = t$, $y = f(t)$ or in the vector form $\mathbf{r}(t) = \langle t, f(t) \rangle$. The parametric equations of the tangent line can therefore be written in the form $x = t = x_0 + (t - t_0)$, $y = y_0 + f'(t_0)(t - t_0)$, where $x_0 = t_0$. Put $\mathbf{r}_0 = \langle x_0, y_0 \rangle$. Then the tangent line is traversed by the linear vector function $\mathbf{L}(t) = \mathbf{r}_0 + \mathbf{r}'(t_0)(t - t_0)$ because $\mathbf{r}'(t_0) = \langle 1, f'(t_0) \rangle$.

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, $\mathbf{r}'(t) = d\mathbf{r}/dt$. Recall that the differential dt is an independent variable that describes infinitesimal variations of t such that higher powers of dt can be neglected. In this sense, the definition of the differential is the linearization of (12.2) in $dt = \Delta t$. At any particular $t = t_0$, the differential $d\mathbf{r}(t_0) = \mathbf{r}'(t_0) dt \neq \mathbf{0}$ defines the tangent line

$$\mathbf{L}(t) = \mathbf{r}(t_0) + d\mathbf{r}(t_0) = \mathbf{r}(t_0) + \mathbf{r}'(t_0) dt, \quad t = t_0 + dt.$$

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 $h \rightarrow 0$, $\overrightarrow{P_0P_h}$ 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

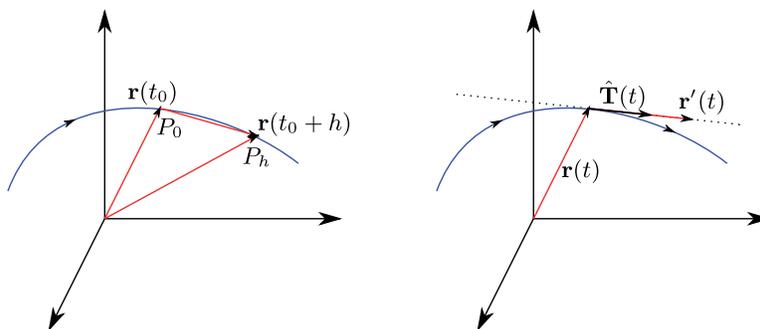


FIGURE 12.6. **Left:** A secant line through two points of the curve, P_0 and P_h . As h gets smaller, the direction of the vector $\overrightarrow{P_0P_h} = \mathbf{r}(t_0 + h) - \mathbf{r}(t_0)$ becomes closer to the tangent to the curve at P_0 . **Right:** The derivative $\mathbf{r}'(t)$ defines a 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 12.6. 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 $x = 2 + 2t$, $y = 2t$, $z = 3 + 5t$. \square

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 C in space is called a smooth curve if there is a simple, continuously differentiable parametric curve $\mathbf{r}(t)$ whose range coincides with C and whose derivative does not vanish.

A smooth parametric curve $\mathbf{r}(t)$ is *oriented* by the direction of the unit tangent vector $\hat{\mathbf{T}}(t)$. Note that if $\mathbf{r}'(t)$ is continuous and never 0, then $\hat{\mathbf{T}}(t)$ is continuous. In particular, with the definition above, a smooth curve does indeed have a continuous unit tangent vector. Therefore, if a curve does not have a continuous unit tangent vector, it cannot be smooth. This enables us to conclude that some curves are not smooth, based on properties deduced from a single parametrization. This is important because one cannot possibly test all parameterizations to see whether one of them meets the conditions in Definition 12.8.

Consider the planar curve $\mathbf{r}(t) = \langle t^3, t^2, 0 \rangle$. 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 $t = 0$. Solving the equation $x = t^3$ for t , $t = x^{1/3}$, and substituting the latter into $y = t^2$, it is concluded that the curve traversed by $\mathbf{r}(t)$ is the graph $y = x^{2/3}$, which has a *cusp* at $x = 0$. The

curve is not smooth at the origin. The tangent line is the vertical line $x = 0$ because $y'(x) = (2/3)x^{-1/3} \rightarrow \pm\infty$ as $x \rightarrow 0^\pm$. The graph lies in the positive half-plane $y \geq 0$ and approaches the y axis, 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 $\mathbf{r}'(0) = \mathbf{0}$. This vector function 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 $t = 0$.

EXAMPLE 12.7. *Determine whether the cycloid C parameterized by $x = a(t - \sin t)$, $y = a(1 - \cos t)$ is smooth, where $a > 0$ is a parameter. If it is not smooth at particular points, investigate its behavior near those points.*

SOLUTION: Following the remark after Definition 12.8, the existence of a continuous unit tangent vector has to be verified. Let $\mathbf{r}(t) = \langle x(t), y(t) \rangle$. Since $x'(t) = a(1 - \cos t) \geq 0$ for all t , and $x'(t) = 0$ only when t is a multiple of 2π , $x(t)$ is monotonically increasing. In particular, $x(t)$ is one-to-one, so C is simple. Since $y'(t) = a \cos t$, the derivatives $x'(t)$ and $y'(t)$ vanish simultaneously if and only if $t = 2\pi n$ for some integer n . Thus, $\mathbf{r}'(t) \neq \mathbf{0}$ unless $t = 2\pi n$, so C is smooth except possibly at the points $\mathbf{r}(2\pi n) = \langle 2\pi na, 0 \rangle$; that is, the portion of C between two consecutive such points is smooth, but it is not yet known whether C is smooth at those points. Since $\|\mathbf{r}'(t)\| = a\sqrt{2(1 - \cos t)} = a\sqrt{4\sin^2(t/2)} = 2a|\sin(t/2)|$, the components of the unit tangent vector for $t \neq 2\pi n$ are

$$T_1(t) = \frac{x'(t)}{\|\mathbf{r}'(t)\|} = |\sin(t/2)|, \quad T_2(t) = \frac{y'(t)}{\|\mathbf{r}'(t)\|} = \frac{\sin t}{2|\sin(t/2)|}.$$

Owing to the periodicity of the sine and cosine functions, it is sufficient to investigate the point corresponding to $t = 0$. If there exists a continuous unit tangent vector, then the limit $\lim_{t \rightarrow 0} \hat{\mathbf{T}}(t)$ should exist and be the unit tangent vector at the point corresponding to $t = 0$. By Theorem 12.1, the limits of the components $T_1(t)$ and $T_2(t)$ should exist as $t \rightarrow 0$. Evidently, $T_1(t) \rightarrow 0$ as $t \rightarrow 0$, but the limit $\lim_{t \rightarrow 0} T_2(t)$

does not exist. Indeed, by l'Hospital's rule the left and right limits are different:

$$\begin{aligned}\lim_{t \rightarrow 0^+} T_2(t) &= \lim_{t \rightarrow 0^+} \frac{\sin t}{2 \sin(t/2)} = \lim_{t \rightarrow 0^+} \frac{\cos t}{\cos(t/2)} = 1 \\ \lim_{t \rightarrow 0^-} T_2(t) &= \lim_{t \rightarrow 0^-} \frac{\sin t}{-2 \sin(t/2)} = - \lim_{t \rightarrow 0^+} \frac{\cos t}{\cos(t/2)} = -1\end{aligned}$$

Therefore $\hat{\mathbf{T}}(t) \rightarrow \langle 0, 1 \rangle$ as $t \rightarrow 0^+$, but $\hat{\mathbf{T}}(t) \rightarrow \langle 0, -1 \rangle$ as $t \rightarrow 0^-$. Thus $\hat{\mathbf{T}}(t)$ cannot be continuously extended across the point $(0, 0)$, so C is not smooth there (as well as at $(2\pi n, 0)$), and, in fact, has a cusp there. \square

A local behavior of the cycloid near $(0, 0)$ may be investigated as follows. Using the Taylor polynomial approximation near $t = 0$, $\sin t \approx t - t^3/6$ and $\cos t \approx 1 - t^2/2$, the cycloid is approximated by the curve $x = at^3/6$, $y = at^2/2$. Expressing $t = (6x/a)^{1/3}$ and substituting it into the other equation, it is concluded that $y = cx^{2/3}$, where $c = (9a/2)^{1/3}$. This curve has a cusp at $x = 0$ as noted above.

80.4. Study Problem.

Problem 12.8. *Prove that, for any smooth curve on a sphere, a tangent vector at any point P is orthogonal 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

$$\mathbf{r}'(t) \cdot (\mathbf{r}(t) - \mathbf{r}_0) = 0 \quad \iff \quad \mathbf{r}'(t) \perp \mathbf{r}(t) - \mathbf{r}_0.$$

If $\mathbf{r}(t)$ is the position vector of P and O is the center of the sphere, then $\overrightarrow{OP} = \mathbf{r}(t) - \mathbf{r}_0$, and hence the tangent vector $\mathbf{r}'(t)$ at P is orthogonal to \overrightarrow{OP} for any t or at any point P of the curve. \square

80.5. Exercises.

(1) Find the derivatives and differentials of each of the following vector functions:

- (i) $\mathbf{r}(t) = \langle 1, 1 + t, 1 + t^3 \rangle$
- (ii) $\mathbf{r}(t) = \langle \cos t, \sin^2(t), t^2 \rangle$
- (iii) $\mathbf{r}(t) = \langle \ln(t), e^{2t}, te^{-t} \rangle$

(iv) $\mathbf{r}(t) = \langle \sqrt[3]{t-2}, \sqrt{t^2-4}, t \rangle$

(v) $\mathbf{r}(t) = \mathbf{a} + \mathbf{b}t^2 - \mathbf{c}e^t$

(vi) $\mathbf{r}(t) = t\mathbf{a} \times (\mathbf{b} - \mathbf{c}e^t)$

(2) Sketch the curve traversed by the vector function $\mathbf{r}(t) = \langle 2, t - 1, t^2 + 1 \rangle$. Indicate the direction in which the curve is traversed by $\mathbf{r}(t)$ with increasing t . Sketch the position vectors $\mathbf{r}(0)$, $\mathbf{r}(1)$, $\mathbf{r}(2)$ and the vectors $\mathbf{r}'(0)$, $\mathbf{r}'(1)$, $\mathbf{r}'(2)$. Repeat the procedure for the vector function $\mathbf{R}(t) = \mathbf{r}(-t) = \langle 2, -t - 1, t^2 + 1 \rangle$ for $t = -2, -1, 0$.

(3) 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) $\mathbf{r}(t) = \langle t, t^2, t^3 \rangle, 0 \leq t \leq 1$

(ii) $\mathbf{r}(t) = \langle t^2, t^3, 2 \rangle, -1 \leq t \leq 1$

(iii) $\mathbf{r}(t) = \langle t^{1/3}, t, t^3 \rangle, -1 \leq t \leq 1$

(iv) $\mathbf{r}(t) = \langle t^5, t^3, t^4 \rangle, -1 \leq t \leq 1$

(v) $\mathbf{r}(t) = \langle \sin^3 t, 1, t^2 \rangle, -\pi/2 \leq t \leq \pi/2$

(4) 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) $\mathbf{r}(t) = \langle \ln t, 2\sqrt{t}, t^2 \rangle, P_0 = (0, 2, 1)$

(5) Find the unit tangent vector to the curve traversed by the specified vector function at the given point P_0 :

(i) $\mathbf{r}(t) = \langle 2t + 1, 2 \tan^{-1} t, e^{-t} \rangle, P_0(1, 0, 1)$

(ii) $\mathbf{r}(t) = \langle \cos(\omega t), \cos(3\omega t), \sin(\omega t) \rangle, P_0(1/2, -1, 1/\sqrt{3})$

(6) Find $\mathbf{r}'(t) \cdot \mathbf{r}''(t)$ and $\mathbf{r}'(t) \times \mathbf{r}''(t)$ if $\mathbf{r}(t) = \langle t, t^2 - 1, t^3 + 2 \rangle$.

(7) 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.

(8) Let $\mathbf{r}(t) = \langle e^t, 2 \cos t, \sin(2t) \rangle$. Use the best linear approximation $\mathbf{L}(t)$ near $t = 0$ to estimate $\mathbf{r}(0.2)$. Use a calculator to assess the accuracy $\|\mathbf{r}(0.2) - \mathbf{L}(0.2)\|$ of the estimate. Repeat the procedure for $\mathbf{r}(0.7)$ and $\mathbf{r}(1.2)$. Compare the errors in all three cases.

(9) Find the point of intersection of the plane $y + z = 3$ and the curve $\mathbf{r}(t) = \langle \ln t, t^2, 2t \rangle$. Find the angle between the normal of the plane and the tangent line to the curve at the point of intersection.

(10) 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. *Hint:* Express the distance between a point on the curve and the plane as a function of t , then solve the extreme value problem.

(11) Find the point of intersection of two curves $\mathbf{r}_1(t) = \langle 1, 1 - t, 3 + t^2 \rangle$ and $\mathbf{r}_2(s) = \langle 3 - s, s - 2, s^2 \rangle$. 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.

(12) State the condition under which the tangent lines to the curve $\mathbf{r}(t)$ at two distinct points $\mathbf{r}(t_1)$ and $\mathbf{r}(t_2)$ are intersecting, or skew, or parallel. Let $\mathbf{r}(t) = \langle 2 \sin(\pi t), \cos(\pi t), \sin(\pi t) \rangle$, $t_1 = 0$, and $t_2 = 1/2$. Determine whether the tangent lines at these points are intersecting and, if so, find the point of intersection.

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

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

(15) Suppose that $\mathbf{r}(t)$ is differentiable three times. Show that $[\mathbf{r}(t) \cdot (\mathbf{r}'(t) \times \mathbf{r}''(t))] = \mathbf{r}(t) \cdot (\mathbf{r}'(t) \times \mathbf{r}'''(t))$.

(16) Let $\mathbf{r}(t)$ be a differentiable vector function. Show that $(\|\mathbf{r}(t)\|)' = \mathbf{r}(t) \cdot \mathbf{r}'(t) / \|\mathbf{r}(t)\|$.

(17) A space warship can fire a laser cannon forward along the tangent line to its trajectory. If the trajectory is traversed by the vector function $\mathbf{r}(t) = \langle t, t, t^2 + 4 \rangle$ in the direction of increasing t and the target is the sphere $x^2 + y^2 + z^2 = 1$, find the part of the trajectory in which the laser cannon can hit the target.

(18) A plane *normal* to a curve at a point P_0 is the plane through P_0 whose normal is tangent to the curve at P_0 . For each of the following curves, find suitable parametric equations, the tangent line, and the normal plane at a specified point:

- (i) $y = x, z = x^2, P_0 = (1, 1, 1)$
- (ii) $x^2 + z^2 = 10, y^2 + z^2 = 10, P_0 = (1, 1, 3)$
- (iii) $x^2 + y^2 + z^2 = 6, x + y + z = 0, P_0 = (1, -2, 1)$

(19) Show that tangent lines to a circular helix have a constant angle with the axis of the helix.

(20) Consider a line through the origin. Any such line sweeps a circular cone when rotated about the z axis and, for this reason, is called a *generating* line of a cone. Prove that the curve $\mathbf{r}(t) = (e^t \cos t, e^t \sin t, e^t)$ intersects all generating lines of the cone $x^2 + y^2 = z^2$ at the same angle. *Hint:* Show that parametric equations of a generating line are $x = s \cos \theta, y = s \sin \theta, z = s$. Define the points of intersection of the line and the curve and find the angle at which they intersect.

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

$$(12.3) \quad \int_a^b \mathbf{r}(t) dt = \left\langle \int_a^b x(t) dt, \int_a^b y(t) dt, \int_a^b z(t) dt \right\rangle.$$

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 $[a, b]$.

EXAMPLE 12.8. Find the integral of $\mathbf{r}(t) = \langle t/\pi, \sin t, \cos t \rangle$ over the interval $[0, \pi]$.

SOLUTION: The components of $\mathbf{r}(t)$ are continuous on $[0, \pi]$. Therefore, by the fundamental theorem of calculus,

$$\begin{aligned} \int_0^\pi \mathbf{r}(t) dt &= \left\langle \int_0^\pi (t/\pi) dt, \int_0^\pi \sin t dt, \int_0^\pi \cos t dt \right\rangle \\ &= \left\langle \frac{t^2}{2\pi} \Big|_0^\pi, -\cos t \Big|_0^\pi, \sin t \Big|_0^\pi \right\rangle = \langle \pi/2, 2, 0 \rangle. \end{aligned}$$

□

DEFINITION 12.10. (Indefinite Integral of a Vector Function).

A vector function $\mathbf{R}(t)$ is called an indefinite integral or an antiderivative 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.

Recall that, for a function $x(t)$ continuous on $[a, b]$, its particular antiderivative derivative is given by

$$X(t) = \int_a^t x(u) du, \quad a \leq t \leq b.$$

Therefore, a particular antiderivative of a continuous vector function $\mathbf{r}(t)$ is

$$\mathbf{R}(t) = \int_a^t \mathbf{r}(u) du, \quad a \leq t \leq b.$$

The vector function $\mathbf{R}(t)$ is *differentiable* on (a, b) and satisfies the condition $\mathbf{R}(a) = \mathbf{0}$. A general antiderivative is obtained by adding a constant vector, $\mathbf{R}(t) \rightarrow \mathbf{R}(t) + \mathbf{c}$. This observation allows us to extend the fundamental theorem of calculus to vector functions.

THEOREM 12.4. (Fundamental Theorem of Calculus for Vector Functions).

If $\mathbf{r}(t)$ is continuous on $[a, b]$, then

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

where $\mathbf{R}(t)$ is any antiderivative of $\mathbf{r}(t)$, that is, a vector function such that $\mathbf{R}'(t) = \mathbf{r}(t)$.

EXAMPLE 12.9. 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, $\mathbf{c} = \langle 2, 1, 0 \rangle - \langle 1, 1, 3 \rangle = \langle 1, 0, -3 \rangle$ and $\mathbf{r}(t) = \langle t^2 + 1, t, 3t^3 - 3 \rangle$. \square

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

$$\mathbf{r}'(t) = \mathbf{v}(t) \quad \text{and} \quad \mathbf{r}(t_0) = \mathbf{r}_0 \quad \Rightarrow \quad \mathbf{r}(t) = \mathbf{r}_0 + \int_{t_0}^t \mathbf{v}(u) du$$

if $\mathbf{v}(t)$ is a continuous vector function. As noted above, 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, $\mathbf{r}'(t) = (d/dt) \int_{t_0}^t \mathbf{v}(u) du = \mathbf{v}(t)$, 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 time is measured in seconds, the length in meters, and the mass in kilograms, then the force is given in newtons, $1 \text{ N} = 1 \text{ kg} \cdot \text{m/s}^2$.

If the force is known as a vector function of time, then Newton's second law 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:

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

if the force is a continuous vector function of time.

Remark. If the force is a function of a particle's position, then 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 12.10. (Motion Under a Constant Force).

Prove that the trajectory of motion under a constant force is a parabola if the initial velocity is not parallel to the force.

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

$$\mathbf{v}(t) = \mathbf{v}_0 + \frac{1}{m} \int_0^t \mathbf{F} \, du = \mathbf{v}_0 + \frac{t}{m} \mathbf{F},$$

$$\mathbf{r}(t) = \mathbf{r}_0 + \int_0^t \mathbf{v}(u) \, du = \mathbf{r}_0 + t\mathbf{v}_0 + \frac{t^2}{2m} \mathbf{F}.$$

If the vectors \mathbf{v}_0 and \mathbf{F} are parallel, then they are proportional, $\mathbf{v}_0 = c\mathbf{F}$. In this particular case, the trajectory $\mathbf{r}(t) = \mathbf{r}_0 + (ct + t^2/(2m))\mathbf{F} = \mathbf{r}_0 + s\mathbf{F}$ lies in the straight line through \mathbf{r}_0 and parallel to \mathbf{F} . The parameter $s = ct + t^2/(2m)$ defines the position of the particle on the line as a function of time. Otherwise, the vector $\mathbf{r}(t) - \mathbf{r}_0$ is a linear combination of two nonparallel vectors \mathbf{v}_0 and \mathbf{F} and hence must be orthogonal 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 orthogonal 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 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 . The parabola is concave in the direction of the force. In Figure 12.7, the force vector points downward and the trajectory is concave downward. \square

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 Example 12.10, 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

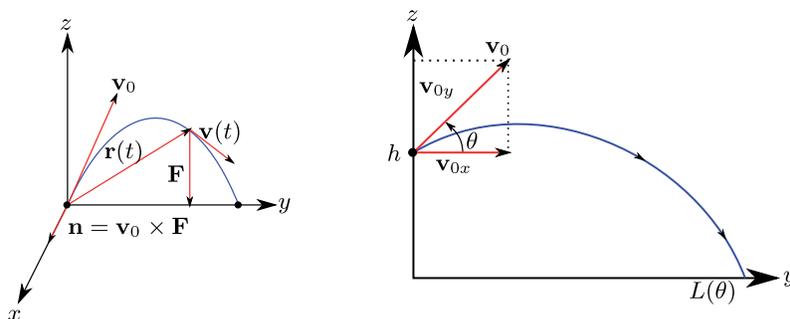


FIGURE 12.7. **Left:** Motion under a constant force \mathbf{F} . The trajectory is a parabola that lies in the plane through the initial point of the motion \mathbf{r}_0 and orthogonal to the vector $\mathbf{n} = \mathbf{v}_0 \times \mathbf{F}$, where the initial velocity \mathbf{v}_0 is assumed to be nonparallel to the force \mathbf{F} . **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 $L(\theta)$.

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 Example 12.10, $F = -mg$ (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 12.9. *The acceleration of a particle is $\mathbf{a} = \langle 2, 6t, 0 \rangle$. Find the position vector of the particle and its velocity in 2 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 $\mathbf{v}(2) = \langle 4, 14, 1 \rangle$. 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 $\mathbf{r}(0) = \langle -1, -4, 1 \rangle$, 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 $\mathbf{r}(2) = \langle 3, 8, 3 \rangle$. \square

Problem 12.10. *Show that if the velocity and position vectors of a particle remain 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. \square

Problem 12.11. 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 magnitude of the magnetic field 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

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

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

Remark. The rate at which the helix rises along the magnetic field is determined by the magnitude (speed) of the initial velocity component v_{\parallel} 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 specified intervals for each of the following functions:

- (i) $\mathbf{r}(t) = \langle 1, 2t, 3t^2 \rangle, 0 \leq t \leq 2$
- (ii) $\mathbf{r}(t) = \langle \sin t, t^3, \cos t \rangle, -\pi \leq t \leq \pi$
- (iii) $\mathbf{r}(t) = \langle t^2, t\sqrt{t-1}, \sqrt{t} \rangle, 0 \leq t \leq 1$
- (iv) $\mathbf{r}(t) = \langle t \ln t, t^2, e^{2t} \rangle, 0 \leq t \leq 1$
- (v) $\mathbf{r}(t) = \langle 2 \sin t \cos t, 3 \sin t \cos^2 t, 3 \sin^2 t \cos t \rangle, 0 \leq t \leq \pi/2$
- (vi) $\mathbf{r}(t) = \mathbf{a} + \cos(t)\mathbf{b}, 0 \leq t \leq \pi$
- (vii) $\mathbf{r}(t) = \mathbf{a} \times (\mathbf{u}'(t) + \mathbf{b}), 0 \leq t \leq 1$ if $\mathbf{u}(0) = \mathbf{a}$ and $\mathbf{u}(1) = \mathbf{a} - \mathbf{b}$

(2) Find $\mathbf{r}(t)$ if the derivatives $\mathbf{r}'(t)$ and $\mathbf{r}(t_0)$ are given:

- (i) $\mathbf{r}'(t) = \langle 1, 2t, 3t^2 \rangle$, $\mathbf{r}(0) = \langle 1, 2, 3 \rangle$
(ii) $\mathbf{r}'(t) = \langle t - 1, t^2, \sqrt{t} \rangle$, $\mathbf{r}(1) = \langle 1, 0, 1 \rangle$
(iii) $\mathbf{r}'(t) = \langle \sin(2t), 2 \cos t, \sin^2 t \rangle$, $\mathbf{r}(\pi) = \langle 1, 2, 3 \rangle$
- (3) Find $\mathbf{r}(t)$ if
- (i) $\mathbf{r}''(t) = \langle 0, 2, 6t \rangle$, $\mathbf{r}(0) = \langle 1, 2, 3 \rangle$, $\mathbf{r}'(0) = \langle 1, 0, -1 \rangle$
(ii) $\mathbf{r}''(t) = \langle t^{1/3}, t^{1/2}, 6t \rangle$, $\mathbf{r}(1) = \langle 1, 0, -1 \rangle$, $\mathbf{r}'(0) = \langle 1, 2, 0 \rangle$
(iii) $\mathbf{r}''(t) = \langle -\sin t, \cos t, 1/t \rangle$, $\mathbf{r}(\pi) = \langle 1, -1, 0 \rangle$, $\mathbf{r}'(\pi) = \langle -1, 0, 2 \rangle$
(iv) $\mathbf{r}''(t) = \langle 0, 2, 6t \rangle$, $\mathbf{r}(0) = \langle 1, 2, 3 \rangle$, $\mathbf{r}(1) = \langle 1, 0, -1 \rangle$
- (4) Solve the equation $\mathbf{r}''(t) = \mathbf{a}$, where \mathbf{a} is a constant vector if $\mathbf{r}(0) = \mathbf{b}$ and $\mathbf{r}(t_0) = \mathbf{c}$ for some $t = t_0 \neq 0$.
- (5) Find the most general vector function whose n th derivative vanishes, $\mathbf{r}^{(n)}(t) = 0$.
- (6) Show that a continuously differentiable vector function $\mathbf{r}(t)$ satisfying the equation $\mathbf{r}'(t) \times \mathbf{r}(t) = \mathbf{0}$ traverses a straight line (or a part of it).
- (7) 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.
- (8) 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 ?
- (9) A particle of mass 1 kg was initially at rest. Then during 2 seconds a constant force of magnitude 3 N was applied to the particle in the direction of $\langle 1, 2, 2 \rangle$. How far is the particle from its initial position in 4 seconds?
- (10) 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.
- (11) 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.
- (12) 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 ?
- (13) 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?

(14) 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

Let a vector function $\mathbf{r}(t)$, $a \leq t \leq b$, traverse a space curve C . Consider a partition of the interval $[a, b]$, $a = t_0 < t_1 < t_2 < \cdots < t_{N-1} < t_N = b$. This partition induces a *partition of the curve*, which is a collection of points of C , P_k , $k = 0, 1, \dots, N$, whose position vectors are $\mathbf{r}(t_k)$. In particular, P_0 and P_N are the endpoints of the curve (see Figure 12.8, left panel). Let $D_N = \max_k(t_k - t_{k-1})$ be the maximal length among all the partition intervals. A partition is said to be *refined* if $D_{N'} < D_N$ for $N' > N$. Under a refinement of a partition, $D_N \rightarrow 0$ as $N \rightarrow \infty$. A refinement is obtained by adding a partition point in each partition interval whose length is D_N (at least one such interval is always present).

DEFINITION 12.11. (Arc Length of a Curve).

Let $\mathbf{r}(t)$, $a \leq t \leq b$, be a vector function traversing a curve C . Let a collection of points P_k be a partition of C , $k = 0, 1, \dots, N$, and let $|P_{k-1}P_k|$ be the distance between two neighboring partition points. The arc length of a curve C is the limit

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

where the partition is refined as $N \rightarrow \infty$, 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, \dots, 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.

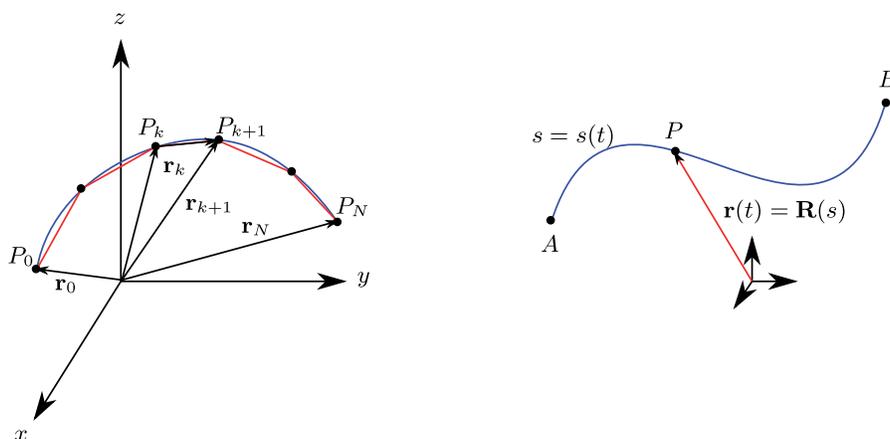


FIGURE 12.8. **Left:** The arc length of a curve is defined as the limit of the sequence of lengths of polygonal paths through partition points of the curve. **Right:** Natural parameterization of a curve. Given a point A of the curve, the arc length s is counted from it to any point P of the curve. 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 $\mathbf{r}(t) = \mathbf{R}(s(t))$.

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

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

PROOF. Owing to the one-to-one correspondence between $[a, b]$ and C , given a partition t_k of $[a, b]$ such that $t_0 = a < t_1 < \cdots < t_{N-1} < t_N = b$, there is a unique polygonal path with vertices P_k on C whose length is

$$\sum_{k=1}^N |P_{k-1}P_k| = \sum_{k=1}^N \|\mathbf{r}_k - \mathbf{r}_{k-1}\|.$$

where $\mathbf{r}_k = \mathbf{r}(t_k)$. Put $\Delta t_k = t_k - t_{k-1} > 0$, $k = 1, 2, \dots, N$. Under a refinement of the partition, $D_N = \max_k \Delta t_k \rightarrow 0$ as $N \rightarrow \infty$ and therefore $\Delta t_k \rightarrow 0$ for all k as $N \rightarrow \infty$. 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 $\mathbf{u}_k \rightarrow \mathbf{0}$ as $\Delta t_k \rightarrow 0$ for every k (cf. (12.2)). By the triangle inequality (11.7),

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

The lower and upper bounds for the length of the polygonal path are obtained by taking the sum over k in this inequality. Next, it is shown that these bounds converge to the Riemann integral of $\|\mathbf{r}'(t)\|$ over $[a, b]$, and the assertion follows from the squeeze principle.

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

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

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

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

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

Remark. Assume $\mathbf{r}(t)$ is 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.

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 L of the curve, $D \geq L$.

EXAMPLE 12.11. Find the arc length of the curve $\mathbf{r}(t) = \langle t^2, 2t, \ln t \rangle$, $1 \leq t \leq 2$.

SOLUTION: The derivative $\mathbf{r}'(t) = \langle 2t, 2, 1/t \rangle$ is continuous on $[1, 2]$. Its norm is

$$\|\mathbf{r}'(t)\| = \sqrt{4t^2 + 4 + \frac{1}{t^2}} = \sqrt{\left(2t + \frac{1}{t}\right)^2} = 2t + \frac{1}{t}.$$

Therefore, by Theorem 12.5,

$$L = \int_1^2 \|\mathbf{r}'(t)\| dt = \int_1^2 \left(2t + \frac{1}{t}\right) dt = t^2 \Big|_1^2 + \ln t \Big|_1^2 = 3 + \ln 2.$$

□

EXAMPLE 12.12. 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 (see Study Problem 12.1). 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 to 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. This consideration also shows that the length does not depend on whether the helix winds about its axis clockwise or counterclockwise. □

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 traversed by different vector functions. These vector functions are different parameterizations of the same curve. For example, a semicircle of radius R is traversed by the vector functions

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

They are related to one another by the composition rule:

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

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. It merely reflects a simple fact that there are many different vector functions that traverse the same space curve.

DEFINITION 12.12. (Reparameterization of a Curve).

Let $\mathbf{r}(t)$ traverse a curve C if $t \in [a, b]$. Let $g(u)$ be a continuous one-to-one function on an interval $[a', b']$ whose range is the interval $[a, b]$; that is, for every $a \leq t \leq b$, there is just one $a' \leq u \leq b'$ such that $t = g(u)$ and vice versa. The vector function $\mathbf{R}(u) = \mathbf{r}(g(u))$ is called a reparameterization of C .

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 parametric equations 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 $s(t)$.

For a smooth curve, the function $\mathbf{r}(t)$ is continuously differentiable, and hence $\|\mathbf{r}'(t)\|$ is continuous on $[a, b]$. Therefore, the derivative $s'(t)$

exists and is obtained by differentiating the integral with respect to its upper limit: $s'(t) = \|\mathbf{r}'(t)\| > 0$. It is positive because $\mathbf{r}'(t) \neq \mathbf{0}$ for a smooth curve. The existence of the inverse function $s(t)$ is guaranteed by the inverse function theorem proved in Calculus I:

THEOREM 12.6. (Inverse Function Theorem).

Let $s(t)$, $a \leq t \leq b$, have a continuous derivative such that $s'(t) > 0$ for $a < t < b$. Then there exists an inverse differentiable function $t = t(s)$, $c < s < d$, and $t'(s) = 1/s'(t)$, where $t = t(s)$ on the right side.

Thus, the condition $s'(t) = \|\mathbf{r}'(t)\| > 0$ guarantees the existence of a one-to-one correspondence between the variables s and t and the existence of the differentiable inverse function $t = t(s)$. Let $\mathbf{r}(t) = \langle x(t), y(t), z(t) \rangle$ be parametric equations of a smooth curve C . Then the parametric equations of C in the natural parameterization have the form

$$\mathbf{R}(s) = \langle x(t(s)), y(t(s)), z(t(s)) \rangle.$$

EXAMPLE 12.13. Reparameterize the helix from Example 12.12, $\mathbf{r}(t) = \langle R \cos t, R \sin t, ht/(2\pi) \rangle$, with respect to the arc length measured from the point $(R, 0, 0)$ in the direction of increasing t .

SOLUTION: The point $(R, 0, 0)$ corresponds to $t = 0$. Then

$$s(t) = \int_0^t \|\mathbf{r}'(u)\| du = \frac{L}{2\pi} \int_0^t du = \frac{Lt}{2\pi} \quad \Rightarrow \quad t(s) = \frac{2\pi s}{L},$$

where $L = \sqrt{(2\pi R)^2 + h^2}$ is the arc length of one turn of the helix (see Example 12.12). Therefore,

$$\mathbf{R}(s) = \mathbf{r}(t(s)) = \langle R \cos(2\pi s/L), R \sin(2\pi s/L), hs/L \rangle$$

In particular, $\mathbf{R}(0) = \langle R, 0, 0 \rangle$ and $\mathbf{R}(L) = \langle R, 0, h \rangle$ are the position vectors of the endpoints of one turn of the helix as required. \square

EXAMPLE 12.14. 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: If $\mathbf{R}(s)$ is the natural parameterization of the helix where s is counted from the point $(4, 0, 0)$, then the position vector of the point in question is given by $\mathbf{R}(5\pi/3)$. Thus, the first task is to find $\mathbf{R}(s)$. One has

$$\mathbf{r}'(u) = \langle -4\pi \sin(\pi u), 4\pi \cos(\pi u), 3\pi \rangle \quad \Rightarrow \quad \|\mathbf{r}'(u)\| = 5\pi.$$

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

$$s(t) = \int_0^t \|\mathbf{r}'(u)\| du = \int_0^t 5\pi du = 5\pi t \quad \Rightarrow \quad t(s) = \frac{s}{5\pi}.$$

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 $\mathbf{R}(5\pi/3) = \langle 2, 2\sqrt{3}, \pi \rangle$. However, this is not a complete answer to the problem because 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. Note that $s(t)$ defined above is the arc length parameter counted from $(4, 0, 0)$ in the direction of *increasing* t (upward along the helix, $t > 0$). Accordingly, $s(t)$ can be counted in the direction of *decreasing* t (downward along the helix, $t < 0$). In this case, $s(t) = -5\pi t > 0$. Hence, the position vector of the other point is $\mathbf{R}(-5\pi/3) = \langle 2, -2\sqrt{3}, -\pi \rangle$. \square

It follows from Theorem 12.6 that *the derivative of a vector function that traverses a smooth curve C with respect to the natural parameter, the arc length, is a unit tangent vector to the curve*. Indeed, by the chain rule applied to the components of the vector function:

$$\begin{aligned} \frac{d\mathbf{r}(t)}{ds} &= \left\langle \frac{dx(t)}{ds}, \frac{dy(t)}{ds}, \frac{dz(t)}{ds} \right\rangle = \langle x'(t)t'(s), y'(t)t'(s), z'(t)t'(s) \rangle \\ &= t'(s) \langle x'(t), y'(t), z'(t) \rangle = \frac{1}{s'(t)} \mathbf{r}'(t) = \frac{1}{\|\mathbf{r}'(t)\|} \mathbf{r}'(t) = \hat{\mathbf{T}}(t). \end{aligned}$$

Thus, for a natural parameterization $\mathbf{r}(s)$ of a smooth curve C , the derivative $\mathbf{r}'(s)$ is a unit tangent vector to C , $\|\mathbf{r}'(s)\| = 1$.

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, by the inverse function theorem, $s'(t) = \|\mathbf{r}'(t)\|$ and $ds = s'(t) dt = \|\mathbf{r}'(t)\| dt$. Thus,

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

for any parameterization of the curve C .

82.3. Exercises.

(1) Find the arc length of each of the following curves:

(i) $\mathbf{r}(t) = \langle 3 \cos t, 2t, 3 \sin t \rangle, -2 \leq t \leq 2$

(ii) $\mathbf{r}(t) = \langle 2t, t^3/3, t^2 \rangle, 0 \leq t \leq 1$

(iii) $\mathbf{r}(t) = \langle 3t^2, 4t^{3/2}, 3t \rangle, 0 \leq t \leq 2$

(iv) $\mathbf{r}(t) = \langle e^t, \sqrt{2}t, e^{-t} \rangle, -1 \leq t \leq 1$

(v) $\mathbf{r}(t) = \langle \cosh t, \sinh t, t \rangle, 0 \leq t \leq 1$

(vi) $\mathbf{r}(t) = \langle \cos t + t \sin t, \sin t + t \cos t, t^2 \rangle, 0 \leq t \leq 2\pi$

Hint: Find the decomposition $\mathbf{r}(t) = \mathbf{v}(t) - t\mathbf{w}(t) + t^2\hat{\mathbf{e}}_3$, where \mathbf{v} , \mathbf{w} , and $\hat{\mathbf{e}}_3$ are mutually orthogonal and $\mathbf{v}'(t) = \mathbf{w}(t)$, $\mathbf{w}'(t) = -\mathbf{v}(t)$. Use the Pythagorean theorem to calculate $\|\mathbf{r}'(t)\|$.

(2) Find the arc length of the curve $\mathbf{r}(t) = \langle e^{-t} \cos t, e^{-t} \sin t, e^{-t} \rangle$, $0 \leq t < \infty$. *Hint:* Put $\mathbf{r}(t) = e^{-t}\mathbf{u}(t)$, differentiate, show that $\mathbf{u}(t)$ is orthogonal to $\mathbf{u}'(t)$, and use the Pythagorean theorem to calculate $\|\mathbf{r}'(t)\|$.

(3) 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 $x^2 + y^2 + z^2 = 2$.

(4) 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 $z = 24$.

(5) Find the arc length of the portion of the curve $\mathbf{r}(t) = \langle \ln t, t^2, 2t \rangle$ that lies between the points of intersection of the curve with the plane $y - 2z + 3 = 0$.

(6) 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 $(36, 18, 6)$.

(7) For each of the following curves defined by the given equations with a parameter a , find suitable parametric equations and evaluate the arc length between a given point A and a generic point $B = (x_0, y_0, z_0)$:

(i) $y = a \sin^{-1}(x/a)$, $z = (a/4) \ln[(a-x)/(a+x)]$, $A = (0, 0, 0)$

(ii) $(x-y)^2 = a(x+y)$, $x^2 - y^2 = 9z^2/8$, $A = (0, 0, 0)$

Hint: Use the new variables $u = x + y$ and $v = x - y$ to find the parametric equations.

(iii) $x^2 + y^2 = az$, $y/x = \tan(z/a)$, $A = (0, 0, 0)$

Hint: Use the polar coordinates in the xy plane to find the parametric equations.

(iv) $x^2 + y^2 + z^2 = a^2$, $\sqrt{x^2 + y^2} \cosh(\tan^{-1}(y/x)) = a$, $A = (a, 0, 0)$

Hint: Represent the second equation as a polar graph.

(8) 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) $\mathbf{r} = \frac{2t}{t^2+1}\hat{\mathbf{e}}_1 + \left(\frac{2}{t^2+1} - 1\right)\hat{\mathbf{e}}_3$

$$(iii) \mathbf{r}(t) = \langle \cosh t, \sinh t, t \rangle$$

$$(iv) x = a(t - \sin t), y = a(1 - \cos t), a > 0$$

(9) 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.

(10) A particle travels along a curve traversed by the vector function $\mathbf{r}(u) = \langle u, \cosh u, \sinh u \rangle$ from the point $(0, 1, 0)$ with a constant speed $\sqrt{2}$ m/s so that its x coordinate increases. Find the position of the particle in 1 second.

(11) Let C be a smooth closed curve whose arc length is L . Let $\mathbf{r}(t)$ be a vector function that traverses C only once for $a \leq t \leq b$. Prove that there is a number $a \leq t^* \leq b$ such that $\|\mathbf{r}'(t^*)\| = L/(b-a)$. *Hint:* Recall the integral mean value theorem.

(12) A particle travels in space a distance D in time T . Show that there is a moment of time $0 \leq t \leq T$ at which the speed of the particle coincides with the average speed D/T .

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 is not the case for a general curve. From the definition of the derivative

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

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

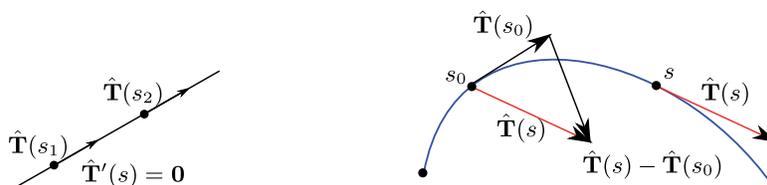


FIGURE 12.9. **Left:** A straight line does not bend. The unit tangent vector has zero rate of change relative to the arc length parameter s . **Right:** Curvature of a smooth curve. The more a smooth curve bends, the larger the rate of change of the unit tangent vector relative to the arc length parameter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

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 $\|\hat{\mathbf{T}}'(s_0)\|$ 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 its unit tangent vector $\hat{\mathbf{T}}(s)$ be a differentiable function of the arc length counted from a particular point of C . The number

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

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

Let $\mathbf{r}(s)$ be the natural parameterization of a smooth curve (the parameter s is the arc length measured from a particular point on the curve). Then $\mathbf{r}'(s) = \hat{\mathbf{T}}(s)$, as shown in the previous section, and therefore $\kappa(s) = \|\mathbf{r}''(s)\|$.

EXAMPLE 12.15. Find the curvature of a helix of radius R that rises a distance h per turn.

SOLUTION: In Example 12.13, the natural parameterization of the helix is obtained

$$\mathbf{r}(s) = \langle R \cos(2\pi s/L), R \sin(2\pi s/L), hs/L \rangle.$$

where $L = \sqrt{(2\pi R)^2 + h^2}$ is the arc length of one turn. Differentiating this vector function twice with respect to the arc length parameter s ,

$$\begin{aligned}\mathbf{r}''(s) &= \langle -(2\pi/L)^2 R \cos(2\pi s/L), -(2\pi/L)^2 R \sin(2\pi s/L), 0 \rangle \\ &= -(2\pi/L)^2 R \langle \cos(2\pi s/L), \sin(2\pi s/L), 0 \rangle, \\ \kappa(s) &= \|\mathbf{r}''(s)\| = (2\pi/L)^2 R = \frac{R}{R^2 + (h/2\pi)^2},\end{aligned}$$

where the relation $\|\langle \cos u, \sin u, 0 \rangle\| = 1$ has been used. So the helix has a constant curvature. \square

In practice, finding the natural parameterization of a smooth curve might be a tedious technical task. Therefore, a question of interest is to develop a method to calculate the curvature in any parameterization. Let $\mathbf{r}(t)$ be a vector function in $[a, b]$ that traces out a smooth curve C . 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. If $s = s(t)$ is the arc length as a function of t (see Definition 12.13), then, by the inverse function theorem (Theorem 12.6), there exists an inverse differentiable function $t = t(s)$ that expresses the parameter t as a function of the arc length s and $dt(s)/ds = 1/(ds(t)/dt) = 1/\|\mathbf{r}'(t)\|$. By the chain rule,

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

and therefore

$$(12.4) \quad \kappa(t) = \frac{\|\hat{\mathbf{T}}'(t)\|}{\|\mathbf{r}'(t)\|}.$$

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 rather technical task, too. The following theorem provides a more convenient way to calculate the curvature.

THEOREM 12.7. (Curvature of a Curve).

Let a smooth curve be traced out by a twice-differentiable vector function $\mathbf{r}(t)$. Then the curvature is

$$(12.5) \quad \kappa(t) = \frac{\|\mathbf{r}'(t) \times \mathbf{r}''(t)\|}{\|\mathbf{r}'(t)\|^3}.$$

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

$$(12.6) \quad \mathbf{r}''(t) = v'(t)\hat{\mathbf{T}}(t) + v(t)\hat{\mathbf{T}}'(t) = \frac{v'(t)}{v(t)}\mathbf{r}'(t) + v(t)\hat{\mathbf{T}}'(t).$$

Since the cross product of two parallel vectors vanishes, it follows from (12.6) that

$$(12.7) \quad \mathbf{r}'(t) \times \mathbf{r}''(t) = v(t)\left(\mathbf{r}'(t) \times \hat{\mathbf{T}}'(t)\right).$$

Now recall that $\|\mathbf{a} \times \mathbf{b}\| = \|\mathbf{a}\|\|\mathbf{b}\|\sin\theta$, where θ is the angle between vectors \mathbf{a} and \mathbf{b} . Therefore,

$$(12.8) \quad \|\mathbf{r}'(t) \times \mathbf{r}''(t)\| = v(t)\|\mathbf{r}'(t) \times \hat{\mathbf{T}}'(t)\| = \|\mathbf{r}'(t)\|^2\|\hat{\mathbf{T}}'(t)\|\sin\theta,$$

where θ is the angle between $\hat{\mathbf{T}}'(t)$ and the tangent vector $\mathbf{r}'(t)$. Since $\hat{\mathbf{T}}(t)$ is a unit vector, one has $\|\hat{\mathbf{T}}(t)\|^2 = \hat{\mathbf{T}}(t) \cdot \hat{\mathbf{T}}(t) = 1$. By taking the derivative of both sides of the latter relation, it is concluded that the vectors $\hat{\mathbf{T}}'(t)$ and $\mathbf{r}'(t)$ are orthogonal:

$$\hat{\mathbf{T}}'(t) \cdot \hat{\mathbf{T}}(t) = 0 \Leftrightarrow \hat{\mathbf{T}}'(t) \perp \hat{\mathbf{T}}(t) \Leftrightarrow \hat{\mathbf{T}}'(t) \perp \hat{\mathbf{r}}'(t) \Leftrightarrow \theta = \frac{\pi}{2}$$

because $\mathbf{r}'(t)$ is parallel to $\hat{\mathbf{T}}(t)$. Hence, $\sin\theta = 1$. The substitution of the latter relation and $\|\hat{\mathbf{T}}'(t)\| = \kappa(t)\|\mathbf{r}'(t)\|$ (see (12.4)) into (12.8) yields $\|\mathbf{r}' \times \mathbf{r}''\| = \|\mathbf{r}'\|^3\kappa$ from which (12.5) follows. \square

EXAMPLE 12.16. 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 $\kappa(1)$:

$$\begin{aligned} \mathbf{r}'(1) &= \langle t^{-1}, 2t, 2 \rangle \Big|_{t=1} = \langle 1, 2, 2 \rangle \Rightarrow \|\mathbf{r}'(1)\| = 3, \\ \mathbf{r}''(1) &= \langle -t^{-2}, 2, 0 \rangle \Big|_{t=1} = \langle -1, 2, 0 \rangle, \\ \mathbf{r}'(1) \times \mathbf{r}''(1) &= \langle -4, -2, 4 \rangle = 2\langle -2, -1, 2 \rangle, \\ \kappa(1) &= \frac{\|\mathbf{r}'(1) \times \mathbf{r}''(1)\|}{\|\mathbf{r}'(1)\|^3} = \frac{2\|\langle -2, -1, 2 \rangle\|}{3^3} = \frac{6}{27} = \frac{2}{9}. \end{aligned}$$

\square

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:

$\mathbf{r}' \times \mathbf{r}'' = \langle 0, 0, x'y'' - x''y' \rangle$. The substitution of this relation into (12.5) leads to the following result.

COROLLARY 12.1. (Curvature of a Planar Curve).

The curvature of a planar smooth curve $\mathbf{r}(t) = \langle x(t), y(t), 0 \rangle$ is

$$\kappa = \frac{|x'y'' - x''y'|}{[(x')^2 + (y')^2]^{3/2}}.$$

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 Corollary 12.1, $x'(t) = 1$, $x''(t) = 0$, 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

$$\kappa(x) = \frac{|f''(x)|}{[1 + (f'(x))^2]^{3/2}}.$$

83.1. Geometrical Significance of the Curvature. Let us calculate the curvature of a circle of radius R . One has

$$\begin{aligned} x(t) = R \cos t & \Rightarrow x'(t) = -R \sin t & \Rightarrow x''(t) = -R \cos t \\ y(t) = R \sin t & \Rightarrow y'(t) = R \cos t & \Rightarrow y''(t) = -R \sin t \end{aligned}$$

By Corollary 12.1,

$$\begin{aligned} (x')^2 + (y')^2 = R^2 & \Rightarrow \kappa = \frac{R^2}{R^3} = \frac{1}{R}. \\ x'y'' - x''y' = R^2 & \end{aligned}$$

Therefore, the curvature is constant along the circle and equals a reciprocal of its radius. 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 $\rho(t) = 1/\kappa(t)$.

The curvature radius is a function of a point on the curve. 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 the curve better than the tangent line at

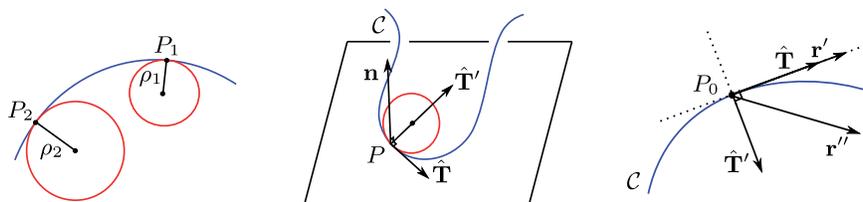


FIGURE 12.10. **Left:** Curvature radius. A smooth curve near a point P can be approximated by a portion of a circle of radius $\rho = 1/\kappa$. The curve bends in the same way as a circle of radius that is the reciprocal of the curvature. A large curvature at a point corresponds to a small curvature 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 orthogonal 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 at 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 .

P because the circle and the curve are equally bent at P and do not just have the same unit tangent vector 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 space curve, not every circle of radius $\rho = 1/\kappa$ that passes through P would approximate well the curve near P . Let $\mathbf{r}(s)$ be a natural parameterization of a curve such that the position vector of P is $\mathbf{r}(0)$ (i.e., the arc length parameter is measured from P). Consider a plane containing the tangent line to the curve at P . Let $\mathbf{R}(s)$ be a natural parameterization of the circle of radius ρ that lies in the plane and passes through P so that $\mathbf{r}(0) = \mathbf{R}(0)$. The plane and the circle in it can be rotated about the tangent line. The deviation $\|\mathbf{r}(s) - \mathbf{R}(s)\|$ (or the approximation error) in a small fixed interval depends on the orientation of the plane. Think about a portion of the curve of length s and a portion of the circle of the same length that have a common endpoint (the point P) so that the part of the circle can be rigidly rotated about the tangent line through P . For s small enough, the error is roughly determined by the distance between the other ends. Now

recall from Calculus I that a twice-differentiable function can be well approximated by its second Taylor polynomial in a sufficiently small neighborhood of any point: $x(s) \approx x(0) + x'(0)s + x''(0)s^2/2 = T_2(s)$ and the approximation error decreases to 0 faster than s^2 with decreasing s , that is, $(x(s) - T_2(s))/s^2 \rightarrow 0$ as $s \rightarrow 0$. Using the Taylor approximation for each component of the vector functions $\mathbf{r}(s)$ and $\mathbf{R}(s)$, one finds

$$\begin{aligned}\mathbf{r}(s) &\approx \mathbf{r}(0) + \mathbf{r}'(0)s + \mathbf{r}''(0)s^2/2, \\ \mathbf{R}(s) &\approx \mathbf{R}(0) + \mathbf{R}'(0)s + \mathbf{R}''(0)s^2/2.\end{aligned}$$

Since the curve and the circle have a common point at $s = 0$, $\mathbf{r}(0) = \mathbf{R}(0)$. Furthermore, they have the same unit tangent vector at the common point and, hence $\mathbf{r}'(0) = \mathbf{R}'(0)$. For a natural parameterization, the unit tangent vectors to the curve and the circle are given by the derivatives $\mathbf{r}'(s)$ and $\mathbf{R}'(s)$. Consequently, the approximation error $\|\mathbf{r}(s) - \mathbf{R}(s)\| \approx \|\mathbf{r}''(0) - \mathbf{R}''(0)\|s^2/2$ will be minimal if the curve and the circle have the same derivative of the unit tangent vector at the common point, $\mathbf{r}''(0) = \mathbf{R}''(0)$. More accurately, the approximation error will decrease to 0 faster than s^2 with decreasing s in this case. Since the unit tangent vector and its derivative are orthogonal (see the proof of Theorem 12.7), they lie in one particular plane through the point $\mathbf{r}(0)$. Thus, the best approximation of the curve by a circle of radius $1/\kappa(0)$ at P is achieved when the circle lies in the plane through P that is parallel to the tangent vector $\hat{\mathbf{T}}(0) = \mathbf{r}'(0)$ and its derivative $\hat{\mathbf{T}}'(0) = \mathbf{r}''(0)$ at P . The normal of this plane is $\mathbf{n} = \hat{\mathbf{T}}(0) \times \hat{\mathbf{T}}'(0)$. By the geometrical interpretation of the derivative, $\hat{\mathbf{T}}'(0)$ should point in the direction in which the curve bends (see Figure 12.9). Therefore, the center of the circle must be in the direction of $\hat{\mathbf{T}}'(0)$ from P , not in the opposite one.

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 .

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. \square

EXAMPLE 12.17. 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 $t = 1$. Then

$$\begin{aligned}\mathbf{r}'(1) &= \langle 1, 2t, 3t^2 \rangle \Big|_{t=1} = \langle 1, 2, 3 \rangle, \\ \mathbf{r}''(1) &= \langle 0, 2, 6t \rangle \Big|_{t=1} = \langle 0, 2, 6 \rangle.\end{aligned}$$

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 $3x - 3y + z = 1$. \square

EXAMPLE 12.18. Find the osculating circle for the graph $y = \cos(2x)$ at the point $(0, 1)$.

SOLUTION: For a planar curve, the osculating plane is the plane in which the curve lies. Since $y'(0) = -2\sin 0 = 0$, the tangent line to the graph is horizontal, $y = 1$, and the y axis is the line normal to the tangent line at $(0, 1)$. Therefore, the center of the osculating circle lies on the y axis down from $y = 1$ because the graph of $\cos(2x)$ is concave downward. The curvature of the graph at $x = 0$ is found by Corollary 12.2: $y'(x) = -2\sin(2x)$, $y''(x) = -4\cos(2x)$, and $\kappa(0) = |-4| = 4$. So the curvature radius is $\rho(0) = 1/\kappa(0) = 1/4$. The center of the osculating circle lies $1/4$ of unit length down from $y = 1$, that is, at $(0, 3/4)$. The equation of the osculating circle is $x^2 + (y - 3/4)^2 = (1/4)^2$. \square

83.2. Study Problems.

Problem 12.12. Show that any smooth curve whose curvature vanishes is a straight line.

SOLUTION: Let $\mathbf{r}(s)$ be a natural parameterization of a smooth curve. Then the derivative is a unit tangent vector to the curve, $\hat{\mathbf{T}}(s) = \hat{\mathbf{r}}'(s)$. By the definition of the curvature, $\kappa(s) = \|\hat{\mathbf{T}}'(s)\| = \|\mathbf{r}''(s)\|$. If $\kappa(s) = 0$, then $\mathbf{r}''(s) = \mathbf{0}$ for all s . Therefore, the unit tangent vector $\mathbf{r}'(s) = \hat{\mathbf{T}}$ is a constant vector. The integration of this relation yields

$\mathbf{r}(s) = \mathbf{r}_0 + \hat{\mathbf{T}}s$, which is the vector equation of a line through the point \mathbf{r}_0 and parallel to $\hat{\mathbf{T}}$. \square

Problem 12.13. 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 calculated by Corollary 12.2: $\kappa(x) = e^x/(1 + e^{2x})^{3/2}$. Critical points are determined by $\kappa'(x) = 0$ or

$$\begin{aligned} \kappa'(x) = \frac{e^x(1 + e^{2x})^{1/2}[2e^{2x} - 1]}{(1 + e^{2x})^3} = 0 &\Rightarrow 2e^{2x} - 1 = 0 \\ &\Rightarrow x = -\frac{\ln 2}{2}. \end{aligned}$$

From the shape of the graph of the exponential, it is clear that the found critical point corresponds to the (absolute) maximum of $\kappa(x)$ (maximal bending) and $\kappa_{\max} = \kappa(-\ln(2)/2) = 2/3^{3/2}$. \square

Problem 12.14. (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, $\mathbf{R}(t) - \mathbf{R}_0 = a(t)\hat{\mathbf{N}}_0 + b(t)\hat{\mathbf{T}}_0$. Since $\mathbf{R}(t)$ traverses a circle of radius ρ_0 centered at \mathbf{R}_0 , the length of $\mathbf{R}(t) - \mathbf{R}_0$ must be ρ_0 for any t . Owing to the orthogonality of the unit vectors $\hat{\mathbf{T}}_0$ and $\hat{\mathbf{N}}_0$, this condition yields:

$$\|\mathbf{R}(t) - \mathbf{R}_0\|^2 = \rho_0^2 \iff a^2(t) + b^2(t) = \rho_0^2$$

by the Pythagorean theorem. Parametric equations of the circle can be taken in the form $a(t) = -\rho_0 \cos t$ and $b(t) = \rho_0 \sin t$. The vector function that traces out the osculating circle is

$$\mathbf{R}(t) = \mathbf{r}_0 + \rho_0(1 - \cos t)\hat{\mathbf{N}}_0 + \rho_0 \sin t \hat{\mathbf{T}}_0,$$

where $t \in [0, 2\pi]$. The above choice of $a(t)$ and $b(t)$ has been made so that $\mathbf{R}(0) = \mathbf{r}_0$. \square

Problem 12.15. 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 $\omega \rightarrow \infty$.

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 $\omega \rightarrow \infty$, 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, $L(\omega) \rightarrow 2\pi R$ (the circumference) and $\kappa(\omega) \rightarrow 1/R$ (the curvature of the circle) as $\omega \rightarrow \infty$.

A calculus approach requires a lot more work to establish this result:

$$\begin{aligned} L(\omega) &= \int_0^{2\pi/\omega} \|\mathbf{r}'(t)\| dt = \frac{2\pi}{\omega} \sqrt{(R\omega)^2 + h^2} \\ &= 2\pi \sqrt{R^2 + (h/\omega)^2} \rightarrow 2\pi R, \\ \kappa(\omega) &= \frac{\|\mathbf{r}'(t) \times \mathbf{r}''(t)\|}{\|\mathbf{r}'(t)\|^3} = \frac{R\omega^2[(R\omega)^2 + h^2]^{1/2}}{[(R\omega)^2 + h^2]^{3/2}} \\ &= \frac{R}{R^2 + (h/\omega)^2} \rightarrow \frac{1}{R} \end{aligned}$$

as $\omega \rightarrow \infty$. □

83.3. Exercises.

(1) Find the curvature of each of the following curves as a function of the parameter and the curvature radius at a specified point P :

- (i) $\mathbf{r}(t) = \langle t, 1 - t, t^2 + 1 \rangle$, $P(1, 0, 2)$
- (ii) $\mathbf{r}(t) = \langle t^2, t, 1 \rangle$, $P(4, 2, 1)$
- (iii) $y = \sin(x/2)$, $P(\pi, 1)$
- (iv) $\mathbf{r}(t) = \langle 4t^{3/2}, -t^2, t \rangle$, $P(4, -1, 1)$
- (v) $x = 1 + t^2$, $y = 2 + t^3$, $P(2, 1)$
- (vi) $x = e^t \cos t$, $y = 0$, $z = e^t \sin t$, $P(1, 0, 0)$
- (vii) $\mathbf{r}(t) = \langle \ln t, \sqrt{t}, t^2 \rangle$, $P(0, 1, 1)$
- (viii) $\mathbf{r}(t) = \langle \cosh t, \sinh t, 2 + t \rangle$, $P(1, 0, 2)$
- (ix) $\mathbf{r}(t) = \langle e^t, \sqrt{2}t, e^{-t} \rangle$, $P(1, 0, 1)$
- (x) $\mathbf{r}(t) = \langle \sin t - t \cos t, t^2, \cos t + t \sin t \rangle$, $P(0, 0, 1)$

(2) Find the curvature of $\mathbf{r}(t) = \langle t, t^2/2, t^3/3 \rangle$ at the point of its intersection with the surface $z = 2xy + 1/3$.

- (3) Find the maximal and minimal curvatures of the graph $y = \cos(ax)$ and the points at which they occur. Sketch the graph for $a = 1$ and mark the points of the maximal and minimal curvatures, local maxima and minima of $\cos x$, and the inflection points.
- (4) 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.
- (5) Let $f(x)$ be a twice continuously differentiable function and let $\kappa(x)$ be the curvature of the graph $y = f(x)$.
- (i) Does κ attain a local maximum value at every local minimum and maximum of f ? If not, state an additional condition on f under which the answer to this question is affirmative.
- (ii) Prove that $\kappa = 0$ at inflection points of the graph.
- (iii) Show by an example that the converse is not true, that is, that the curvature vanishes at $x = x_0$ does not imply that the point $(x_0, f(x_0))$ is an inflection point.
- (6) Let f be twice differentiable at x_0 . Let $T_2(x)$ be its Taylor polynomial of the second order about $x = x_0$. Compare the curvatures of the graphs $y = f(x)$ and $y = T_2(x)$ at $x = x_0$.
- (7) Find the equation of the osculating circles to the parabola $y = x^2$ at the points $(0, 0)$ and $(1, 1)$.
- (8) Find the maximal and minimal curvature of the ellipse $x^2/a^2 + y^2/b^2 = 1$, $a > b$, and the points where they occur. Give the equations of the osculating circles at these points.
- (9) Let $\mathbf{r}(t) = \langle t^3, t^2, 0 \rangle$. This curve is not smooth and has a cusp at $t = 0$. Find the curvature for $t \neq 0$ and investigate its limit as $t \rightarrow 0$.
- (10) Find an equation of the osculating plane for each of the following curves at a specified point:
- (i) $\mathbf{r}(t) = \langle 4t^{3/2}, -t^2, t \rangle$, $P(4, -1, 1)$
- (ii) $\mathbf{r}(t) = \langle \ln t, \sqrt{t}, t^2 \rangle$, $P(0, 1, 1)$
- (11) 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.
- (12) Is there a point on the curve $\mathbf{r}(t) = \langle t, t^2, t^3 \rangle$ where the osculating plane is parallel to the plane $3x + 3y + z = 1$?
- (13) Prove that the trajectory of a particle has a constant curvature if the particle moves so that the magnitudes of its velocity and acceleration vectors are constant.
- (14) Consider a graph $y = f(x)$ such that $f''(x_0) \neq 0$. At a point

(x_0, y_0) on the curve, where $y_0 = f(x_0)$, find the equation of the osculating circle in the form $(x - a)^2 + (y - b)^2 = R^2$. *Hint:* Show first that the vector $\langle 1, f'(x_0) \rangle$ is tangent to the graph and a vector orthogonal to it is $\langle -f'(x_0), 1 \rangle$. Then consider two cases $f''(x_0) > 0$ and $f''(x_0) < 0$.

(15) Find the osculating circle for the cycloid $x = a(t - \sin t)$, $y = a(1 - \cos t)$ at the point $t = \pi/2$.

(16) Let a smooth curve $\mathbf{r} = \mathbf{r}(t)$ be planar and lie in the xy plane. At a point (x_0, y_0) on the curve, find the equation of the osculating circle in the form $(x - a)^2 + (y - b)^2 = R^2$. *Hint:* Use the result of Study Problem 12.14 to express the constants a , b , and R via x_0 , y_0 , and the curvature at (x_0, y_0) .

84. Applications to Mechanics and Geometry

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 orthogonal 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 $\hat{\mathbf{N}}(t)$. The differentiation of the relation $\mathbf{v}(t) = v(t)\hat{\mathbf{T}}(t)$ (see (12.6)) shows that that acceleration always lies in the osculating plane:

$$\mathbf{a} = v'\hat{\mathbf{T}} + v\hat{\mathbf{T}}' = v'\hat{\mathbf{T}} + v\|\hat{\mathbf{T}}'\|\hat{\mathbf{N}}.$$

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

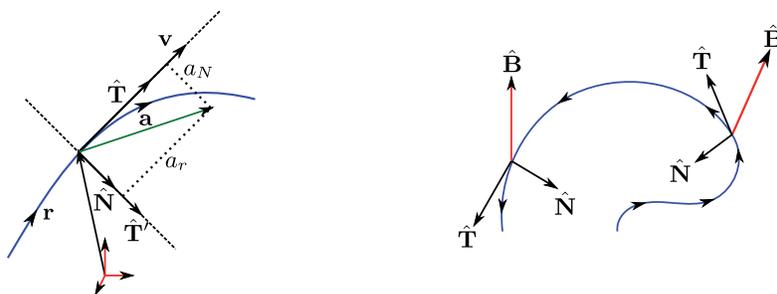


FIGURE 12.11. **Left:** Decomposition of the acceleration \mathbf{a} of a particle into normal and tangential components. The tangential component a_T is the scalar projection of \mathbf{a} onto the unit tangent vector $\hat{\mathbf{T}}$. The normal component is the scalar projection of \mathbf{a} onto 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.

the normal acceleration is constant, $a_N = v_0^2/R$, where R is the circle radius. Indeed, taking the derivative of the relation $\mathbf{v} \cdot \mathbf{v} = v_0^2$, it is concluded that $\mathbf{v}' \cdot \mathbf{v} = 0$ or $\mathbf{a} \cdot \mathbf{v} = 0$ or $a_T = 0$. Since the curvature of a circle is the reciprocal of its radius, $a_N = \kappa v^2 = v_0^2/R$.

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*. By Newton's law, its magnitude is

$F = ma_N = mv^2/\rho$, where m is the mass of a moving object (e.g., a car). 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 12.19. *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 $a_N(x) \geq a_m$ yields

$$\frac{v_0^2}{R} \frac{1}{[1 + (x/R)^2]^{3/2}} \geq a_m \quad \Rightarrow \quad |x| \leq R\sqrt{\nu - 1}, \quad \nu = \left(\frac{v_0^2}{Ra_m}\right)^{2/3}.$$

The constant $\nu = a_N(0)/a_m$ always exceeds 1 if $a_N(0) = v_0^2/R > a_m$. The car can skid off the road when moving on its part corresponding to the interval $-R(\nu - 1)^{1/2} \leq x \leq R(\nu - 1)^{1/2}$. Conversely, a suggested speed limit sign can be placed $v_0 < \sqrt{Ra_m}$ for a part of the road that contains $x = 0$. \square

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 of mutually orthogonal unit 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 $\hat{\mathbf{e}}_1$, $\hat{\mathbf{e}}_2$, and $\hat{\mathbf{e}}_3$, respectively. Indeed, $\hat{\mathbf{T}}$ and $\hat{\mathbf{N}}$ can always be made parallel to $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$. Then, owing to the relation $\hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_2 = \hat{\mathbf{e}}_3$, the binormal must be parallel to $\hat{\mathbf{e}}_3$. In other words, $\hat{\mathbf{T}}$, $\hat{\mathbf{N}}$, and $\hat{\mathbf{B}}$ define a *right-handed* coordinate system. 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

$$\frac{d\hat{\mathbf{B}}(s)}{ds} = -\tau(s)\hat{\mathbf{N}}(s),$$

where 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. *A planar curve has no torsion.* 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 $\hat{\mathbf{e}}_2 = \hat{\mathbf{e}}_3 \times \hat{\mathbf{e}}_1$) 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 or equations*:

$$(12.9) \quad \hat{\mathbf{T}}'(s) = \kappa(s)\hat{\mathbf{N}}(s),$$

$$(12.10) \quad \hat{\mathbf{N}}'(s) = -\kappa(s)\hat{\mathbf{T}}(s) + \tau(s)\hat{\mathbf{B}}(s),$$

$$(12.11) \quad \hat{\mathbf{B}}'(s) = -\tau(s)\hat{\mathbf{N}}(s).$$

The Frenet-Serret equations form a system of *differential equations* for the components of $\hat{\mathbf{T}}(s)$, $\hat{\mathbf{N}}(s)$, and $\hat{\mathbf{B}}(s)$. If the curvature and torsion are continuous functions on an interval $0 \leq s \leq L$, then the system can be proved to have a unique solution on this interval for every given set of the vectors $\hat{\mathbf{T}}(0)$, $\hat{\mathbf{N}}(0)$, and $\hat{\mathbf{B}}(0)$ at an initial point of the curve. Given a coordinate system, the initial point of a curve is specified by a translation of the origin, and the orientation of $\hat{\mathbf{T}}(0)$, $\hat{\mathbf{N}}(0)$, and $\hat{\mathbf{B}}(0)$ is determined by a rotation of unit coordinate vectors. Therefore, the following assertion about the shape of a curve holds.

THEOREM 12.9. (Shape of a Smooth Curve in Space).

Given the curvature and torsion as continuous functions along a curve, the curve is uniquely determined by them up to rigid rotations and translations of the curve as a whole.

A proof of Theorem 12.9 requires a proof of the uniqueness of a solution to the Frenet-Serret equations, which goes beyond the scope of this course. However, in some specific examples, the Frenet-Serret equations can be explicitly integrated. For example, consider curves with the vanishing curvature and torsion, $\kappa(s) = \tau(s) = 0$. Then $\hat{\mathbf{T}}(s) = \hat{\mathbf{T}}(0)$, $\hat{\mathbf{N}}(s) = \hat{\mathbf{N}}(0)$, and $\hat{\mathbf{B}}(s) = \hat{\mathbf{B}}(0)$. If $\mathbf{r}(s)$ is a natural parameterization of a curve, then $\mathbf{r}'(s) = \hat{\mathbf{T}}(s) = \hat{\mathbf{T}}(0)$. The integration of this equation yields $\mathbf{r}(s) = \mathbf{r}_0 + \hat{\mathbf{T}}(0)s$, where \mathbf{r}_0 is a constant vector, which is a straight line.

EXAMPLE 12.20. Use the Frenet-Serret equations to 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: A vector function $\mathbf{r}(s)$ that satisfies the Frenet-Serret equations is sought in the basis of the initial tangent, normal, and binormal vectors: $\hat{\mathbf{e}}_1 = \hat{\mathbf{T}}(0)$, $\hat{\mathbf{e}}_2 = \hat{\mathbf{N}}(0)$, and $\hat{\mathbf{e}}_3 = \hat{\mathbf{B}}(0)$. Since the torsion is 0, the binormal does not change along the curve, $\hat{\mathbf{B}}(s) = \hat{\mathbf{e}}_3$. The curve is planar and lies in a plane orthogonal to $\hat{\mathbf{e}}_3$. Any unit vector $\hat{\mathbf{T}}$ orthogonal to $\hat{\mathbf{e}}_3$ can be written as $\hat{\mathbf{T}} = \cos \varphi \hat{\mathbf{e}}_1 + \sin \varphi \hat{\mathbf{e}}_2$ where $\varphi = \varphi(s)$ such that $\varphi(0) = 0$. Owing to the relations $\hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_2 = -\hat{\mathbf{e}}_2 \times \hat{\mathbf{e}}_1 = \hat{\mathbf{e}}_3$, a unit vector $\hat{\mathbf{N}}$ orthogonal to $\hat{\mathbf{T}}$ such that $\hat{\mathbf{T}} \times \hat{\mathbf{N}} = \hat{\mathbf{B}} = \hat{\mathbf{e}}_3$ must have the form $\hat{\mathbf{N}} = -\sin \varphi \hat{\mathbf{e}}_1 + \cos \varphi \hat{\mathbf{e}}_2$. Equation (12.9) gives

$$\hat{\mathbf{T}}' = -\varphi' \sin \varphi \hat{\mathbf{e}}_1 + \varphi' \cos \varphi \hat{\mathbf{e}}_2 = \varphi' \hat{\mathbf{N}} = \kappa_0 \hat{\mathbf{N}} \Rightarrow \varphi'(s) = \kappa_0,$$

and therefore $\varphi(s) = \kappa_0 s$ because $\varphi(0) = 0$. For a natural parameterization of the curve, $\mathbf{r}'(s) = \hat{\mathbf{T}}(s)$. Hence,

$$\begin{aligned} \mathbf{r}'(s) &= \cos(\kappa_0 s) \hat{\mathbf{e}}_1 + \sin(\kappa_0 s) \hat{\mathbf{e}}_2, \\ \mathbf{r}(s) &= \mathbf{r}_0 + \kappa_0^{-1} \sin(\kappa_0 s) \hat{\mathbf{e}}_1 - \kappa_0^{-1} \cos(\kappa_0 s) \hat{\mathbf{e}}_2, \end{aligned}$$

where \mathbf{r}_0 is a constant vector. By the Pythagorean theorem, the distance between any point of the curve and a fixed point \mathbf{r}_0 is constant: $\|\mathbf{r}(s) - \mathbf{r}_0\|^2 = 1/\kappa_0^2 = R^2$. Since the curve is planar, it is a circle (or its portion) of radius $R = 1/\kappa_0$. \square

THEOREM 12.10. (Torsion of a Curve).

Let $\mathbf{r}(t)$ be a three times differentiable vector function that traverses a smooth curve whose curvature does not vanish. Then the torsion of the curve is

$$\tau(t) = \frac{(\mathbf{r}'(t) \times \mathbf{r}''(t)) \cdot \mathbf{r}'''(t)}{\|\mathbf{r}'(t) \times \mathbf{r}''(t)\|^2}.$$

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,

$$(12.12) \quad \mathbf{r}'' = v' \hat{\mathbf{T}} + \kappa v^2 \hat{\mathbf{N}},$$

and by (12.7) and the definition of the binormal,

$$(12.13) \quad \mathbf{r}' \times \mathbf{r}'' = v \hat{\mathbf{T}} \times \mathbf{r}'' = \kappa v^3 \hat{\mathbf{B}}.$$

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,

$$(12.14) \quad \mathbf{r}''' = (v'' - \kappa^2 v^3) \hat{\mathbf{T}} + (3\kappa v v' + \kappa' v^2) \hat{\mathbf{N}} + \kappa \tau v^3 \hat{\mathbf{B}}.$$

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, $\kappa = \|\mathbf{r}' \times \mathbf{r}''\|/v^3$. \square

Remark. Relation (12.13) shows that $\hat{\mathbf{B}}$ is the unit vector in the direction of $\mathbf{r}' \times \mathbf{r}''$. This observations offers a more convenient way for calculating the unit binormal vector 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

$$\hat{\mathbf{T}}(t_0) = \frac{\mathbf{r}'(t_0)}{\|\mathbf{r}'(t_0)\|}, \quad \hat{\mathbf{B}}(t_0) = \frac{\mathbf{r}'(t_0) \times \mathbf{r}''(t_0)}{\|\mathbf{r}'(t_0) \times \mathbf{r}''(t_0)\|}, \quad \hat{\mathbf{N}}(t_0) = \hat{\mathbf{B}}(t_0) \times \hat{\mathbf{T}}(t_0).$$

EXAMPLE 12.21. Find the unit tangent, normal, and binormal vectors and the torsion of the curve $\mathbf{r}(t) = \langle \ln t, t, t^2/2 \rangle$ at the point $(0, 1, 1/2)$.

SOLUTION: The point in question corresponds to $t = 1$. Therefore,

$$\mathbf{r}'(1) = \langle t^{-1}, 1, t \rangle \Big|_{t=1} = \langle 1, 1, 1 \rangle \quad \Rightarrow \quad \|\mathbf{r}'(1)\| = \sqrt{3},$$

$$\mathbf{r}''(1) = \langle -t^{-2}, 0, 1 \rangle \Big|_{t=1} = \langle -1, 0, 1 \rangle,$$

$$\mathbf{r}'''(1) = \langle 2t^{-3}, 0, 0 \rangle \Big|_{t=1} = \langle 2, 0, 0 \rangle,$$

$$\mathbf{r}'(1) \times \mathbf{r}''(1) = \langle 1, -2, 1 \rangle \quad \Rightarrow \quad \|\mathbf{r}'(1) \times \mathbf{r}''(1)\| = \sqrt{6},$$

$$\begin{aligned}\hat{\mathbf{T}}(1) &= \frac{1}{\sqrt{3}} \langle 1, 1, 1 \rangle, \\ \hat{\mathbf{B}}(1) &= \frac{1}{\sqrt{6}} \langle 1, -2, 1 \rangle, \\ \hat{\mathbf{N}}(1) &= \frac{1}{\sqrt{6}\sqrt{3}} \langle 1, -2, 1 \rangle \times \langle 1, 1, 1 \rangle = \frac{1}{3\sqrt{2}} \langle -3, 0, 3 \rangle, \\ &= \frac{1}{\sqrt{2}} \langle -1, 0, 1 \rangle \\ \tau(1) &= \frac{(\mathbf{r}'(1) \times \mathbf{r}''(1)) \cdot \mathbf{r}'''(1)}{\|\mathbf{r}'(1) \times \mathbf{r}''(1)\|^2} = \frac{2}{6} = \frac{1}{3}.\end{aligned}$$

□

84.3. Approximations of a Smooth Space Curve. A smooth curve C has a unit tangent vector at a point P . So a small part of the curve (a part of a small arc length Δs) containing P can be approximated by a piece of the tangent line of the same length Δs . If the curve C has a nonzero curvature at P , then a better approximation can be obtained by a part of the osculating circle of arc length Δs (see Study Problem 12.14). If the curve C has a nonzero torsion at P , an even more accurate approximation is provided by a curve through P that has the same unit tangent vector at P , and constant curvature and torsion equal to the curvature and torsion of the curve C at P . By Theorem 12.9, such a curve is unique. As shown in Study Problem 12.18, it is a helix whose radius and length of each turn are uniquely determined by the curvature and torsion. These three successively more accurate approximations do not refer to any particular coordinate system or any particular parameterization of C as the approximation curves are fully determined as the point sets in space by the geometrical invariants of the curve C at P : the unit tangent vector, curvature, and torsion. An analogy can be made with the Taylor polynomial approximation of a function at a particular point. The tangent line is the analog of the first-order Taylor polynomial (a linear approximation), the osculating circle is the analog of the second-order Taylor polynomial (a quadratic approximation), and the helix is the analog of the third-order Taylor polynomial (a cubic approximation). Given $\hat{\mathbf{T}}$, $\hat{\mathbf{N}}$, and $\hat{\mathbf{B}}$ of the curve C at P , the Frenet-Serret equations can be used to obtain unique higher-order approximations of C near P by approximating the curvature $\kappa(s)$ and the torsion $\tau(s)$ of C near P . The helix approximation uses the constant approximations of the curvature and torsion by their values at P . If, for example, the curvature and torsion of C is known at

two points near P , then $\kappa(s)$ and $\tau(s)$ can be approximated by linear functions near P that attain the two known values. The corresponding (unique) solution of the Frenet-Serret equations would generally provide a more accurate approximation than the helix approximation.

84.4. Study Problems.

Problem 12.16. 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 $t = 0$). \square

Problem 12.17. 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 12.16, 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, $\tau(1) = (\mathbf{v} \times \mathbf{a}) \cdot \mathbf{b} / \|\mathbf{v} \times \mathbf{a}\|^2 = -8/36 = -2/9$. \square

Problem 12.18. (Curves with Constant Curvature and Torsion).

Prove that all curves with a constant curvature $\kappa(s) = \kappa_0 \neq 0$ and a constant torsion $\tau(s) = \tau_0 \neq 0$ are helices by integrating the Frenet-Serret equations.

SOLUTION: 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 $\mathbf{w}' = \tau \hat{\mathbf{T}}' + \kappa \hat{\mathbf{B}}' = (\tau\kappa - \tau\kappa)\hat{\mathbf{N}} = \mathbf{0}$. By the Pythagorean theorem, $\|\mathbf{w}\| = (\kappa_0^2 + \tau_0^2)^{1/2}$. Consider two new unit vectors orthogonal to $\hat{\mathbf{N}}$:

$$\hat{\mathbf{w}} = \frac{1}{\|\mathbf{w}\|} \mathbf{w} = \sin \alpha \hat{\mathbf{T}} + \cos \alpha \hat{\mathbf{B}}, \quad \hat{\mathbf{u}} = \cos \alpha \hat{\mathbf{T}} - \sin \alpha \hat{\mathbf{B}},$$

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,

$$\hat{\mathbf{u}} \times \hat{\mathbf{w}} = \cos^2 \alpha \hat{\mathbf{T}} \times \hat{\mathbf{B}} - \sin^2 \alpha \hat{\mathbf{B}} \times \hat{\mathbf{T}} = (\cos^2 \alpha + \sin^2 \alpha)\hat{\mathbf{N}} = \hat{\mathbf{N}}.$$

By differentiating the vector $\hat{\mathbf{u}}$ and using the Frenet-Serret equations,

$$\hat{\mathbf{u}}' = \cos \alpha \hat{\mathbf{T}}' - \sin \alpha \hat{\mathbf{B}}' = (\kappa_0 \cos \alpha + \tau_0 \sin \alpha)\hat{\mathbf{N}} = \omega \hat{\mathbf{N}}.$$

Since $\hat{\mathbf{w}}(s) = \hat{\mathbf{w}}(0)$ is a constant unit vector, it is convenient to seek a solution in an orthonormal basis such that $\hat{\mathbf{e}}_3 = \hat{\mathbf{w}}(0)$ and $\hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_2 = \hat{\mathbf{e}}_3$. In this basis, $\hat{\mathbf{u}} = \cos \varphi \hat{\mathbf{e}}_1 + \sin \varphi \hat{\mathbf{e}}_2$, where $\varphi = \varphi(s)$, as a unit vector in the plane orthogonal to $\hat{\mathbf{e}}_3$. The orientation of the basis vectors in the plane orthogonal to $\hat{\mathbf{e}}_3$ is defined up to a general rotation about $\hat{\mathbf{e}}_3$. This freedom is used to set $\hat{\mathbf{e}}_1 = \hat{\mathbf{u}}(0)$, which implies that the function $\varphi(s)$ satisfies the condition $\varphi(0) = 0$. Then the unit normal vector in this basis is

$$\hat{\mathbf{N}} = \hat{\mathbf{u}} \times \hat{\mathbf{w}} = \cos \varphi \hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_3 + \sin \varphi \hat{\mathbf{e}}_2 \times \hat{\mathbf{e}}_3 = -\sin \varphi \hat{\mathbf{e}}_1 + \cos \varphi \hat{\mathbf{e}}_2$$

and

$$\hat{\mathbf{u}}' = -\varphi' \sin \varphi \hat{\mathbf{e}}_1 + \varphi' \cos \varphi \hat{\mathbf{e}}_2 = \varphi' \hat{\mathbf{N}}.$$

Hence, $\varphi'(s) = \omega$ or $\varphi(s) = \omega s$ owing to the condition $\varphi(0) = 0$. 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 12.20)

$$\mathbf{r}'(s) = \hat{\mathbf{T}}(s) = \frac{\kappa_0}{\omega} \cos(\omega s) \hat{\mathbf{e}}_1 + \frac{\kappa_0}{\omega} \sin(\omega s) \hat{\mathbf{e}}_2 + \frac{\tau_0}{\omega} \hat{\mathbf{e}}_3,$$

where $\mathbf{r}(s)$ is a natural parameterization of the curve. The integration of this equation gives

$$\mathbf{r}(s) = \mathbf{r}_0 + R \sin(\omega s) \hat{\mathbf{e}}_1 - R \cos(\omega s) \hat{\mathbf{e}}_2 + hs \hat{\mathbf{e}}_3, \quad R = \frac{\kappa_0}{\omega^2}, \quad h = \frac{\tau_0}{\omega}.$$

This is a helix of radius R whose axis goes through the point \mathbf{r}_0 parallel to $\hat{\mathbf{e}}_3$; the helix climbs along its axis by $2\pi h/\omega$ per each turn. \square

Problem 12.19. (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 12.11, 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, $\mathbf{B} = B\hat{\mathbf{e}}_3$, 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, $\mathbf{a} = \mu\mathbf{v} \times \mathbf{B} = \mu B\mathbf{v} \times \hat{\mathbf{e}}_3$, where $\mu = e/(mc)$. First, note that $v'_3 = a_3 = \hat{\mathbf{e}}_3 \cdot \mathbf{a} = 0$. Hence, $v_3 = v_{\parallel} = \text{const}$. Second, by the geometrical property of the cross product 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 $\mathbf{v} = \mathbf{v}_{\perp} + v_{\parallel}\hat{\mathbf{e}}_3$, 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, $\|\mathbf{v}_{\perp}\| = v_{\perp} = (v_0^2 - v_{\parallel}^2)^{1/2}$. The velocity vector can therefore be written in the form $\mathbf{v} = \langle v_{\perp} \cos \varphi, v_{\perp} \sin \varphi, v_{\parallel} \rangle$, where the function $\varphi = \varphi(t)$ is to be determined by the equations of motion:

$$\begin{aligned}\mathbf{a} &= \mu B \mathbf{v} \times \hat{\mathbf{e}}_3 = \mu B \langle -v_{\perp} \sin \varphi, v_{\perp} \cos \varphi, 0 \rangle, \\ \mathbf{a} &= \mathbf{v}' = \varphi' \langle -v_{\perp} \sin \varphi, v_{\perp} \cos \varphi, 0 \rangle.\end{aligned}$$

It follows from the comparison of these expressions that $\varphi'(t) = \mu B$ or $\varphi(t) = \mu B t + \varphi_0 = \omega t + \varphi_0$, where $\omega = eB/(mc)$ is the so-called cyclotron frequency and the integration constant φ_0 is determined by the initial velocity: $\mathbf{v}(0) = \langle v_{\perp} \cos \varphi_0, v_{\perp} \sin \varphi_0, v_{\parallel} \rangle$, that is, $\tan \varphi_0 = v_2(0)/v_1(0)$. Integration of the equation

$$\mathbf{r}'(t) = \mathbf{v}(t) = \langle v_{\perp} \cos(\omega t + \varphi_0), v_{\perp} \sin(\omega t + \varphi_0), v_{\parallel} \rangle$$

yields the trajectory of motion:

$$\mathbf{r}(t) = \mathbf{r}_0 + \langle R \sin(\omega t + \varphi_0), -R \cos(\omega t + \varphi_0), v_{\parallel} t \rangle,$$

where $R = v_{\perp}/\omega$. 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 $h = 2\pi v_{\parallel}/\omega$. In particular, if the initial

velocity is orthogonal to the magnetic field (i.e., $v_{\parallel} = 0$), then the trajectory is a circle of radius R .

The Polar Lights. The Sun produces a stream of charged particles (the solar wind). The magnetic field of the Earth plays the role of a shield from the solar wind as it traps the particles, forcing them to travel along its force lines that are arcs connecting the magnetic poles of the Earth (which approximately coincide with the south and north poles). As a result, the solar wind particles can penetrate the lower atmosphere only near the magnetic poles of the Earth, causing a spectacular phenomenon, the polar lights, by colliding with molecules of the oxygen and nitrogen in the atmosphere. \square

Problem 12.20. *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. \square

Problem 12.21. (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. The mass of the Sun is much larger than the mass of a planet; therefore, a displacement of the Sun due to the gravitational pull from a planet can be neglected (e.g., the Sun is about 332,946 times heavier than the Earth). 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 gravitational force is a central force, and, by Study Problem 12.20, 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

$$(12.15) \quad (\mathbf{v} \times \mathbf{l})' = GM\hat{\mathbf{r}}' \implies \mathbf{v} \times \mathbf{l} = GM\hat{\mathbf{r}} + \mathbf{c},$$

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 orthogonal to the constant vector \mathbf{l} because $\mathbf{l} = \mathbf{r} \times \mathbf{v}$ must be orthogonal to \mathbf{r} . It also follows from (12.15) and $\mathbf{l} \cdot \hat{\mathbf{r}} = 0$ that the constant vectors \mathbf{l} and \mathbf{c} are orthogonal because $\mathbf{l} \cdot \mathbf{c} = 0$. 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,

$$\mathbf{r} \cdot (\mathbf{v} \times \mathbf{l}) = \mathbf{l} \cdot (\mathbf{r} \times \mathbf{v}) = \mathbf{l} \cdot \mathbf{l} = l^2,$$

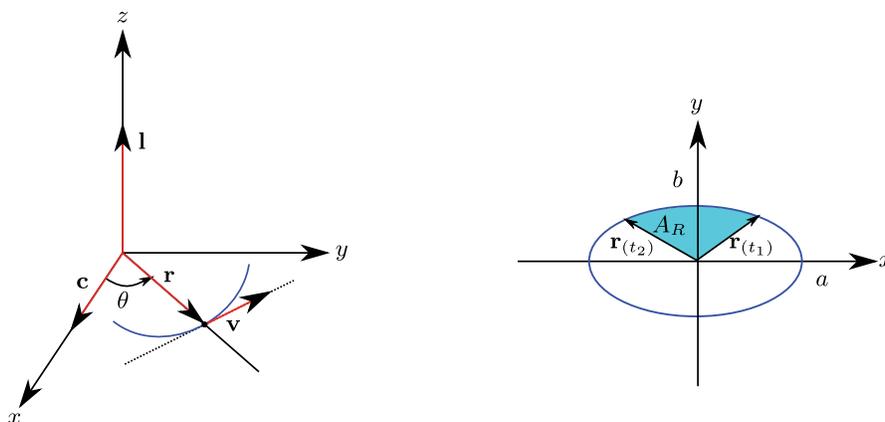


FIGURE 12.12. **Left:** The setup of the coordinate system for the derivation of Kepler's first law. **Right:** An illustration to the derivation of Kepler's second law.

where $l = \|\mathbf{l}\|$ is the length of \mathbf{l} . The comparison of the last two equations yields the equation for the trajectory:

$$l^2 = r(GM + b \cos \theta) \quad \implies \quad r = \frac{ed}{1 + e \cos \theta},$$

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,

$$\mathbf{l} = r^2(\hat{\mathbf{r}} \times \hat{\mathbf{r}}') = \langle 0, 0, r^2\theta' \rangle \quad \implies \quad l = r^2\theta'.$$

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 $b^2 = eda = l^2 a/(GM)$. Hence,

$$T^2 = \frac{4\pi^2 a^2 b^2}{l^2} = \frac{4\pi^2}{GM} a^3.$$

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). \square

84.5. Exercises.

(1) For each of the following trajectories of a particle, find the velocity, speed, and normal and tangential accelerations as functions of time and their values at a specified point P :

- (i) $\mathbf{r}(t) = \langle t, 1 - t, t^2 + 1 \rangle$, $P(1, 0, 2)$
- (ii) $\mathbf{r}(t) = \langle t^2, t, 1 \rangle$, $P(4, 2, 1)$
- (iii) $\mathbf{r}(t) = \langle 4t^{3/2}, -t^2, t \rangle$, $P(4, -1, 1)$
- (iv) $\mathbf{r}(t) = \langle \ln t, \sqrt{t}, t^2 \rangle$, $P(0, 1, 1)$
- (v) $\mathbf{r}(t) = \langle \cosh t, \sinh t, 2 + t \rangle$, $P(1, 0, 2)$
- (vi) $\mathbf{r}(t) = \langle e^t, \sqrt{2}t, e^{-t} \rangle$, $P(1, 0, 1)$
- (vii) $\mathbf{r}(t) = \langle \sin t - t \cos t, t^2, \cos t + t \sin t \rangle$, $P(0, 0, 1)$

(2) Find the normal and tangential accelerations of a particle with the position vector $\mathbf{r}(t) = \langle t^2 + 1, t, t^2 - 1 \rangle$ when the particle is closest to the origin.

(3) Find the tangential and normal accelerations of a particle with the position vector $\mathbf{r}(t) = \langle R \sin(\omega t + \varphi_0), -R \cos(\omega t + \varphi_0), v_0 t \rangle$, where R , ω , φ_0 , and v_0 are constants (see Study Problem 12.19).

(4) The shape of a winding road can be approximated by the graph $y = L \cos(x/L)$, where the coordinates are in meters and $L = 40\text{m}$. The condition of the road is such that if the normal acceleration of a car on it exceeds 10 m/s^2 , the car may skid off the road. Recommend

a speed limit for this portion of the road.

(5) A particle moves along the curve $y = x^2 + x^3$. If the acceleration of the particle at the point $(1, 2)$ is $\mathbf{a} = \langle 3, -1 \rangle$, find its normal and tangential accelerations.

(6) Suppose that a particle moves so that its tangential acceleration a_T is constant, while the normal acceleration a_N remains 0. What is the trajectory of the particle?

(7) Suppose that a particle moves in a plane so that its tangential acceleration a_T remains 0, while the normal acceleration a_N is constant. What is the trajectory of the particle? *Hint:* Investigate the curvature of the trajectory.

(8) A race car moves with a constant speed v_0 along an elliptic track $x^2/a^2 + y^2/b^2 = 1$, $a > b$. Find the maximal and minimal values of the magnitude of its acceleration and the points where they occur.

(9) Does there exist a curve with zero curvature and nonzero torsion? Explain the answer.

(10) For each of the following curves, find the unit tangent, normal, and binormal vectors and the torsion at a specified point P :

(i) $\mathbf{r}(t) = \langle t, 1 - t, t^2 + 1 \rangle$, $P(1, 0, 2)$

(ii) $\mathbf{r}(t) = \langle t^3, t^2, 1 \rangle$, $P(8, 4, 1)$

(iii) $\mathbf{r}(t) = \langle 4t^{3/2}, -t^2, t \rangle$, $P(4, -1, 1)$

(iv) $\mathbf{r}(t) = \langle \ln t, \sqrt{t}, t^2 \rangle$, $P(0, 1, 1)$

(v) $\mathbf{r}(t) = \langle \cosh t, \sinh t, 2 + t \rangle$, $P(1, 0, 2)$

(11) Let $\mathbf{r}(t) = \langle \cos t + t \sin t, \sin t + t \cos t, t^2 \rangle$. Find the speed, the tangential and normal accelerations, the curvature and torsion, and the unit tangent, normal, and binormal vectors as functions of time t .

Hint: To simplify calculations, find the decomposition $\mathbf{r}(t) = \mathbf{v}(t) - t\mathbf{w}(t) + t^2\hat{\mathbf{e}}_3$, where \mathbf{v} , \mathbf{w} , and $\hat{\mathbf{e}}_3$ are mutually orthogonal unit vectors such that $\mathbf{v}'(t) = \mathbf{w}(t)$, $\mathbf{w}'(t) = -\mathbf{v}(t)$. Use the properties of the cross products of mutually orthogonal unit vectors.

(12) Let C be the curve of intersection of an ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$ with the plane $2x - 2y + z = 0$. Find the torsion and the binormal $\hat{\mathbf{B}}$ along C .

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 $C = C(x, y, z, t)$ depends on four variables, three spatial coordinates and the time t . In general, if the value of a quantity f depends on values of several other quantities, say, x_1, x_2, \dots, x_n , this dependence is indicated by writing $f = f(x_1, x_2, \dots, x_n)$. In other words, $f = f(x_1, x_2, \dots, x_n)$ indicates a rule that assigns a number f to each ordered n -tuple of real numbers (x_1, x_2, \dots, x_n) . Each number in the n -tuple may be of a different nature and measured in different units. 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 (measured in units of length) and t is time (measured in units of time). All ordered n -tuples form an n -dimensional Euclidean space, much like all ordered doublets (x, y) form a plane, and all ordered triples (x, y, z) form a space.

85.1. 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, $\mathbf{a} = \mathbf{b} \Leftrightarrow \|\mathbf{a} - \mathbf{b}\| = 0$. 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 $\|\mathbf{a}\|^2 = \mathbf{a} \cdot \mathbf{a}$. Since points and vectors are described by the same mathematical object, an ordered triple (or pair) of numbers, it is not necessary to make a distinction between them. So, in what follows, the same notation is used for a point and a vector, for example, $\mathbf{r} = (x, y, z)$. These observations can be extended to ordered n -tuples for any n and lead to the notion of a *Euclidean space*.

DEFINITION 13.1. (Euclidean Space).

For each positive integer n , consider the set of all ordered n -tuples of real numbers. For any two elements $\mathbf{a} = (a_1, a_2, \dots, a_n)$ and $\mathbf{b} = (b_1, b_2, \dots, b_n)$ and a number c , put

$$\begin{aligned}\mathbf{a} + \mathbf{b} &= (a_1 + b_1, a_2 + b_2, \dots, a_n + b_n), \\ c\mathbf{a} &= (ca_1, ca_2, \dots, ca_n), \\ \mathbf{a} \cdot \mathbf{b} &= a_1b_1 + a_2b_2 + \dots + a_nb_n, \\ \|\mathbf{a}\| &= \sqrt{\mathbf{a} \cdot \mathbf{a}} = \sqrt{a_1^2 + a_2^2 + \dots + a_n^2} .\end{aligned}$$

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 $i = 1, 2, \dots, n$. 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 $\mathbf{c} = (0, 0, \dots, 0)$. 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 11.2).

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, $a^2 - 2tc + t^2b^2 \geq 0$, where $c = \mathbf{a} \cdot \mathbf{b}$. Completing the squares on the left side of this inequality,

$$\left(bt - \frac{c}{b}\right)^2 - \frac{c^2}{b^2} + a^2 \geq 0,$$

shows that the left side attains its absolute minimum when the expression in the parentheses vanishes, that is, at $t = c/b^2$. Since the inequality is valid for any t , it is satisfied for $t = c/b^2$, that is, $a^2 - c^2/b^2 \geq 0$ or $c^2 \leq a^2b^2$ 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 $\mathbf{a} = t\mathbf{b}$. \square

It follows from the Cauchy-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 two 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.

85.2. Real-Valued Functions of Several Variables.

DEFINITION 13.2. (Real-Valued Function of Several Variables).

Let D be a set of ordered n -tuples of real numbers (x_1, x_2, \dots, 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 $f(x_1, x_2, \dots, x_n)$. The set D is the domain of f , and its range is the set of values that f takes on it, that is, $\{f(x_1, x_2, \dots, x_n) \mid (x_1, x_2, \dots, x_n) \in D\}$.

This definition is illustrated in Figure 13.1. 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 $P_i \in D$, $i = 1, 2, \dots, N$, 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

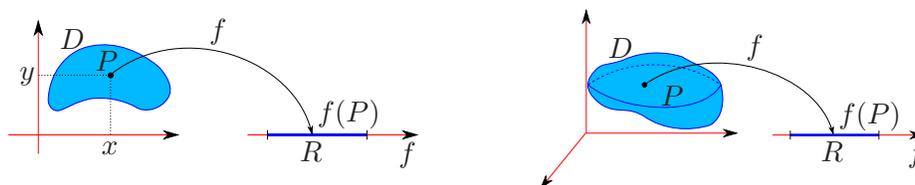


FIGURE 13.1. **Left:** A function f of two variables is a rule that assigns a number $f(P)$ to every point P of a planar region D . The set R of all numbers $f(P)$ is the range of f . The region D is the domain of f . **Right:** A function f of three variables is a rule that assigns a number $f(P)$ to every point P of a solid region D .

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 doublet (x, y) must be such that $1 - x^2 - y^2 > 0$ or $x^2 + y^2 < 1$. Hence, $D = \{(x, y) \mid x^2 + y^2 < 1\}$. 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 $0 < 1 - x^2 - y^2 < 1$. So the range of f is the set of values of the logarithm in the interval $(0, 1]$, which is $-\infty < f \leq 0$. \square

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, $D = \{(x, y, z) \mid z \geq x^2 + y^2\}$. 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 $0 \leq f(x, y, z) < \infty$. \square

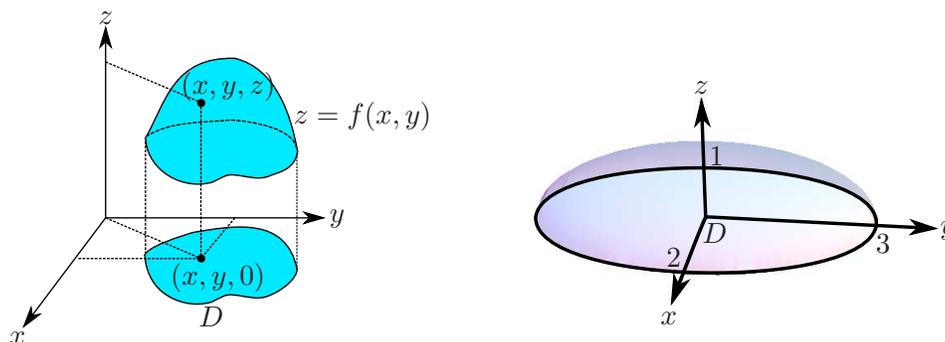


FIGURE 13.2. **Left:** The graph of a function of two variables is the surface defined by the equation $z = f(x, y)$. It is obtained from the domain D of f by moving each point $(x, y, 0)$ in D along the z axis to the point $(x, y, f(x, y))$. **Right:** The graph of the function studied in Example 13.3.

In general, the domain of a function of n variables is viewed as a subset of an n -dimensional Euclidean space. It is also convenient to adopt the vector notation of the argument:

$$f(x_1, x_2, \dots, x_n) = f(\mathbf{r}), \quad \mathbf{r} = (x_1, x_2, \dots, x_n).$$

For example, the domain of the function

$$f(\mathbf{r}) = (1 - x_1^2 - x_2^2 - \dots - x_n^2)^{1/2} = (1 - \|\mathbf{r}\|^2)^{1/2}$$

is the set of points in the n -dimensional Euclidean space whose distance from the origin (the zero vector) does not exceed 1, $D = \{\mathbf{r} \mid \|\mathbf{r}\| \leq 1\}$; 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.

85.3. 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 $\{(x, y) \mid y = f(x)\}$. 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 13.3. (Graph of a Function of Two Variables).

The graph of a function $f(x, y)$ with domain D is the point set in space

$$\{(x, y, z) \mid z = f(x, y), (x, y) \in D\}.$$

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 (see Figure 13.2, left panel). 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 $z \geq 0$ as depicted in the right panel of Figure 13.2. \square

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.

85.4. Level Sets. 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 13.4. (Level Sets).

The level sets of a function f are subsets of the domain of f on which the function has a fixed value; that is, they are determined by the equation $f(\mathbf{r}) = k$, where k is a number from the range of f .

For functions of two variables, the equation $f(x, y) = k$ generally defines a curve, but not necessarily so. For example, if $f(x, y) = x^2 + y^2$, then the equation $x^2 + y^2 = k$ defines concentric circles of radii \sqrt{k} for any $k > 0$. However, for $k = 0$, the level set consists of a single point $(x, y) = (0, 0)$. If f is a constant function on D , then it has just one level set that coincides with the entire domain D . A level set is called a *level curve* if the equation $f(x, y) = k$ defines a curve. Recall that

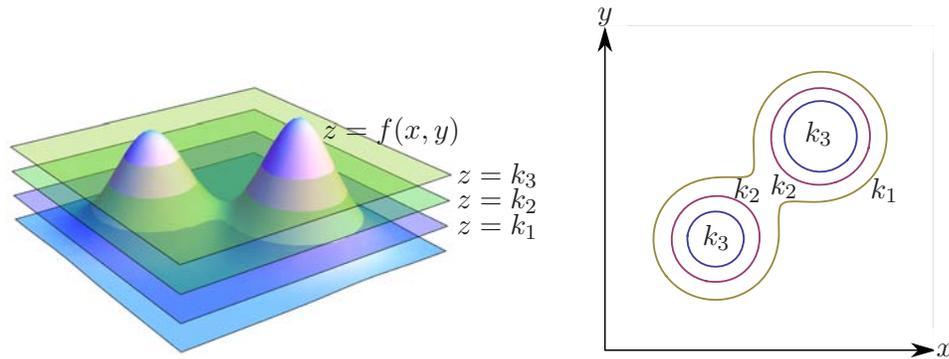


FIGURE 13.3. **Left:** Cross sections of the graph $z = f(x, y)$ by horizontal planes $z = k_i$, $i = 1, 2, 3$, are level curves $f(x, y) = k_i$ of the function f . **Right:** The contour map of the function f consists of level curves $f(x, y) = k_i$. The number k_i indicates the value of f along each level curve.

a curve in a plane can be described by parametric equations $x = x(t)$, $y = y(t)$, where $x(t)$ and $y(t)$ are continuous functions on an interval $a \leq t \leq b$. Therefore, the equation $f(x, y) = k$ defines a curve if there exist continuous functions $x(t)$ and $y(t)$ such that $f(x(t), y(t)) = k$ for all values of t from an interval. In general, a level set of a function may contain curves, isolated points, and even portions of the domain with nonzero area.

DEFINITION 13.5. (Contour Map).

A collection of level curves is called a contour map of the function f .

The concepts of level curves and a contour map of a function of two variables are illustrated in Figure 13.3. The contour map of the function in Example 13.3 consists of ellipses. Indeed, the range is the interval $[0, 1]$. For any $0 \leq k < 1$, 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 $b = 3\sqrt{1 - k^2}$. The level set for $k = 1$ consists of a single point, the origin.

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 , $i = 1, 2, \dots$, $f(x, y) = k_i$, 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. Contour maps are used in topography to indicate the steepness of mountains on maps.

EXAMPLE 13.4. Describe the level sets (a contour map) of the function $f(x, y) = (a^2 - (x^2 + y^2/4))^2$.

SOLUTION: The function depends on the combination $u^2 = x^2 + y^2/4$, $f(x, y) = (a^2 - u^2)^2$, and therefore the level sets $f(x, y) = k \geq 0$ are ellipses. The level set $k = 0$ is the ellipse $x^2/a^2 + y^2/(2a)^2 = 1$. The level sets $0 < k < a^4$ contain two ellipses because the equation $(a^2 - u^2)^2 = k$ has two solutions in this case: $u^2 = a^2 \pm \sqrt{k} > 0$. For $k = a^4$, the level set consists of the ellipse $u^2 = 2a^2$ and the point $(x, y) = (0, 0)$. The level sets for $k > a^4$ are ellipses $u^2 = \sqrt{k} - a^2$. So the contour map contains the ellipse $x^2/a^2 + y^2/(2a)^2 = 1$ along which the function attains its absolute minimum $f(x, y) = 0$. As the value of f increases, this ellipse splits into smaller and larger ellipses. At $f(x, y) = a^4$ (a local maximum of f attained at the origin), the smaller ellipses collapse to a point and disappear, while the larger ellipses keep expanding in size. The graph of f looks like a Mexican hat. \square

85.5. Level Surfaces. In contrast to the graph, the method of level curves uses only the domain of a function of two variables to study its behavior. Therefore, the concept of level sets can be useful to study the qualitative behavior of functions of three variables. In general, the equation $f(x, y, z) = k$ defines a surface in space, but not necessarily so as in the case of functions of two variables. The level sets of the function $f(x, y, z) = x^2 + y^2 + z^2$ are concentric spheres $x^2 + y^2 + z^2 = k$ for $k > 0$, but the level set for $k = 0$ contains just one point, the origin.

Intuitively, a surface in space can be obtained by a continuous deformation (without breaking) of a part of a plane, just like a curve is obtained by a continuous deformation of a line segment. Let S be a nonempty point set in space. A *neighborhood* of a point P of S is a collection of all points of S whose distance from P is less than a number $\delta > 0$. In particular, a neighborhood of a point in a plane is a disk centered at that point, and the boundary circle does not belong to the neighborhood. If every point of a subset D of a plane has a neighborhood that is contained in D , then the set D is called *open*. In other words, for every point P of an open region D in a plane, there is a disk of a sufficiently small radius that is centered at P and contained

in D . A point set S is a surface in space if every point of S has a neighborhood that can be obtained by a continuous deformation (or a deformation without breaking) of an open set in a plane and this deformation has a continuous inverse. This is analogous to the definition of a curve as a point set in space given in Section 79.3.

When the level sets of a function of three variables are surfaces, they are called *level surfaces*. 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 , $f(x, y, z) = k_i$, $k_{i+1} - k_i = \Delta k$, $i = 1, 2, \dots$, can be depicted in the domain of f . If P_0 is a point on S_i and P is the point on S_{i+1} that is the closest to P_0 , then the ratio $\Delta k/|P_0P|$ determines the maximal rate of change of f at P . So the closer the level surfaces S_i are to one another, the faster the function changes (see the right panel of Figure 13.4).

EXAMPLE 13.5. 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; that is, they are concave upward (see the right panel of Figure 13.4). For $k < 0$, the paraboloids are below the xy plane (i.e., they are concave downward). \square

85.6. Exercises.

(1) Find and sketch the domain of each of the following functions:

- (i) $f(x, y) = x/y$
- (ii) $f(x, y) = x/(x^2 + y^2)$
- (iii) $f(x, y) = x/(y^2 - 4x^2)$
- (iv) $f(x, y) = \ln(9 - x^2 - (y/2)^2)$
- (v) $f(x, y) = \sqrt{1 - (x/2)^2 - (y/3)^2}$
- (vi) $f(x, y) = \sqrt{4 - x^2 - y^2} + 2x \ln y$
- (vii) $f(x, y) = \sqrt{4 - x^2 - y^2} + x \ln y^2$
- (viii) $f(x, y) = \sqrt{4 - x^2 - y^2} + \ln(1 - x^2 - (y/2)^2)$
- (ix) $f(x, y, z) = x/(yz)$
- (x) $f(x, y, z) = x/(x - y^2 - z^2)$
- (xi) $f(x, y, z) = \ln(1 - z + x^2 + y^2)$
- (xii) $f(x, y, z) = \sqrt{x^2 - y^2 - z^2} + \ln(1 - x^2 - y^2 - z^2)$

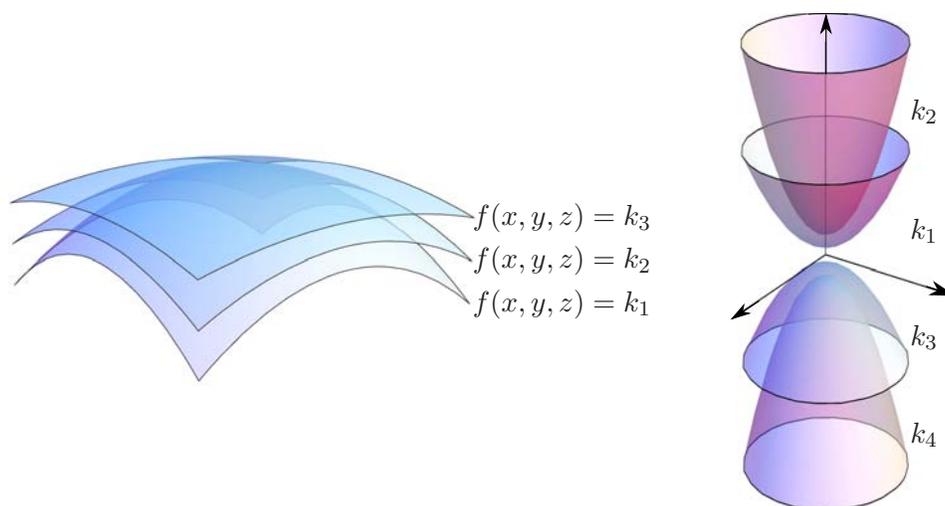


FIGURE 13.4. **Left:** A level surface of a function f of three variables is a surface in the domain of f on which the function attains a constant value k ; that is, it is defined by the equation $f(x, y, z) = k$. Here three level surfaces are depicted. **Right:** Level surfaces of the function studied in Example 13.5. Here $k_2 > k_1 > 0$ and $k_4 < k_3 < 0$. The level surface $f(x, y, z) = 0$ is the xy plane, $z = 0$.

(xiii) $f(t, \mathbf{r}) = (t^2 - \|\mathbf{r}\|^2)^{-1}$, $\mathbf{r} = (x_1, x_2, \dots, x_n)$

(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 sets 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) $f(x, y, z) = x^2 + y^2 - z^2$

(v) $f(x, y, z) = \ln(x^2 + y^2 - z^2)$

(vi) $f(x, y, z) = \ln(z^2 - x^2 - y^2)$

(4) Sketch the level sets of each of the following functions. Here $\min(a, b)$ and $\max(a, b)$ denote the smallest number and the largest number of a and b , respectively, and $\min(a, a) = \max(a, a) = a$.

(i) $f(x, y) = |x| + y$

- (ii) $f(x, y) = |x| + |y| - |x + y|$
 (iii) $f(x, y) = \min(x, y)$
 (iv) $f(x, y) = \max(|x|, |y|)$
 (v) $f(x, y) = \text{sign}(\sin(x)\sin(y))$; here $\text{sign}(a)$ is the sign function, it has the values 1 and -1 for positive and negative a , respectively
 (vi) $f(x, y, z) = (x + y)^2 + z^2$
 (vii) $f(x, y) = \tan^{-1}\left(\frac{2ay}{x^2 + y^2 - a^2}\right)$, $a > 0$
- (5) 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
- (6) Given a function f , sketch the graphs of $g(x, y)$ defined in exercise 5. Analyze carefully various cases for values of the constants, for example, $m > 0$, $m < 0$, $p > 1$, $0 < p < 1$, and $p = -1$.
- (i) $f(x, y) = x^2 + y^2$
 (ii) $f(x, y) = xy$
 (iii) $f(x, y) = (a^2 - x^2 - y^2)^2$
- (7) Find $f(u)$ if $f(x/y) = \sqrt{x^2 + y^2}/x$, $x > 0$.
 (8) Find $f(x, y)$ if $f(x + y, y/x) = x^2 - y^2$.
 (9) Let $z = \sqrt{y} + f(\sqrt{x} - 1)$. Find the functions z and f if $z = x$ when $y = 1$.
 (10) Graph the function $F(t) = f(\cos t, \sin t)$, where $f(x, y) = 1$ if $y \geq x$ and $f(x, y) = 0$ if $y < x$. Give a geometrical interpretation of the graph of F via the intersection of two surfaces.
 (11) Let $f(u)$ be a continuous function for all real u . Investigate the relation between the shape of the graph of f and the shape of the following surfaces:
- (i) $z = f(y - ax)$
 (ii) $z = f(\sqrt{x^2 + y^2})$
 (iii) $z = f(-\sqrt{x^2 + y^2})$
 (iv) $z = f(x/y)$

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 $x \rightarrow 0$. It is known (see Calculus I) that $\sin(x)/x \rightarrow 1$ as $x \rightarrow 0$. 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 $(x, y) = (0, 0)$. 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 an n -dimensional Euclidean space. Two points $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$ coincide if and only if the distance between them

$$\|\mathbf{x} - \mathbf{y}\| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

vanishes.

DEFINITION 13.6. *A point \mathbf{r} is said to approach a fixed point \mathbf{r}_0 in a Euclidean space if the distance $\|\mathbf{r} - \mathbf{r}_0\|$ tends to 0. The limit $\|\mathbf{r} - \mathbf{r}_0\| \rightarrow 0$ is also denoted by $\mathbf{r} \rightarrow \mathbf{r}_0$.*

In the above example, the limit $(x, y) \rightarrow (0, 0)$ means that $\sqrt{x^2 + y^2} \rightarrow 0$ or $x^2 + y^2 \rightarrow 0$. Therefore,

$$\frac{\sin(x^2 + y^2)}{x^2 + y^2} = \frac{\sin u}{u} \rightarrow 1 \quad \text{as} \quad x^2 + y^2 = u \rightarrow 0.$$

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 $(x, y) \rightarrow (0, 0)$ can be defined, whereas the limit of f as $(x, y) \rightarrow (-1, -1)$ 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 13.7. (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 $N_\delta(\mathbf{r}_0) = \{\mathbf{r} \mid 0 < \|\mathbf{r} - \mathbf{r}_0\| < \delta\}$ (with the center \mathbf{r}_0 removed) contains a point of D .

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 $\mathbf{r} \rightarrow \mathbf{r}_0$ and $\mathbf{r} \in D$ because, no matter how small δ is, one can always find a point $\mathbf{r} \in D$ that does not coincide with \mathbf{r}_0 and whose distance from \mathbf{r}_0 is less than δ . In other words, an intersection of any ball $N_\delta(\mathbf{r}_0)$ centered at a limit point of D with the set D , denoted as $N_\delta(\mathbf{r}_0) \cap D$, is always

nonempty. In the above example of D being the first quadrant, the limit $(x, y) \rightarrow (0, 0)$ is understood as $x^2 + y^2 \rightarrow 0$ while $(x, y) \neq (0, 0)$ and $x \geq 0, y \geq 0$. The intersection $N_\delta \cap D$ is the part of the disk $0 < x^2 + y^2 < \delta^2$ that lies in the first quadrant.

86.1. Limits of Functions of Several Variables.

DEFINITION 13.8. (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 $\mathbf{r} \rightarrow \mathbf{r}_0$ 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

$$\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = f_0.$$

The number $|f(\mathbf{r}) - f_0|$ determines the deviation of the value of f from the number f_0 . The existence of the limit means that no matter how small the number ε is, there is a neighborhood $N_\delta(\mathbf{r}_0) \cap D$, which contains all points of D whose distance from \mathbf{r}_0 is less than a number δ and in which the values of the function f deviate from the limit value f_0 no more than ε , that is, $f_0 - \varepsilon < f(\mathbf{r}) < f_0 + \varepsilon$ for all $\mathbf{r} \in N_\delta(\mathbf{r}_0) \cap D$. The condition $0 < \|\mathbf{r} - \mathbf{r}_0\|$ ensures that \mathbf{r} does not coincide with \mathbf{r}_0 . Note that f is not even defined at \mathbf{r}_0 if \mathbf{r}_0 is not in D . In the case of a function of two variables, this definition is illustrated in the right panel of Figure 13.5.

EXAMPLE 13.6. Show that

$$\lim_{(x,y,z) \rightarrow (0,0,0)} (x^2y + yz^2 - 6z^3) = 0.$$

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 $|z| \leq R$. Let us find an upper bound on the deviation of values of the function from the limit value $f_0 = 0$ in terms of R :

$$|f(\mathbf{r}) - f_0| = |x^2y + yz^2 - 6z^3| \leq |x^2y| + |yz^2| + 6|z^3| \leq 8R^3,$$

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, for all points $\mathbf{r} \neq \mathbf{0}$ in the domain of the function for which $R < \delta = \sqrt[3]{\varepsilon}/2$, the function differs from 0 no more than ε . For example, put $\varepsilon = 10^{-6}$. Then, in the interior of a ball of radius

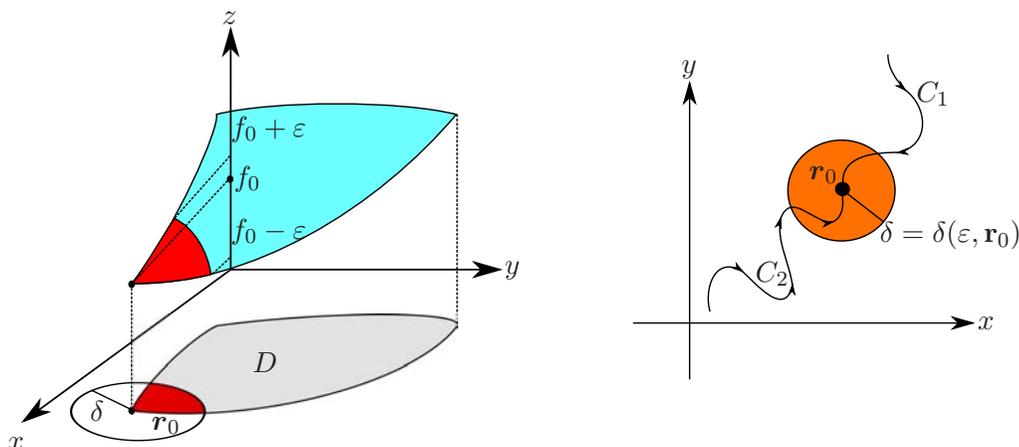


FIGURE 13.5. **Left:** An illustration of Definition 13.8 in the case of a function f of two variables. Given a positive number ε , consider two horizontal planes $z = f_0 - \varepsilon$ and $z = f_0 + \varepsilon$. Then one can always find a corresponding number $\delta > 0$ and the disk N_δ centered at \mathbf{r}_0 such that the portion of the graph $z = f(\mathbf{r})$ above the intersection $N_\delta \cap D$ lies between the planes: $f_0 - \varepsilon < f(\mathbf{r}) < f_0 + \varepsilon$. The radius $\delta > 0$ of N_δ depends on the choice of ε and, generally, on the limit point \mathbf{r}_0 . **Right:** The independence of the limit of a path along which the limit point \mathbf{r}_0 is approached. For every path leading to \mathbf{r}_0 , there is a part of it that lies in N_δ . The values of f along this part of the path deviate from f_0 no more than any preassigned number $\varepsilon > 0$.

$\delta = 0.005$, the values of the function can deviate from $f_0 = 0$ no more than 10^{-6} . \square

The radius δ of a neighborhood in which a function f deviates no more than ε from the value of the limit depends on ε and, in general, on the limit point \mathbf{r}_0 .

EXAMPLE 13.7. Let $f(x, y) = xy$. Show that

$$\lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y) = x_0 y_0$$

for any point (x_0, y_0) .

SOLUTION: The distance between $\mathbf{r} = (x, y)$ and $\mathbf{r}_0 = (x_0, y_0)$ is $R = \sqrt{(x - x_0)^2 + (y - y_0)^2}$. Therefore, $|x - x_0| \leq R$ and $|y - y_0| \leq R$. Consider the identity

$$xy - x_0 y_0 = (x - x_0)(y - y_0) + x_0(y - y_0) + (x - x_0)y_0.$$

Put $a = (|x_0| + |y_0|)/2$. Then the deviation of f from the limit value $f_0 = x_0y_0$ is bounded as

$$\begin{aligned} |f(x, y) - f_0| &\leq |x - x_0||y - y_0| + |x_0||y - y_0| + |x - x_0||y_0| \\ &\leq R^2 + (|x_0| + |y_0|)R = R^2 + 2aR \\ &= (R + a)^2 - a^2. \end{aligned}$$

Now fix $\varepsilon > 0$ and assume that R is such that $(R + a)^2 - a^2 < \varepsilon$ or $0 < R < \sqrt{\varepsilon + a^2} - a$. Therefore, the function f deviates from f_0 no more than ε in a neighborhood of \mathbf{r}_0 of radius $\delta = \sqrt{\varepsilon + a^2} - a$, which depends on ε and the limit point \mathbf{r}_0 . \square

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 path 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 $f_0 - \varepsilon < f(\mathbf{r}) < f_0 + \varepsilon$ for all points \mathbf{r} in the ball and hence for all points of the portion of the path in the ball (see Figure 13.5, right panel). 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 $x \rightarrow x_0$, then the right $x \rightarrow x_0^+$ and left $x \rightarrow x_0^-$ limits exist and are equal (and vice versa).

86.2. 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 $\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = f_0$ and $\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} g(\mathbf{r}) = g_0$. Then the following properties hold:

$$\begin{aligned} \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} (cf(\mathbf{r})) &= c \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = cf_0, \\ \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} (g(\mathbf{r}) + f(\mathbf{r})) &= \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} g(\mathbf{r}) + \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = g_0 + f_0, \\ \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} (g(\mathbf{r})f(\mathbf{r})) &= \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} g(\mathbf{r}) \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = g_0f_0, \\ \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} \frac{g(\mathbf{r})}{f(\mathbf{r})} &= \frac{\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} g(\mathbf{r})}{\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r})} = \frac{g_0}{f_0} \quad \text{if } f_0 \neq 0. \end{aligned}$$

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 13.6 employs a rather general strategy to verify whether a particular number f_0 is the limit of $f(\mathbf{r})$ as $\mathbf{r} \rightarrow \mathbf{r}_0$.

THEOREM 13.3. (Squeeze Principle).

Let the functions of several variables g , f , and h have a common domain D and let $g(\mathbf{r}) \leq f(\mathbf{r}) \leq h(\mathbf{r})$ for any $\mathbf{r} \in D$. If the limits of $g(\mathbf{r})$ and $h(\mathbf{r})$ as $\mathbf{r} \rightarrow \mathbf{r}_0$ exist and equal a number f_0 , then the limit of $f(\mathbf{r})$ as $\mathbf{r} \rightarrow \mathbf{r}_0$ exists and equals f_0 , that is,

$$g(\mathbf{r}) \leq f(\mathbf{r}) \leq h(\mathbf{r}) \text{ and } \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} g(\mathbf{r}) = \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} h(\mathbf{r}) = f_0 \Rightarrow \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = f_0.$$

PROOF. From the hypothesis of the theorem, it follows that $0 \leq f(\mathbf{r}) - g(\mathbf{r}) \leq h(\mathbf{r}) - g(\mathbf{r})$. Put $F(\mathbf{r}) = f(\mathbf{r}) - g(\mathbf{r})$ and $H(\mathbf{r}) = h(\mathbf{r}) - g(\mathbf{r})$. Then $0 \leq F(\mathbf{r}) \leq H(\mathbf{r})$ implies $|F(\mathbf{r})| \leq |H(\mathbf{r})|$ (the positivity of F is essential for this conclusion). By the hypothesis of the theorem and the basic properties of the limit, $H(\mathbf{r}) = h(\mathbf{r}) - g(\mathbf{r}) \rightarrow f_0 - f_0 = 0$ as $\mathbf{r} \rightarrow \mathbf{r}_0$. Hence, for any $\varepsilon > 0$, there is a corresponding number δ such that $0 \leq |F(\mathbf{r})| \leq |H(\mathbf{r})| < \varepsilon$ whenever $0 < \|\mathbf{r} - \mathbf{r}_0\| < \delta$. By Definition 13.8, this means that $\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} F(\mathbf{r}) = 0$. By the basic properties of the limit, it is then concluded that $f(\mathbf{r}) = F(\mathbf{r}) + g(\mathbf{r}) \rightarrow 0 + f_0 = f_0$ as $\mathbf{r} \rightarrow \mathbf{r}_0$. \square

A particular case of the squeeze principle is also useful.

COROLLARY 13.1. (Simplified Squeeze Principle).

If there exists a function h of one variable such that

$$|f(\mathbf{r}) - f_0| \leq h(R) \rightarrow 0 \quad \text{as} \quad \|\mathbf{r} - \mathbf{r}_0\| = R \rightarrow 0^+,$$

then $\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = f_0$.

The condition $|f(\mathbf{r}) - f_0| \leq h(R)$ is equivalent to $f_0 - h(R) \leq f(\mathbf{r}) \leq f_0 + h(R)$, which is a particular case of the hypothesis in the squeeze principle. In Example 13.6, $h(R) = 8R^3$. In general, the condition $h(R) \rightarrow 0$ as $R \rightarrow 0^+$ 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 $\|\mathbf{r} - \mathbf{r}_0\| = R < \delta(\varepsilon)$.

EXAMPLE 13.8. *Show that*

$$\lim_{(x,y) \rightarrow (0,0)} f(x,y) = 0, \quad \text{where} \quad f(x,y) = \frac{x^3y - 3x^2y^2}{x^2 + y^2 + x^4}.$$

SOLUTION: Let $R = \sqrt{x^2 + y^2}$ (the distance from the limit point $(0, 0)$). Then $|x| \leq R$ and $|y| \leq R$. Therefore,

$$\frac{|x^3y - 3x^2y^2|}{x^2 + y^2 + x^4} \leq \frac{|x|^3|y| + 3x^2y^2}{x^2 + y^2 + x^4} \leq \frac{4R^4}{R^2 + x^4} = \frac{4R^2}{1 + (x^4/R^2)} \leq 4R^2.$$

It follows from this inequality that $-4(x^2 + y^2) \leq f(x, y) \leq 4(x^2 + y^2)$, and, by the squeeze principle, $f(x, y)$ must tend to 0 because $\pm 4(x^2 + y^2) = \pm 4R^2 \rightarrow 0$ as $R \rightarrow 0$. In Definition 13.8, given $\varepsilon > 0$, the corresponding number δ is $\delta = \sqrt{\varepsilon}/2$. \square

86.3. Continuity of Functions of Several Variables. Suppose that $\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = f_0$. If the limit point \mathbf{r}_0 lies in the domain of the function f , then the function has a value $f(\mathbf{r}_0)$, which may or may not coincide with the limit value f_0 . In fact, the limit value f_0 does not generally give any information about the possible value of the function at the limit point. For example, if $f(\mathbf{r}) = 1$ everywhere except one point \mathbf{r}_0 at which $f(\mathbf{r}_0) = c$, then, in every neighborhood $0 < \|\mathbf{r} - \mathbf{r}_0\| < \delta$, $f(\mathbf{r}) = 1$ and hence the limit of f as $\mathbf{r} \rightarrow \mathbf{r}_0$ exists and equals $f_0 = 1$. When $c \neq 1$, the limit value does not coincide with the value of the function at the limit point. The values of f suffer a jump *discontinuity* when \mathbf{r} reaches \mathbf{r}_0 , and one says that f is *discontinuous* at \mathbf{r}_0 . A discontinuity also occurs when the limit of f as $\mathbf{r} \rightarrow \mathbf{r}_0$ does not exist while f has a value at the limit point.

DEFINITION 13.9. (Continuity).

A function f of several variables with domain D is said to be continuous at a point $\mathbf{r}_0 \in D$ if

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

The function f is said to be continuous on D if it is continuous at every point of D .

EXAMPLE 13.9. Let $f(x, y) = 1$ if $y \geq x$ and let $f(x, y) = 0$ if $y < x$. Determine the region on which f is continuous.

SOLUTION: 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, $\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = f(\mathbf{r}_0) = 1$. The same line of reasoning 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. \square

The following theorem is a simple consequence of the basic properties of the limit.

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 $g(\mathbf{r}) \neq 0$.

The use of the definition to establish the continuity of a function defined by an algebraic rule is not always convenient. The following two theorems are helpful when studying the continuity of a given function. For an ordered n -tuple $\mathbf{r} = (x_1, x_2, \dots, x_n)$, the function $x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n}$, where k_1, k_2, \dots, 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*. For example, the function $f(x, y, z) = 1 + y - 2xz + z^4$ is a polynomial of three variables. 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 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, \dots, x_n) is changed to $(x_1 + a_1, x_2 + a_2, \dots, x_n + a_n)$, where a_1, a_2, \dots, 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. A monomial function $f(\mathbf{r}) = x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n}$ of degree $N = k_1 + k_2 + \cdots + k_n > 0$ vanishes at the origin $\mathbf{r}_0 = \mathbf{0}$, $f(\mathbf{r}_0) = 0$. Put $R = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$. Then $|x_i| \leq R$ for any element of the n -tuple. Hence,

$$|f(\mathbf{r}) - f(\mathbf{r}_0)| = |x_1|^{k_1} |x_2|^{k_2} \cdots |x_n|^{k_n} \leq R^{k_1 + k_2 + \cdots + k_n} = R^N \rightarrow 0$$

as $R \rightarrow 0$. By the squeeze principle, $f(\mathbf{r}) \rightarrow 0 = f(\mathbf{r}_0)$. The rational function $f(\mathbf{r})/g(\mathbf{r})$ is continuous as the ratio of two continuous functions if $g(\mathbf{r}) \neq 0$. \square

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.

EXAMPLE 13.10. Find the limit

$$\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{e^{xz} \cos(xy + z^2)}{x + yz + 3xz^4 + (xyz - 2)^2}.$$

SOLUTION: The function is a ratio. The denominator is a polynomial and hence continuous. Its limit value is $(-2)^2 = 4 \neq 0$. The function e^{xz} is a composition of the exponential e^u and the polynomial $u = xz$. So it is continuous. Its value is 1 at the limit point. Similarly, $\cos(xy + z^2)$ is continuous as a composition of $\cos u$ and the polynomial $u = xy + z^2$. Its value is 1 at the limit point. The ratio of continuous functions is continuous and the limit is $1/4$. \square

86.4. Exercises.

(1) Use the definition of the limit to verify each of the following limits (i.e., given $\varepsilon > 0$, find the corresponding $\delta(\varepsilon)$):

(i) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{x^3 - 4y^2x + 5y^3}{x^2 + y^2} = 0$

(ii) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{x^3 - 4y^2x + 5y^3}{3x^2 + 4y^2} = 0$

(iii) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{x^3 - 4y^4 + 5y^3x^2}{3x^2 + 4y^2} = 0$

$$(iv) \lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{x^3 - 4y^2x + 5y^3}{3x^2 + 4y^2 + y^4} = 0$$

$$(v) \lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{3x^3 + 4y^4 - 5z^5}{x^2 + y^2 + z^2} = 0$$

(2) Use the squeeze principle to prove the following limits and find a neighborhood of the limit point in which the deviation of the function from the limit value does not exceed a small given number ε :

$$(i) \lim_{\mathbf{r} \rightarrow \mathbf{0}} y \sin(x/\sqrt{y}) = 0$$

$$(ii) \lim_{\mathbf{r} \rightarrow \mathbf{0}} [1 - \cos(y/x)]x = 0$$

$$(iii) \lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{\cos(xy) \sin(4x\sqrt{y})}{\sqrt{xy}} = 0$$

Hint: $|\sin u| \leq |u|$.

(3) Suppose that $\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = 2$ and \mathbf{r}_0 is in the domain of f . If nothing else is known about the function, what can be said about the value $f(\mathbf{r}_0)$? If, in addition, f is known to be continuous at \mathbf{r}_0 , what can be said about the value $f(\mathbf{r}_0)$?

(4) Find the points of discontinuity of each of the following functions:

$$(i) f(x, y) = yx/(x^2 + y^2) \text{ if } (x, y) \neq (0, 0) \text{ and } f(0, 0) = 1$$

$$(ii) f(x, y, z) = yxz/(x^2 + y^2 + z^2) \text{ if } (x, y, z) \neq (0, 0, 0) \text{ and } f(0, 0, 0) = 0$$

$$(iii) f(x, y) = \sin(\sqrt{xy})$$

$$(iv) f(x, y) = \cos(\sqrt{xyz})/(x^2y^2 + 1)$$

$$(v) f(x, y) = (x^2 + y^2) \ln(x^2 + y^2) \text{ if } (x, y) \neq (0, 0) \text{ and } f(0, 0) = 0$$

$$(vi) f(x, y) = 1 \text{ if either } x \text{ or } y \text{ is rational and } f(x, y) = 0 \text{ elsewhere}$$

$$(vii) f(x, y) = (x^2 - y^2)/(x - y) \text{ if } x \neq y \text{ and } f(x, x) = 2x$$

$$(viii) f(x, y) = (x^2 - y^2)/(x - y) \text{ if } x \neq y \text{ and } f(x, x) = x$$

$$(ix) f(x, y, z) = 1/[\sin(x) \sin(z - y)]$$

$$(x) f(x, y) = \sin\left(\frac{1}{xy}\right)$$

(5) Each of the following functions has the value at the origin $f(0, 0) = c$. Determine whether there is a particular value of c at which the function is continuous at the origin if, for $(x, y) \neq (0, 0)$,

$$(i) f(x, y) = \sin(1/(x^2 + y^2))$$

$$(ii) f(x, y) = (x^2 + y^2)^\nu \sin(1/(x^2 + y^2)), \nu > 0$$

(iii) $f(x, y) = x^n y^m \sin(1/(x^2 + y^2))$, $n \geq 0$, $m \geq 0$, and $n + m > 0$

(6) Use the properties of continuous functions to find the following limits

$$(i) \lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{(1 + x + yz^2)^{1/3}}{2 + 3x - 4y + 5z^2}$$

$$(ii) \lim_{\mathbf{r} \rightarrow \mathbf{0}} \sin(x\sqrt{y})$$

$$(iii) \lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{\sin(x\sqrt{y})}{\cos(x^2y)}$$

$$(iv) \lim_{\mathbf{r} \rightarrow \mathbf{0}} [e^{xyz} - 2 \cos(yz) + 3 \sin(xy)]$$

$$(v) \lim_{\mathbf{r} \rightarrow \mathbf{0}} \ln(1 + x^2 + y^2z^2)$$

87. A General Strategy for Studying Limits

The definition of the limit gives only the criterion for whether a number f_0 is the limit of $f(\mathbf{r})$ as $\mathbf{r} \rightarrow \mathbf{r}_0$. 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 $\mathbf{r}_0 = (0, 0, \dots, 0)$. 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,

$$\lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y) = \lim_{(x,y) \rightarrow (0,0)} f(x + x_0, y + y_0).$$

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:

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

For example,

$$\lim_{(x,y) \rightarrow (1,2)} \frac{xy}{x^3 - y^2} = -\frac{2}{3}$$

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 continuous at t_0 . Suppose that the function f is the composition $f(\mathbf{r}) = g(h(\mathbf{r}))$ so that \mathbf{r}_0 is a limit point of the domain of f and $h(\mathbf{r}) \rightarrow t_0$ as $\mathbf{r} \rightarrow \mathbf{r}_0$. Then

$$\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = \lim_{t \rightarrow t_0} g(t) = g(t_0).$$

The proof is omitted as it is similar to the proof of the composition rule for limits of single-variable functions given in Calculus I. The significance of this theorem is that, under the hypotheses 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 limits.*

EXAMPLE 13.11. Find

$$\lim_{(x,y) \rightarrow (0,0)} \frac{\cos(xy) - 1}{x^2y^2}.$$

SOLUTION: The function in question is $g(t) = (\cos t - 1)/t^2$ for $t \neq 0$, where the argument t is replaced by the function $h(x, y) = xy$. The function h is a polynomial and hence continuous. In particular, $h(x, y) \rightarrow h(0, 0) = 0$ as $(x, y) \rightarrow (0, 0)$. The function $g(t)$ is continuous for all $t \neq 0$ and its value at $t = 0$ is not defined. Using l'Hospital's rule twice,

$$\lim_{t \rightarrow 0} \frac{\cos t - 1}{t^2} = \lim_{t \rightarrow 0} \frac{-\sin t}{2t} = \lim_{t \rightarrow 0} \frac{-\cos t}{2} = -\frac{1}{2}.$$

So, by setting $g(0) = -1/2$, the function $g(t)$ becomes continuous at $t = 0$, and the hypotheses of the composition rule are fulfilled. Therefore, the two dimensional limit in question exists and equals $-1/2$. \square

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 $x \rightarrow x_0$ exists and equals f_0 if and only if the corresponding right and left limits of $f(x)$ exist and equal f_0 :

$$\lim_{x \rightarrow x_0^+} f(x) = \lim_{x \rightarrow x_0^-} f(x) = f_0 \iff \lim_{x \rightarrow x_0} f(x) = f_0.$$

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 path along which the limit point may be approached.*

DEFINITION 13.10. (Parametric Curve in a Euclidean Space).

A parametric curve in a Euclidean space is a set of points $\mathbf{r}(t) = (x_1(t), x_2(t), \dots, x_n(t))$, where $x_i(t)$, $i = 1, 2, \dots, n$, are continuous functions of a variable $t \in [a, b]$.

This is a natural generalization of the concept of a parametric curve in a plane or space as a vector function defined by the parametric equations $x_i = x_i(t)$, $i = 1, 2, \dots, n$.

DEFINITION 13.11. (Limit Along a Curve).

Let \mathbf{r}_0 be a limit point of the domain D of a function f . Let $\mathbf{r}(t) = (x_1(t), x_2(t), \dots, x_n(t))$, $t \geq t_0$, be a parametric curve C in D such that $\mathbf{r}(t) \rightarrow \mathbf{r}_0$ as $t \rightarrow t_0^+$. Let $F(t) = f(\mathbf{r}(t))$, $t > t_0$, be the values of f on the curve C . The limit

$$\lim_{t \rightarrow t_0^+} F(t) = \lim_{t \rightarrow t_0^+} f(x_1(t), x_2(t), \dots, x_n(t))$$

is called the limit of f along the curve C if it exists.

Suppose that the limit of $f(\mathbf{r})$ as $\mathbf{r} \rightarrow \mathbf{r}_0$ exists and equals f_0 . Let C be a curve such that $\mathbf{r}(t) \rightarrow \mathbf{r}_0$ as $t \rightarrow t_0^+$. Fix $\varepsilon > 0$. By the existence of the limit, there is a neighborhood $N_\delta(\mathbf{r}_0) = \{\mathbf{r} \mid \mathbf{r} \in D, 0 < \|\mathbf{r} - \mathbf{r}_0\| < \delta\}$ in which the values of f deviate from f_0 no more than ε , $|f(\mathbf{r}) - f_0| < \varepsilon$. Since the curve C 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 $t \in (t_0, t_0 + \delta')$, which is merely the definition of the limit $\mathbf{r}(t) \rightarrow \mathbf{r}_0$ as $t \rightarrow t_0^+$. 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 $F(t) \rightarrow f_0$ as $t \rightarrow t_0$ 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 $\mathbf{r} \rightarrow \mathbf{r}_0$, then the limit of f along any curve leading to \mathbf{r}_0 from within the domain of f exists, and its value is independent of the curve.

An immediate consequence of this theorem is a useful criterion for the nonexistence of a multivariable limit.

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 $\mathbf{r}(t) \rightarrow \mathbf{r}_0$ as $t \rightarrow t_0^+$ and the limit $\lim_{t \rightarrow t_0^+} f(\mathbf{r}(t))$ does not exist, then the multivariable limit $\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r})$ 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 $\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r})$ does not exist.

Repeated Limits. Let $(x, y) \neq (0, 0)$. Consider a path C_1 that consists of two straight line segments $(x, y) \rightarrow (x, 0) \rightarrow (0, 0)$ and a path C_2 that consists of two straight line segments $(x, y) \rightarrow (0, y) \rightarrow (0, 0)$. Both paths connect (x, y) with the origin. The limits along C_1 and C_2 ,

$$\lim_{y \rightarrow 0} \left(\lim_{x \rightarrow 0} f(x, y) \right) \quad \text{and} \quad \lim_{x \rightarrow 0} \left(\lim_{y \rightarrow 0} f(x, y) \right),$$

are called the *repeated* limits. If C_1 and C_2 are *within* the domain of f , then Theorem 13.8 and Corollary 13.2 establish the relations between the repeated limits and the two-variable limit $\lim_{(x,y) \rightarrow (0,0)} f(x, y)$. In particular, suppose that $f(x, y) \rightarrow f(0, y)$ as $x \rightarrow 0$ and $f(x, y) \rightarrow f(x, 0)$ as $y \rightarrow 0$ (the function is continuous with respect to x if y is fixed and it is also continuous with respect to y if x is fixed). Then the repeated limits become

$$\lim_{y \rightarrow 0} f(0, y) \quad \text{and} \quad \lim_{x \rightarrow 0} f(x, 0).$$

If at least one of them does not exist or they exist but are not equal, then, by Corollary 13.2, the two-variable limit does not exist. If they exist and are equal, then the two-variable limit *may or may not* exist. A further investigation is needed.

In general, the segment $(x, 0) \rightarrow (0, 0)$ or $(0, y) \rightarrow (0, 0)$ or both may not be in the domain of f , while the repeated limits still make sense (e.g., the function f is defined only for strictly positive x and y so that the half-lines $x = 0, y > 0$ and $y = 0, x > 0$ are limit points of the domain). In this case, the hypotheses of Corollary 13.2 are not fulfilled, and, in particular, *the nonexistence of the repeated limits does not imply the nonexistence of the two-variable limit*. An example is provided in exercise 1, part (iii).

Limits Along Straight Lines. Let the limit point be the origin $\mathbf{r}_0 = (0, 0, \dots, 0)$. The simplest curve leading to \mathbf{r}_0 is a straight line $x_i = v_i t$, where $t \rightarrow 0^+$ for some numbers $v_i, i = 1, 2, \dots, n$, that do not vanish

simultaneously. The limit of a function of several variables f along a straight line, $\lim_{t \rightarrow 0^+} f(v_1 t, v_2 t, \dots, v_n t)$, 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 13.12. *Investigate the two-variable limit*

$$\lim_{(x,y) \rightarrow (0,0)} \frac{xy^3}{x^4 + 2y^4}.$$

SOLUTION: Consider the limits along straight lines $x = t$, $y = at$ (or $y = ax$, where a is the slope) as $t \rightarrow 0^+$:

$$\lim_{t \rightarrow 0^+} f(t, at) = \lim_{t \rightarrow 0^+} \frac{a^3 t^4}{t^4(1 + 2a^4)} = \frac{a^3}{1 + 2a^4}.$$

So the limit along a straight line depends on the slope of the line. Therefore, the two-variable limit does not exist. \square

EXAMPLE 13.13. *Investigate the limit*

$$\lim_{(x,y) \rightarrow (0,0)} \frac{\sin(\sqrt{xy})}{x + y}.$$

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 $t \rightarrow 0$. 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

$$\lim_{t \rightarrow 0^+} f(t, at) = \lim_{t \rightarrow 0^+} \frac{\sin(t\sqrt{a})}{t(1 + a)} = \lim_{t \rightarrow 0^+} \frac{\sqrt{a} \cos(t\sqrt{a})}{1 + a} = \frac{\sqrt{a}}{1 + a},$$

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. \square

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 13.14. *Investigate the limit*

$$\lim_{(x,y) \rightarrow (0,0)} \frac{y^3}{x}.$$

SOLUTION: The domain of the function is the whole plane with the y axis removed ($x \neq 0$). The limit along a straight line

$$\lim_{t \rightarrow 0^+} f(t, at) = \lim_{t \rightarrow 0^+} \frac{a^3 t^3}{t} = a^3 \lim_{t \rightarrow 0^+} t^2 = 0$$

vanishes for any slope; that is, it is independent of the choice of the line. However, the two-variable limit does not exist! Consider the power curve $x = t$, $y = at^{1/3}$ approaching the origin as $t \rightarrow 0^+$. The limit along this curve can attain any value by varying the parameter a :

$$\lim_{t \rightarrow 0^+} f(t, at^{1/3}) = \lim_{t \rightarrow 0^+} \frac{a^3 t}{t} = a^3.$$

Thus, the multivariable limit does not exist. \square

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 $t \rightarrow 0^+$. For instance, in the example considered, put $x = t$ and $y = at^n$. Then $f(t, at^n) = (a^3 t^{3n})/t$. 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 curve 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-after “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 13.15. Find the limit or prove that it does not exist:

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

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 $t \rightarrow 0^+$. The limit along straight lines

$$\begin{aligned} \lim_{t \rightarrow 0^+} f(t, at) &= \lim_{t \rightarrow 0^+} \frac{\sin(a^2 t^3)}{t^2} = \lim_{u \rightarrow 0^+} \frac{\sin(a^2 u^{3/2})}{u} \\ &= \lim_{u \rightarrow 0^+} \frac{(3/2)a^2 u^{1/2} \cos(a^2 u^{3/2})}{1} = 0 \end{aligned}$$

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 $|f(x, y) - f_0| = |f(x, y)| \leq h(R) \rightarrow 0$ as $R = \sqrt{x^2 + y^2} \rightarrow 0$. A key technical trick here is the inequality $|\sin u| \leq |u|$, which holds for any real u . One has

$$|f(x, y) - 0| = \frac{|\sin(xy^2)|}{x^2 + y^2} \leq \frac{|xy^2|}{x^2 + y^2} \leq \frac{R^3}{R^2} = R \rightarrow 0,$$

where the inequalities $|x| \leq R$ and $|y| \leq R$ have been used. Thus, the two-variable limit exists and equals 0. \square

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 $\theta \in [0, 2\pi]$:

$$|f(x, y) - f_0| = |f(x_0 + R \cos \theta, y_0 + R \sin \theta) - f_0| \leq h(R) \rightarrow 0$$

as $R \rightarrow 0^+$. 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 13.15, Step 3 gives $f_0 = 0$ if only the limits along straight lines have been studied. Then

$$|f(x, y) - f_0| = \frac{|y|^3}{|x|} = \frac{|R \sin^3 \theta|}{|R \cos \theta|} = R^2 \sin^2(\theta) |\tan \theta|.$$

Despite that the deviation of f from 0 is proportional to $R^2 \rightarrow 0$ as $R \rightarrow 0^+$, it cannot be made as small as desired uniformly for all θ 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 larger than any 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 $0 < \|\mathbf{r} - \mathbf{r}_0\| < \delta$, and hence the limit does not exist.

Remark. For multivariable limits with $n > 2$, a similar approach exists. If, for simplicity, $\mathbf{r}_0 = (0, 0, \dots, 0)$. Then put $x_i = Ru_i$, where the variables u_i satisfy the condition $u_1^2 + u_2^2 + \dots + u_n^2 = 1$. 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 unit vectors $\hat{\mathbf{e}}_i$ parallel to the coordinate axes, $u_i = \mathbf{r} \cdot \hat{\mathbf{e}}_i / \|\mathbf{r}\|$. Then one has to investigate whether there is $h(R)$ such that

$$|f(Ru_1, Ru_2, \dots, Ru_n) - f_0| \leq h(R) \rightarrow 0, \quad R \rightarrow 0^+.$$

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.

87.5. Infinite Limits and Limits at Infinity. Suppose that the limit of a multivariable function f does not exist as $\mathbf{r} \rightarrow \mathbf{r}_0$. There are two particular cases, which are of interest, when f tends to either positive or negative infinity.

DEFINITION 13.12. (Infinite Limits).

The limit of $f(\mathbf{r})$ as $\mathbf{r} \rightarrow \mathbf{r}_0$ is said to be the positive infinity if, for any number $M > 0$, there exists a number $\delta > 0$ such that $f(\mathbf{r}) > M$ whenever $0 < \|\mathbf{r} - \mathbf{r}_0\| < \delta$. Similarly, the limit is said to be the negative infinity if, for any number $M < 0$, there exists a number $\delta > 0$ such that $f(\mathbf{r}) < M$ whenever $0 < \|\mathbf{r} - \mathbf{r}_0\| < \delta$. In these cases, one writes,

respectively,

$$\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = \infty \quad \text{and} \quad \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r}) = -\infty.$$

For example,

$$\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{1}{x^2 + y^2} = \infty.$$

Indeed, put $R = \sqrt{x^2 + y^2}$. Then, for any $M > 0$, the inequality $f(\mathbf{r}) > M$ can be written in the form $R < 1/\sqrt{M}$. Therefore, the values of f in the disk $0 < \|\mathbf{r}\| < \delta = 1/\sqrt{M}$ are larger than any preassigned positive number M .

Naturally, if the limit is infinite, the function f approaches the infinite value along any curve that leads to the limit point. For example, the limit of $f(x, y) = y/(x^2 + y^2)$ as $(x, y) \rightarrow (0, 0)$ does not exist because, along straight lines $(x, y) = (t, at)$ approaching the origin when $t \rightarrow 0^+$, the function $f(t, at) = c/t$, where $c = a/(1 + a^2)$, tends to $+\infty$ if $a > 0$, to $-\infty$ if $a < 0$, and to 0 if $a = 0$. If, however, the domain of f is restricted to the half-plane $y > 0$, then the limit exists and equals ∞ . Indeed, for all x and $y > 0$, $f(x, y) \geq y/y^2 = 1/y \rightarrow \infty$ as $y \rightarrow 0^+$, and the conclusion follows from the squeeze principle.

For functions of one variable x , one can define the limits at infinity (i.e., when $x \rightarrow +\infty$ or $x \rightarrow -\infty$). Both the limits have a common property that the distance $|x|$ of the “infinite points” $\pm\infty$ from the origin $x = 0$ is infinite. Similarly, in a Euclidean space, the limit at infinity is defined in the sense that $\|\mathbf{r}\| \rightarrow \infty$. If D is an unbounded region, then a neighborhood of the infinite point in D consists of all points of D whose distance from the origin exceeds a number δ , $\|\mathbf{r}\| > \delta$. A smaller neighborhood is obtained by increasing δ .

DEFINITION 13.13. (Limit at Infinity).

Let f be a function on an unbounded region D . A number f_0 is the limit of a function f at infinity,

$$\lim_{\mathbf{r} \rightarrow \infty} f(\mathbf{r}) = f_0$$

if, for any number $\varepsilon > 0$, there exists $\delta > 0$ such that $|f(\mathbf{r}) - f_0| < \varepsilon$ whenever $\|\mathbf{r}\| > \delta$ in D .

Infinite limits at infinity can be defined similarly. The squeeze principle has a natural extension to the infinite limits and limits at infinity. For example, if $g(\mathbf{r}) \leq f(\mathbf{r})$ and $g(\mathbf{r}) \rightarrow \infty$ as $\mathbf{r} \rightarrow \mathbf{r}_0$ (or $\mathbf{r} \rightarrow \infty$), then $f(\mathbf{r}) \rightarrow \infty$.

87.6. Study Problems.

Problem 13.1. Find the limit $\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r})$ or show that it does not exist, where

$$f(\mathbf{r}) = f(x, y, z) = (x^2 + 2y^2 + 4z^2) \ln(x^2 + y^2 + z^2), \quad \mathbf{r}_0 = (0, 0, 0).$$

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 $\mathbf{r}(t) = (at, bt, ct)$ for some constants a, b , and c that do not vanish simultaneously so that they define the direction of the line through the origin. Then

$$f(\mathbf{r}(t)) = At^2 \ln(Bt^2) = 2At^2 \ln t + At^2 \ln B,$$

where $A = a^2 + 2b^2 + 4c^2$ and $B = a^2 + b^2 + c^2 > 0$. By l'Hospital's rule,

$$\lim_{t \rightarrow 0^+} t^2 \ln t = \lim_{t \rightarrow 0^+} \frac{\ln t}{t^{-2}} = \lim_{t \rightarrow 0^+} \frac{t^{-1}}{-2t^{-3}} = -\frac{1}{2} \lim_{t \rightarrow 0^+} t = 0,$$

and therefore $f(\mathbf{r}(t)) \rightarrow 0$ as $t \rightarrow 0^+$. So, if the limit exists, then it must be equal to 0.

Step 4. Put $R^2 = x^2 + y^2 + z^2$. Since the limit $R \rightarrow 0^+$ is of interest, one can always assume that $R < 1$ so that $\ln R^2 = 2 \ln R < 0$. By making use of the inequalities $|x| \leq R$, $|y| \leq R$, and $|z| \leq R$, one has $R^2 \leq x^2 + 2y^2 + 4z^2 \leq 7R^2$. By multiplying the latter inequality by $\ln R^2 < 0$, $R^2 \ln R^2 \geq f(\mathbf{r}) \geq 7R^2 \ln(R^2)$. Since $t \ln t \rightarrow 0$ as $t = R^2 \rightarrow 0^+$, the limit exists and equals 0 by the squeeze principle. \square

Problem 13.2. Prove that the limit $\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r})$ exists, where

$$f(\mathbf{r}) = f(x, y) = \frac{1 - \cos(x^2 y)}{x^2 + 2y^2}, \quad \mathbf{r}_0 = (0, 0),$$

and find a disk centered at \mathbf{r}_0 in which values of f deviate from the limit by 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

$$\begin{aligned}\lim_{t \rightarrow 0^+} f(\mathbf{r}(t)) &= \lim_{t \rightarrow 0^+} \frac{1 - \cos(at^3)}{t^2(1 + 2a^2)} = \frac{1}{1 + 2a^2} \lim_{u \rightarrow 0^+} \frac{1 - \cos(au^{3/2})}{u} \\ &= \frac{1}{1 + 2a^2} \lim_{u \rightarrow 0^+} \frac{au^{1/2} \sin(au^{3/2})}{1} = 0,\end{aligned}$$

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 $1 - \cos u = 2 \sin^2(u/2) \leq u^2/2$, where the inequality $|\sin x| \leq |x|$ has been used. Put $R^2 = x^2 + y^2$. Then, by making use of the above inequality with $u = x^2y$ together with $|x| \leq R$ and $|y| \leq R$, the following chain of inequalities is obtained:

$$|f(\mathbf{r}) - 0| \leq \frac{(x^2y)^2/2}{x^2 + 2y^2} = \frac{(x^2y)^2/2}{R^2 + y^2} \leq \frac{(x^2y)^2/2}{R^2} \leq \frac{1}{2} \frac{R^6}{R^2} = \frac{R^4}{2} \rightarrow 0$$

as $R \rightarrow 0^+$. By the squeeze principle, the limit exists and equals 0. It follows from $|f(\mathbf{r})| \leq R^4/2$ that $|f(\mathbf{r})| < \varepsilon$ whenever $R^4/2 < \varepsilon$ or $R = \|\mathbf{r} - \mathbf{r}_0\| < \delta(\varepsilon) = (2\varepsilon)^{1/4} = 0.1$. \square

Problem 13.3. Find the limit $\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} f(\mathbf{r})$ 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 $f(\mathbf{r}(t)) = at^3/t^2(1 - a^2) = a(1 - a^2)^{-1}t \rightarrow 0$ as $t \rightarrow 0^+$. 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 $0 < \|\mathbf{r} - \mathbf{r}_0\| < \delta$, 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 than 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 $0 < \|\mathbf{r} - \mathbf{r}_0\| < \delta$. 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. Observe that the line $\mathbf{L}(t) = (t, t)$ (or $y = x$) is tangent to the curve $\mathbf{r}(t)$ at the origin because $\mathbf{r}'(0) = (1, 1)$ for $n > 1$. The term $-at^n$ in $\mathbf{r}(t)$ models a small deviation of the curve from the line $y = x$ in the vicinity of which the function f is expected to be unbounded. Then $f(\mathbf{r}(t)) = (t^3 - at^{n+2}) / (2at^{n+1} - a^2t^{2n})$. This function tends to a number as $t \rightarrow 0^+$ 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 $n = 2$, $f(\mathbf{r}(t)) = (t^3 - at^4) / (2at^3 - a^2t^4) = (1 - at^2) / (2a - a^2t) \rightarrow 1/(2a)$ as $t \rightarrow 0^+$ and $f(\mathbf{r}(t))$ diverges for $n > 2$ in this limit. \square

Problem 13.4. Find

$$\lim_{\mathbf{r} \rightarrow \infty} \frac{\ln(x^2 + y^4)}{x^2 + 2y^2}$$

or show that the limit does not exist.

SOLUTION: Step 1. Does not apply.

Step 2. No substitution exists to reduce the limit to a one-variable limit.

Step 3. Put $(x, y) = (t, at)$ and let $t \rightarrow \infty$. Then $\|\mathbf{r}\| \rightarrow \infty$ as $t \rightarrow \infty$. One has $f(t, at) = \ln(t^2 + a^4t^4) / (t^2 + 2a^2t^2)$. For large values of t , $\ln(t^2 + a^4t^4) \approx \ln(a^4t^4) = \ln(t^4) + \ln(a^4) \approx 4 \ln t$ if $a \neq 0$ and $f(t, 0) = 2 \ln t / t^2$. Therefore, $f(t, at)$ behaves as $\ln t / t^2 \rightarrow 0$ as $t \rightarrow \infty$ (by l'Hospital's rule). So the limit along all straight lines is 0.

Step 4. Put $R = \sqrt{x^2 + y^2}$ so that $|x| \leq R$ and $|y| \leq R$. Then, owing to the monotonicity of the logarithm function, $\ln(x^2 + y^4) \leq \ln(R^2 + R^4) \leq \ln(2R^4)$ for $R \geq 1$. The denominator of the ratio f can be estimated from below: $x^2 + 2y^2 = x^2 + y^2 + y^2 = R^2 + y^2 \geq R^2$. Hence, for $R > 1$,

$$|f(x, y) - 0| \leq \frac{\ln(4R^4)}{R^2} = \frac{4 \ln R + \ln 4}{R^2} \rightarrow 0 \quad \text{as } R \rightarrow \infty.$$

Thus, by the squeeze principle the limit is indeed 0. \square

87.7. Exercises.**(1)** Prove the following statements:(i) Let $f(x, y) = (x - y)/(x + y)$. Then

$$\lim_{x \rightarrow 0} \left(\lim_{y \rightarrow 0} f(x, y) \right) = 1, \quad \lim_{y \rightarrow 0} \left(\lim_{x \rightarrow 0} f(x, y) \right) = -1,$$

but the limit of $f(x, y)$ as $(x, y) \rightarrow (0, 0)$ does not exist.(ii) Let $f(x, y) = x^2 y^2 / (x^2 y^2 + (x - y)^2)$. Then

$$\lim_{x \rightarrow 0} \left(\lim_{y \rightarrow 0} f(x, y) \right) = \lim_{y \rightarrow 0} \left(\lim_{x \rightarrow 0} f(x, y) \right) = 0,$$

but the limit of $f(x, y)$ as $(x, y) \rightarrow (0, 0)$ does not exist.(iii) Let $f(x, y) = (x + y) \sin(1/x) \sin(1/y)$. Then the limits

$$\lim_{x \rightarrow 0} \left(\lim_{y \rightarrow 0} f(x, y) \right) \quad \text{and} \quad \lim_{y \rightarrow 0} \left(\lim_{x \rightarrow 0} f(x, y) \right)$$

do not exist, but the limit of $f(x, y)$ exists and equals 0 as $(x, y) \rightarrow (0, 0)$. Does the result contradict Theorem 13.8? Explain.**(2)** Find each of the following limits or show that it does not exist:

(i) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{\cos(xy + z)}{x^4 + y^2 z^2 + 4}$

(ii) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{\sin(xy) - xy}{(xy)^3}$

(iii) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{\sqrt{xy^2 + 1} - 1}{xy^2}$

(iv) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{\sin(xy^3)}{x^2}$

(v) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{x^3 + y^5}{x^2 + 2y^2}$

(vi) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{e^{\|\mathbf{r}\|} - 1 - \|\mathbf{r}\|}{\|\mathbf{r}\|^2}$

(vii) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{x^2 + \sin^2 y}{x^2 + 2y^2}$

(viii) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{xy^2 + x \sin(xy)}{x^2 + 2y^2}$

(ix) $\lim_{(x,y) \rightarrow (1,0)} \frac{\ln(x + e^y)}{\sqrt{x^2 + y^2}}$

(x) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} (x^2 + y^2)^{x^2 y^2}$

(xi) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{1}{xy} \tan\left(\frac{xy}{1 + xy}\right)$

(xii) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \ln\left(\frac{\sin(x^2 - y^2)}{x^2 - y^2}\right)^2$

(xiii) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{\sqrt{xy + 1} - 1}{y\sqrt{x}}$

(xiv) $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{x^b y^a}{x^a + y^b}, \quad 0 < b < a$

(3) Let $f(x, y) = (|x| + |y| - |x + y|)/(x^2 + y^2)^k$, if $x^2 + y^2 \neq 0$ and $f(0, 0) = c$. Find all values of constants c and $k > 0$ at which the function is continuous at the origin.**(4)** Let $f(x, y) = x^2 y / (x^4 + y^2)$ if $x^2 + y^2 \neq 0$ and $f(0, 0) = 0$. Show that f is continuous along any straight line through the origin; that is, $F(t) = f(x(t), y(t))$ is continuous for all t , where $x(t) = t \cos \theta$,

$y(t) = t \sin \theta$ for any fixed θ , but f is not continuous at $(0, 0)$. *Hint:* Investigate the limits of f along power curves leading to the origin.

(5) Let $f(x, y)$ be continuous in a rectangle $a < x < b$, $c < y < d$. Let $g(x)$ be continuous on the interval (a, b) and take values in (c, d) . Prove that the function $F(x) = f(x, g(x))$ is continuous on (a, b) .

(6) Investigate the limits of the function $f(x, y) = x^2 e^{-(x^2 - y)}$ along the rays $x(t) = \cos(\theta)t$, $y(t) = \sin(\theta)t$ as $t \rightarrow \infty$ for all $0 \leq \theta \leq 2\pi$. Are the values of the function arbitrarily small for all $\|\mathbf{r}\| > \delta$ if δ is large enough? Does the limit $\lim_{\mathbf{r} \rightarrow \infty} f(x, y)$ exist?

(7) Find the limit or show that it does not exist:

$$\begin{array}{ll} \text{(i)} \quad \lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{\sin(x^2 + y^2 + z^2)}{x^4 + y^4 + z^4} & \text{(ii)} \quad \lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{x^2 + 2y^2 + 3z^2}{x^4 + y^2 z^4} \\ \text{(iii)} \quad \lim_{\mathbf{r} \rightarrow \infty} \frac{\ln(x^2 y^2 z^2)}{x^2 + y^2 + z^2} & \text{(iv)} \quad \lim_{\mathbf{r} \rightarrow \infty} \frac{e^{3x^2 + 2y^2 + z^2}}{(x^2 + 2y^2 + 3z^2)^{2012}} \\ \text{(v)} \quad \lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{z}{x^2 + y^2 + z^2} & \text{(vi)} \quad \lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{z}{x^2 + y^2 + z^2} \quad \text{if } z < 0 \\ \text{(vii)} \quad \lim_{\mathbf{r} \rightarrow \infty} \frac{x^2 + y^2}{x^2 + y^4} & \text{(viii)} \quad \lim_{\mathbf{r} \rightarrow \infty} \sin\left(\frac{\pi x}{2x + y}\right) \\ \text{(ix)} \quad \lim_{\mathbf{r} \rightarrow \infty} (x^2 + y^2)e^{-|x+y|} & \text{(x)} \quad \lim_{\mathbf{r} \rightarrow \infty} \left(\frac{xy}{x^2 + y^2}\right)^{x^2} \end{array}$$

(8) Find the repeated limits

$$\lim_{x \rightarrow 1} \left(\lim_{y \rightarrow 0} \log_x(x + y) \right) \quad \text{and} \quad \lim_{y \rightarrow 0} \left(\lim_{x \rightarrow 1} \log_x(x + y) \right).$$

What can be said about the corresponding two-variable limit?

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 $h'(x_0)$, and the slope along the south–north direction is $g'(y_0)$. These slopes are called *partial derivatives* of f and denoted as

$$\begin{aligned}\frac{\partial f}{\partial x}(x_0, y_0) &= \left. \frac{d}{dx} f(x, y_0) \right|_{x=x_0}, \\ \frac{\partial f}{\partial y}(x_0, y_0) &= \left. \frac{d}{dy} f(x_0, y) \right|_{y=y_0}.\end{aligned}$$

The partial derivatives are also 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 above analysis of the geometrical significance of partial derivatives is illustrated in Figure 13.6. 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 13.14. (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 $B_\delta(\mathbf{r}_0) = \{\mathbf{r} \mid \|\mathbf{r} - \mathbf{r}_0\| < \delta\}$ 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 $D = \{(x, y) \mid x^2 + y^2 \leq 1\}$, then any point of D that does not lie on the circle $x^2 + y^2 = 1$ is an interior point.

DEFINITION 13.15. (Open Sets).

A set D in a Euclidean space is said to be open if all points of D are interior points of D .

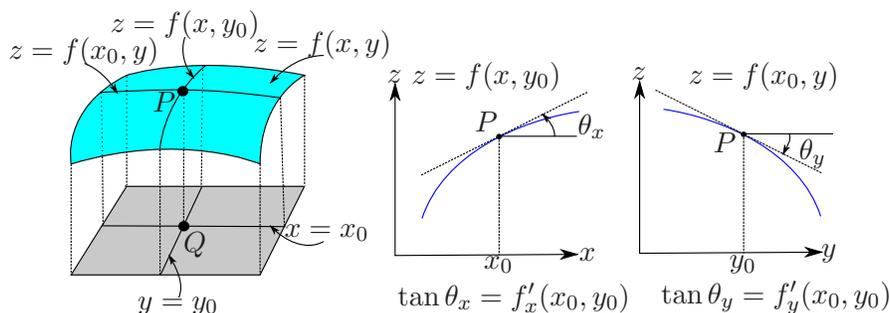


FIGURE 13.6. Geometrical significance of partial derivatives. **Left:** The graph $z = f(x, y)$ and its cross sections by the coordinate planes $x = x_0$ and $y = y_0$. The point $Q = (x_0, y_0, 0)$ is in the domain of f and the point $P = (x_0, y_0, f(x_0, y_0))$ lies on the graph. **Middle:** The cross section $z = f(x, y_0)$ of the graph in the plane $y = y_0$ and the tangent line to it at the point P . The slope $\tan \theta_x$ of the tangent line is determined by the partial derivative $f'_x(x_0, y_0)$ at the point Q . **Right:** The cross section $z = f(x_0, y)$ of the graph in the plane $x = x_0$ and the tangent line to it at the point P . The slope $\tan \theta_y$ of the tangent line is determined by the partial derivative $f'_y(x_0, y_0)$ at the point Q . Here $\theta_y < 0$ as it is counted clockwise.

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

$$\mathbf{r} = (x_1, x_2, \dots, x_n) = x_1\hat{\mathbf{e}}_1 + x_2\hat{\mathbf{e}}_2 + \cdots + x_n\hat{\mathbf{e}}_n,$$

where $\hat{\mathbf{e}}_i$ is the n -tuple whose components are zeros except the i th one, which is equal to 1. Obviously, $\|\hat{\mathbf{e}}_i\| = 1$, $i = 1, 2, \dots, n$.

DEFINITION 13.16. (Partial Derivatives at a Point).

Let f be a function of several variables (x_1, x_2, \dots, x_n) . Let D be the domain of f and let \mathbf{r}_0 be an interior point of D . If the limit

$$f'_{x_i}(\mathbf{r}_0) = \lim_{h \rightarrow 0} \frac{f(\mathbf{r}_0 + h\hat{\mathbf{e}}_i) - f(\mathbf{r}_0)}{h}$$

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 $\mathbf{r}_0 + h\hat{\mathbf{e}}_i$, $i = 1, 2, \dots, n$, 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 $B_\delta(\mathbf{r}_0)$ of sufficiently small radius $\delta = |h|$ are in D .

Remark. It is also common to omit “prime” in the notations for partial derivatives. For example, the partial derivative of f with respect to x is denoted as f_x . In what follows, the notation introduced in Definition 13.16 will be used.

Let $\mathbf{r}_0 = (a_1, a_2, \dots, a_n)$, 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 i th 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

$$(13.1) \quad f'_{x_i}(\mathbf{r}_0) = \lim_{h \rightarrow 0} \frac{F(a_i + h) - F(a_i)}{h} = \left. \frac{dF(x_i)}{dx_i} \right|_{x_i=a_i}$$

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 13.16. Find the partial derivatives of $f(x, y, z) = x^3 - y^2z$ at the point $(1, 2, 3)$.

SOLUTION: By the rule (13.1),

$$\begin{aligned} f'_x(1, 2, 3) &= \left. \frac{d}{dx} f(x, 2, 3) \right|_{x=1} = \left. \frac{d}{dx} (x^3 - 12) \right|_{x=1} = 3, \\ f'_y(1, 2, 3) &= \left. \frac{d}{dy} f(1, y, 3) \right|_{y=2} = \left. \frac{d}{dy} (1 - 3y^2) \right|_{y=2} = -12, \\ f'_z(1, 2, 3) &= \left. \frac{d}{dz} f(1, 2, z) \right|_{z=3} = \left. \frac{d}{dz} (1 - 4z) \right|_{z=3} = -4. \end{aligned}$$

□

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 13.16 defines the temperature in degrees Celsius as a function of the position whose coordinates are given in meters, then, at the point $(1, 2, 3)$, the temperature increases at rate of 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 . 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

$$f'_{x_i}(x_1, x_2, \dots, x_n) = \frac{\partial}{\partial x_i} f(x_1, x_2, \dots, x_n);$$

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 13.17. 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

$$\begin{aligned} f'_x(x, y) &= \frac{\partial}{\partial x} f(x, y) = \left(\frac{\partial}{\partial x} x \right) \sin(xy) + x \frac{\partial}{\partial x} \sin(xy) \\ &= \sin(xy) + xy \cos(xy) \end{aligned}$$

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

$$\begin{aligned} \frac{\partial}{\partial x_i}(cf) &= c \frac{\partial f}{\partial x_i}, & \frac{\partial}{\partial x_i}(f+g) &= \frac{\partial f}{\partial x_i} + \frac{\partial g}{\partial x_i}, \\ \frac{\partial}{\partial x_i}(fg) &= \frac{\partial f}{\partial x_i} g + f \frac{\partial g}{\partial x_i}, & \frac{\partial}{\partial x_i}\left(\frac{f}{g}\right) &= \frac{\frac{\partial f}{\partial x_i} g - f \frac{\partial g}{\partial x_i}}{g^2}. \end{aligned}$$

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

$$(13.2) \quad \frac{\partial f}{\partial x_i} = h'(g) \frac{\partial g}{\partial x_i}.$$

EXAMPLE 13.18. Find the partial derivatives of the function $f(\mathbf{r}) = \|\mathbf{r}\|^{-1}$, where $\mathbf{r} = (x_1, x_2, \dots, x_n)$.

SOLUTION: Put $h(u) = u^{-1/2}$ and $g(\mathbf{r}) = x_1^2 + x_2^2 + \dots + x_n^2 = \|\mathbf{r}\|^2$. 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) $f(x, y) = (x - y)/(x + y)$, $f'_x(1, 2)$, $f'_y(1, 2)$
- (ii) $f(x, y, z) = (xy + z)/(z + y)$, $f'_x(1, 2, 3)$, $f'_y(1, 2, 3)$, $f'_z(1, 2, 3)$
- (iii) $f(\mathbf{r}) = (x_1 + 2x_2 + \dots + nx_n)/(1 + \|\mathbf{r}\|^2)$, $f'_{x_i}(\mathbf{0})$, $i = 1, 2, \dots, n$
- (iv) $f(x, y, z) = x \sin(yz)$, $f'_x(1, 2, \pi/2)$, $f'_y(1, 2, \pi/2)$, $f'_z(1, 2, \pi/2)$
- (v) $f(x, y) = x + (y - 1) \sin^{-1}(\sqrt{x/y})$, $f'_x(1, 1)$, $f'_y(1, 1)$
- (vi) $f(x, y) = (x^3 + y^3)^{1/3}$, $f'_x(0, 0)$, $f'_y(0, 0)$
- (vii) $f(x, y) = \sqrt{|xy|}$, $f'_x(0, 0)$, $f'_y(0, 0)$

(2) Find the partial derivatives of each of the following functions:

- (i) $f(x, y) = (x + y^2)^n$
- (ii) $f(x, y) = x^y$
- (iii) $f(x, y) = xe^{(x+2y)^2}$
- (iv) $f(x, y) = \sin(xy) \cos(x^2 + y^2)$
- (v) $f(x, y, z) = \ln(x + y^2 + z^3)$
- (vi) $f(x, y, z) = xy^2 \cos(z^2x)$
- (vii) $f(\mathbf{r}) = (a_1x_1 + a_2x_2 + \dots + a_nx_n)^m = (\mathbf{a} \cdot \mathbf{r})^m$
- (viii) $f(x, y) = \tan^{-1}(x/y)$
- (ix) $f(x, y) = \sin^{-1}(x/\sqrt{x^2 + y^2})$
- (x) $f(x, y, z) = x^{y^z}$
- (xi) $f(x, y, z) = x^{y/z}$
- (xii) $f(x, y) = \tan(x^2/y)$
- (xiii) $f(x, y, z) = \sin(x \sin(y \sin z))$
- (xiv) $f(x, y) = (x + y^2)/(x^2 + y)$
- (xv) $f(x, y, z) = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{r})$

where \mathbf{a} and \mathbf{b} are constant vectors

(xvi) $f(x, y, z) = \|\mathbf{a} \times \mathbf{r}\|$ where \mathbf{a} is a constant vector

(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) $f(x, y) = xy/(x + y)$, $P_0(1, 2)$
- (ii) $f(x, y) = (x^2 - 2y^2)^{1/3}$, $P_0(1, 1)$
- (iii) $f(x, y) = x^2 \sin(xy)$, $P_0(-1, \pi)$

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 partial derivatives are

$$\begin{aligned} \frac{\partial f}{\partial x} &\longmapsto \frac{\partial}{\partial x} \frac{\partial f}{\partial x} = \frac{\partial^2 f}{\partial x^2}, & \frac{\partial}{\partial y} \frac{\partial f}{\partial x} &= \frac{\partial^2 f}{\partial y \partial x}, \\ \frac{\partial f}{\partial y} &\longmapsto \frac{\partial}{\partial x} \frac{\partial f}{\partial y} = \frac{\partial^2 f}{\partial x \partial y}, & \frac{\partial}{\partial y} \frac{\partial f}{\partial y} &= \frac{\partial^2 f}{\partial y^2}. \end{aligned}$$

Throughout the text, brief notations for higher-order partial 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} . Partial derivatives of the third order are defined as partial derivatives of second-order partial derivatives, and so on.

EXAMPLE 13.19. For the function $f(x, y) = x^4 - x^2y + y^2$, find all second- and third-order partial derivatives.

SOLUTION: The first partial derivatives are $f'_x = 4x^3 - 2xy$ and $f'_y = -x^2 + 2y$. Then the second partial derivatives are

$$\begin{aligned} f''_{xx} &= (4x^3 - 2xy)'_x = 12x^2 - 2y, & f''_{yy} &= (-x^2 + 2y)'_y = 2, \\ f''_{xy} &= (4x^3 - 2xy)'_y = -2x, & f''_{yx} &= (-x^2 + 2y)'_x = -2x. \end{aligned}$$

The third partial derivatives are found similarly:

$$\begin{aligned} f'''_{xxx} &= (12x^2 - 2y)'_x = 24x, & f'''_{yyy} &= (2)'_y = 0, \\ f'''_{xxy} &= (12x^2 - 2y)'_y = -2, & f'''_{xyx} &= f'''_{yxx} = (-2x)'_x = -2, \\ f'''_{yyx} &= (2)'_x = 0, & f'''_{yxy} &= f'''_{xyy} = (-2x)'_y = 0. \end{aligned}$$

□

In contrast to the one-variable case, there are higher-order partial 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 is *independent of the order in which the partial derivatives have been taken*. Is this a peculiarity of the function considered or a general feature of higher-order partial derivatives? The following theorem answers this question.

THEOREM 13.9. (Clairaut's Theorem).

Let f be a function of several variables (x_1, x_2, \dots, x_n) that is defined on an open ball D in a Euclidean space. If the second partial derivatives $f''_{x_i x_j}$ and $f''_{x_j x_i}$, where $j \neq i$, are continuous functions on D , then $f''_{x_i x_j} = f''_{x_j x_i}$ at any point of D .

A consequence of Clairaut's theorem can be proved. It asserts that *the result of partial differentiation does not depend on the order in which the derivatives have been taken if all higher-order partial derivatives in question are continuous*. It is not always necessary to calculate higher-order partial derivatives in all possible orders to verify the hypothesis of Clairaut's theorem (i.e., the continuity of the partial derivatives). Partial derivatives of polynomials are polynomials and hence continuous. By the quotient rule for partial derivatives, rational functions have continuous partial derivatives (where the denominator does not vanish). Derivatives of basic elementary functions like the sine and cosine and exponential functions are continuous. So compositions of these functions with multivariable polynomials or rational functions have continuous partial derivatives of any order. In other words, the continuity of higher-order partial derivatives can often be established by different, simpler means.

EXAMPLE 13.20. *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 partial derivatives. Therefore, 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 partial derivatives in question coincide and are equal to

$$\begin{aligned} f'''_{xyz} &= (f'_{xz})''_{yz} = (2x \cos(x^2 + yz))''_{yz} = (-2xz \sin(x^2 + yz))'_z \\ &= -2x \sin(x^2 + yz) - 2xyz \cos(x^2 + yz). \end{aligned}$$

□

89.1. Reconstruction of a Function from Its Partial Derivatives. One of the standard problems in calculus is finding a function $f(x)$ on an interval I if its derivative $f'(x) = F(x)$ is known. A sufficient condition for the existence of a solution is the continuity of $F(x)$ on I . In this case,

$$f'(x) = F(x) \implies f(x) = \int F(x) dx.$$

The indefinite integral is given by the sum of a particular antiderivative of F and an arbitrary constant (see *Concepts in Calculus I*). A similar problem can be posed for a function of several variables. Given the first partial derivatives

$$(13.3) \quad f'_{x_i}(\mathbf{r}) = F_i(\mathbf{r}), \quad i = 1, 2, \dots, n,$$

find $f(\mathbf{r})$ if it exists. The existence of such f is a more subtle question in the case of several variables. Suppose that the partial derivatives $\partial F_i / \partial x_j$ exist and are continuous functions in an open ball. Then taking the partial derivative $\partial / \partial x_j$ of both sides of (13.3) and applying Clairaut's theorem, one infers that

$$(13.4) \quad f''_{x_i x_j} = f''_{x_j x_i} \implies \frac{\partial F_i}{\partial x_j} = \frac{\partial F_j}{\partial x_i}.$$

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 13.21. 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 with respect to x is a function whose *partial* derivative with respect to x is $3x^2y$. It is easy to verify that x^3y satisfies this requirement. It is obtained by taking an antiderivative of $3x^2y$ 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 an antiderivative, $x^3y + c$ 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 $f(x, y) = x^3y + g(y)$ for some $g(y)$.

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 13.21, 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 13.22. 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

$$f'_x = yz + 2x \implies f(x, y, z) = xyz + x^2 + g(y, z),$$

for some $g(y, z)$. The substitution of f into the second equations yields

$$\begin{aligned} f'_y = xz + 3y^2 &\implies xz + g'_y(y, z) = xz + 3y^2 \\ &\implies g'_y(y, z) = 3y^2 \\ &\implies g(y, z) = y^3 + h(z) \\ &\implies f(x, y, z) = xyz + x^2 + y^3 + h(z), \end{aligned}$$

for some $h(z)$. The substitution of f into the third equation yields

$$\begin{aligned} f'_z = xy + 4z^3 &\implies xy + h'(z) = xy + 4z^3 \\ &\implies h'(z) = 4z^3 \\ &\implies h(z) = z^4 + c \\ &\implies f(x, y, z) = xyz + x^2 + y^3 + z^4 + c, \end{aligned}$$

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 stated in the form of partial differential equations.

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 in 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

$$n'_t = k(n''_{xx} + n''_{yy} + n''_{zz}),$$

where the parameter k is a diffusion constant. So the concentration as a function of the spatial position and time must satisfy the above partial differential equation.

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.

Laplace and Poisson Equations. The equation

$$u''_{xx} + u''_{yy} + u''_{zz} = f,$$

where f is a given nonzero 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 13.23. 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 $q = ct - \hat{\mathbf{n}} \cdot \mathbf{r} = ct - n_1x - n_2y - n_3z$. By the chain rule (13.2), $u'_t = q'_t h'(q)$ and similarly for the other partial derivatives. Therefore, $u'_t = ch'(q)$, $u''_{tt} = c^2 h''(q)$, $u'_x = -n_1 h'(q)$, $u''_{xx} = n_1^2 h''(q)$, and, in the same fashion, $u''_{yy} = n_2^2 h''(q)$, and $u''_{zz} = n_3^2 h''(q)$. 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 . \square

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 reasoning 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 about 5 seconds later. Conversely, if one sees a lightning strike and starts counting seconds until the thunder is heard, then one could estimate the distance to the lightning. The sound travels 1 mile in about 4.7 seconds.

89.3. Study Problems.

Problem 13.5. Find the value of a constant a for which the function

$$u(\mathbf{r}, t) = t^{-3/2} e^{-ar^2/t}, \quad r = \|\mathbf{r}\|,$$

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:

$$\begin{aligned} \frac{\partial u}{\partial x} &= \frac{\partial u}{\partial r^2} \frac{\partial r^2}{\partial x} = 2x \frac{\partial u}{\partial r^2} = -\frac{2ax}{t} u, \\ u''_{xx} &= \frac{\partial}{\partial x} \left(\frac{\partial u}{\partial x} \right) = -\frac{2a}{t} u - \frac{2ax}{t} \frac{\partial u}{\partial x} = \left(-\frac{2a}{t} + \frac{4a^2 x^2}{t^2} \right) u. \end{aligned}$$

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

$$k(u''_{xx} + u''_{yy} + u''_{zz}) = \left(-\frac{6ka}{t} + \frac{4ka^2 r^2}{t^2} \right) u.$$

Using the product rule to calculate the partial derivative with respect to time, one finds for the left side

$$u'_t = -\frac{3}{2} t^{-5/2} e^{-ar^2/t} + t^{-3/2} e^{-ar^2/t} \frac{ar^2}{t^2} = \left(-\frac{3}{2t} + \frac{ar^2}{t^2} \right) u.$$

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: $6ka = 3/2$ (as the equality of the coefficients at $1/t$) and $a = 4ka^2$ (as the equality of the coefficients at r^2/t^2). The only common solution of these conditions is $a = 1/(4k)$. \square

Problem 13.6. Consider the function

$$f(x, y) = \frac{x^3 y - xy^3}{x^2 + y^2} \quad \text{if } (x, y) \neq (0, 0) \quad \text{and} \quad f(0, 0) = 0.$$

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 quotient 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 partial derivative f'_y is obtained from f'_x by changing the sign of the latter and swapping x and y . The partial derivatives at $(0, 0)$ are found by the rule (13.1):

$$f'_x(0, 0) = \left. \frac{d}{dx} f(x, 0) \right|_{x=0} = 0, \quad f'_y(0, 0) = \left. \frac{d}{dy} f(0, y) \right|_{y=0} = 0.$$

The first-order partial derivatives are continuous functions (the proof is left to the reader as an exercise). Next, one has

$$\begin{aligned} f''_{xy}(0, 0) &= \left. \frac{d}{dy} f'_x(0, y) \right|_{y=0} = \lim_{h \rightarrow 0} \frac{f'_x(0, h) - f'_x(0, 0)}{h} \\ &= \lim_{h \rightarrow 0} \frac{-h - 0}{h} = -1, \\ f''_{yx}(0, 0) &= \left. \frac{d}{dx} f'_y(x, 0) \right|_{x=0} = \lim_{h \rightarrow 0} \frac{f'_y(h, 0) - f'_y(0, 0)}{h} \\ &= \lim_{h \rightarrow 0} \frac{h - 0}{h} = 1. \end{aligned}$$

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 quotient 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 $(x, y) \rightarrow (0, 0)$ along the straight line $(x, y) = (t, at)$, $t \rightarrow 0$, one infers that the limit depends on the slope a , and hence the two-dimensional limit does not exist, that is, $\lim_{(x,y) \rightarrow (0,0)} f''_{xy}(x, y) \neq f''_{xy}(0, 0) = -1$, and f''_{xy} is not continuous at $(0, 0)$. The technical details are left to the reader. \square

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) $f(x, y, z) = (x + y)/(x + 2z)$
- (v) $f(x, y) = \cos^{-1}(\sqrt{x/y})$
- (vi) $f(x, y) = x^y$

(2) Explain without explicit calculation of higher-order partial derivatives that the hypothesis of Clairaut's theorem is satisfied for the following functions

- (i) $f(x, y, z) = \sin(x^2 + y - z) \cos(xy)$
- (ii) $f(x, y) = \sin(x + y^2)/(x^2 + y^2)$, $x^2 + y^2 \neq 0$
- (iii) $f(x, y, z) = e^{x^2yz}(y^2 + zx^4)$
- (iv) $f(x, y) = \ln(1 + x^2 + y^4)/(x^2 - y^2)$, $x^2 \neq y^2$
- (v) $f(x, y, z) = (x + yz^2 - xz^5)/(1 + x^2y^2z^4)$

(3) 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) $f(x, y, z) = \sin(xy)e^z$, $\partial f^5/\partial z^5$, $f^{(4)}_{xyzz}$, $f^{(4)}_{zyxz}$, $f^{(4)}_{zxyz}$
- (iv) $f(x, y, z, t) = \sin(x + 2y + 3z - 4t)$, $f^{(4)}_{abcd}$, where $abcd$ denotes all permutations of $xyzt$
- (v) $f(x, y) = e^{xy}(y^2 + x)$, $f^{(4)}_{abcd}$, where $abcd$ denotes all permutations of $xxxy$
- (vi) $f(x, y, z) = \tan^{-1}\left(\frac{x+y+z-xyz}{1-xy-xz-yz}\right)$, $f^{(3)}_{abc}$, where abc denotes all permutations of xyz
- (vii) $f(x, y, z, t) = \ln((x - y)^2 + (z - t)^2)^{-1/2}$, $f^{(4)}_{abcd}$, where $abcd$ are all permutations of $xyzt$.
- (viii) $f(x, y) = e^x \sin(y)$, $\frac{\partial^{n+m} f}{\partial^n x \partial^m y}(0, 0)$

(4) Given partial derivatives, find the function or show that it does not exist:

- (i) $f'_x = 3x^2y$, $f'_y = x^3 + 3y^2$
- (ii) $f'_x = yz + 3x^2$, $f'_y = xz + 4y$, $f'_z = xy + 1$
- (iii) $f'_{x_k} = kx_k$, $k = 1, 2, \dots, n$
- (iv) $f'_x = xy + z$, $f'_y = x^2/2$, $f'_z = x + y$
- (v) $f'_x = \sin(xy) + xy \cos(xy)$, $f'_y = x^2 \cos(xy) + 1$

(5) 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) $f(x, y, t) = g(ct - ax - by) + h(ct + ax + by)$, $f''_{tt} = c^2(f''_{xx} + f''_{yy})$
if $a^2 + b^2 = 1$ and g and h are twice differentiable functions
- (iii) $f(x, y) = \ln(x^2 + y^2)$, $f''_{xx} + f''_{yy} = 0$
- (iv) $f(x, y) = \ln(e^x + e^y)$, $f'_x + f'_y = 1$ and $f''_{xx} f''_{yy} - (f''_{xy})^2 = 0$
- (v) $f(\mathbf{r}) = \exp(\mathbf{a} \cdot \mathbf{r})$, where $\mathbf{a} \cdot \mathbf{a} = 1$, $f''_{x_1 x_1} + f''_{x_2 x_2} + \dots + f''_{x_n x_n} = f$
- (vi) $f(\mathbf{r}) = \|\mathbf{r}\|^{2-n}$, $f''_{x_1 x_1} + f''_{x_2 x_2} + \dots + f''_{x_n x_n} = 0$ for $\|\mathbf{r}\| \neq 0$
- (vii) $f(x, y, z) = \sin(k\|\mathbf{r}\|)/\|\mathbf{r}\|$, $f''_{xx} + f''_{yy} + f''_{zz} + k^2 f = 0$ (Helmholtz equation)

(6) Find a relation between the constants a , b , and c such that the function

$u(x, y, t) = \sin(ax + by + ct)$ satisfies the wave equation $u''_{tt} - u''_{xx} - u''_{yy} = 0$. Give a geometrical description of such a relation, for example, by setting values of c on a vertical axis and the values of a and b on two horizontal axes.

(7) Let $f(x, y, z) = u(t)$, where $t = xyz$. Show that $f^{(3)}_{xyz} = F(t)$ and find $F(t)$.

(8) Find $(f'_x)^2 + (f'_y)^2 + (f'_z)^2$ and $f''_{xx} + f''_{yy} + f''_{zz}$ if

(i) $f = x^3 + y^3 + z^3 - 3xyz$

(ii) $f = (x^2 + y^2 + z^2)^{-1/2}$

(9) Let the action of K on a function f be defined by $Kf = xf'_x + yf'_y$. Find Kf , $K^2f = K(Kf)$, and $K^3f = K(K^2f)$ if

(i) $f = x/(x^2 + y^2)$

(ii) $f = \ln \sqrt{x^2 + y^2}$

(10) Let $f(x, y) = xy/(x^2 + y^2)$ if $(x, y) \neq (0, 0)$ and $f(0, 0) = 0$. Does $f''_{xy}(0, 0)$ exist?

(11) If $f = f(x, y)$ and $g = g(x, y, z)$, solve the following equations:

(i) $f''_{xx} = 0$

(ii) $f''_{xy} = 0$

(iii) $\partial^n f / \partial y^n = 0$

(iv) $g'''_{xyz} = 0$

(12) Find $f(x, y)$ that satisfies:

(i) $f'_y = x^2 + 2y$, $f(x, x^2) = 1$

(ii) $f''_{yy} = 4$, $f(x, 0) = 2$, $f'_y(x, 0) = x$

(iii) $f''_{xy} = x + y$, $f(x, 0) = x$, $f(0, y) = y^2$

90. 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. Put $x = x_0 + \Delta x$. Then, by the definition of the derivative $f'(x_0)$,

$$\begin{aligned} \lim_{\Delta x \rightarrow 0} \frac{f(x) - L(x)}{\Delta x} &= \lim_{\Delta x \rightarrow 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} - f'(x_0) \\ &= f'(x_0) - f'(x_0) = 0. \end{aligned}$$

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,

$$(13.5) \quad f(x) = L(x) + \varepsilon(\Delta x) \Delta x, \quad \text{where } \varepsilon(\Delta x) \rightarrow 0 \text{ as } \Delta x \rightarrow 0.$$

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 $\varepsilon(\Delta x) = \Delta x$.

Conversely, consider a line through the point $(x_0, f(x_0))$ and assume that the condition (13.5) holds. If n is the slope of the line, then $L(x) = f(x_0) + n(x - x_0) = f(x_0) + n\Delta x$ and

$$\lim_{\Delta x \rightarrow 0} \frac{f(x) - L(x)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} - n = 0.$$

By the definition of the derivative $f'(x_0)$, the existence of this limit implies the existence of $f'(x_0)$ and the equality $n = f'(x_0)$. Thus, among all linear approximations of f near x_0 , *only* the line with the slope $n = f'(x_0)$ is a good approximation in the sense that the error of the approximation decreases faster than Δx with decreasing Δx , and *the very existence of a good linear approximation at $x = x_0$ is equivalent to differentiability of f at x_0* . For example, the function $f(x) = |x|$ is not differentiable at $x = 0$. A good linear approximation does not exist at $x_0 = 0$. Indeed, here $\Delta x = x$ and $L(x) = nx$. Hence, $(f(x) - L(x))/\Delta x = (|x| - nx)/x = |x|/x - n$, and no number n exists at which this difference vanishes in the limit $x \rightarrow 0$.

90.1. Differentiability of Multivariable Functions. Consider a function of two variables $f(x, y)$ and a point (x_0, y_0) in its domain. The most general linear function $L(x, y)$ with the property $L(x_0, y_0) = f(x_0, y_0)$ reads $L(x, y) = f(x_0, y_0) + n_1(x - x_0) + n_2(y - y_0)$, where n_1 and n_2 are arbitrary numbers. It defines a linear approximation to $f(x, y)$ near (x_0, y_0) in the sense that $L(x_0, y_0) = f(x_0, y_0)$. More generally, given a multivariable function $f(\mathbf{r})$, a linear function

$$L(\mathbf{r}) = f(\mathbf{r}_0) + \mathbf{n} \cdot (\mathbf{r} - \mathbf{r}_0)$$

is said to be a *linear approximation* to f near \mathbf{r}_0 in the sense that $L(\mathbf{r}_0) = f(\mathbf{r}_0)$. The dot product is defined in an m -dimensional Euclidean space if f is a function of m variables. The vector \mathbf{n} is an arbitrary vector so that $L(\mathbf{r})$ is the most general linear function satisfying the condition $L(\mathbf{r}_0) = f(\mathbf{r}_0)$. Note that in the case of two variables $x_1 = x$ and $x_2 = y$, $\mathbf{n} = (n_1, n_2)$ and $\mathbf{r} - \mathbf{r}_0 = (x - x_0, y - y_0)$ so that $\mathbf{n} \cdot (\mathbf{r} - \mathbf{r}_0) = n_1(x - x_0) + n_2(y - y_0)$.

DEFINITION 13.17. (Differentiable Functions).

The function f of several variables $\mathbf{r} = (x_1, x_2, \dots, x_m)$ on an open set D is said to be differentiable at a point $\mathbf{r}_0 \in D$ if there exists a good

linear approximation $L(\mathbf{r})$, i.e. a linear approximation $L(\mathbf{r})$ for which

$$(13.6) \quad \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} \frac{f(\mathbf{r}) - L(\mathbf{r})}{\|\mathbf{r} - \mathbf{r}_0\|} = 0.$$

If f is differentiable at all points of D , then f is said to be differentiable on D .

By this definition, the differentiability of a function is independent of the coordinate system chosen to label points of D (a linear function remains linear under general rotations and translations of the coordinate system and the distance $\|\mathbf{r} - \mathbf{r}_0\|$ is also invariant under these transformations). For functions of a single variable $f(x)$, the existence of a linear approximation at x_0 with the property (13.6) is equivalent to the existence of the derivative $f'(x_0)$. Indeed, put $x - x_0 = \Delta x$. Then $[f(x) - L(x)]/|\Delta x| = \pm [f(x) - L(x)]/\Delta x$ for all $\Delta x \neq 0$. Therefore, the condition (13.6) is equivalent to (13.5), which, in turn, is equivalent to the existence of $f'(x_0)$ as argued above.

THEOREM 13.10. *A linear approximation L to a multivariable function f near a point \mathbf{r}_0 that satisfies the property (13.6) is unique if it exists.*

PROOF. Let $L_1(\mathbf{r}) = f(\mathbf{r}_0) + \mathbf{n}_1 \cdot (\mathbf{r} - \mathbf{r}_0)$ and $L_2(\mathbf{r}) = f(\mathbf{r}_0) + \mathbf{n}_2 \cdot (\mathbf{r} - \mathbf{r}_0)$ be two linear approximations that satisfy the condition (13.6) for which $\mathbf{n}_2 \neq \mathbf{n}_1$. Making use of the identity

$$L_2(\mathbf{r}) - L_1(\mathbf{r}) = [f(\mathbf{r}) - L_1(\mathbf{r})] - [f(\mathbf{r}) - L_2(\mathbf{r})],$$

it is concluded that

$$\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} \frac{L_2(\mathbf{r}) - L_1(\mathbf{r})}{\|\mathbf{r} - \mathbf{r}_0\|} = \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} \frac{f(\mathbf{r}) - L_1(\mathbf{r})}{\|\mathbf{r} - \mathbf{r}_0\|} - \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} \frac{f(\mathbf{r}) - L_2(\mathbf{r})}{\|\mathbf{r} - \mathbf{r}_0\|} = 0.$$

Note that owing to the existence of the limit (13.6) for both linear functions L_1 and L_2 , the limit of the difference equals the difference of the limits (the basic law of limits). On the other hand, $L_2(\mathbf{r}) - L_1(\mathbf{r}) = (\mathbf{n}_2 - \mathbf{n}_1) \cdot (\mathbf{r} - \mathbf{r}_0)$. Put $\mathbf{n} = \mathbf{n}_2 - \mathbf{n}_1$. By assumption, $\|\mathbf{n}\| \neq 0$. Then

$$0 = \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} \frac{L_2(\mathbf{r}) - L_1(\mathbf{r})}{\|\mathbf{r} - \mathbf{r}_0\|} = \lim_{\mathbf{r} \rightarrow \mathbf{r}_0} \frac{\mathbf{n} \cdot (\mathbf{r} - \mathbf{r}_0)}{\|\mathbf{r} - \mathbf{r}_0\|} = \lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{\mathbf{n} \cdot \mathbf{r}}{\|\mathbf{r}\|}.$$

If a multivariable limit exists, then its value does not depend on a path along which the limit point is approached. In particular, take the straight line parallel to \mathbf{n} , $\mathbf{r} = \mathbf{n}t$, $t \rightarrow 0^+$, in the above relation. Then along this line, $\mathbf{r}/\|\mathbf{r}\| = \mathbf{n}/\|\mathbf{n}\|$ and hence

$$0 = \lim_{t \rightarrow 0^+} \frac{\mathbf{n} \cdot \mathbf{n}}{\|\mathbf{n}\|} = \frac{\mathbf{n} \cdot \mathbf{n}}{\|\mathbf{n}\|} = \|\mathbf{n}\| \quad \Rightarrow \quad \mathbf{n} = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{n}_1 = \mathbf{n}_2,$$

which is a contradiction. Thus, a good linear approximation is unique if it exists, $L_1(\mathbf{r}) = L_2(\mathbf{r})$. \square

90.2. Differentiability and Partial Derivatives. In the one-variable case, a function $f(x)$ is differentiable at x_0 if and only if it has the derivative $f'(x_0)$. Also, the existence of the derivative at x_0 implies continuity at x_0 (recall Calculus I). In the multivariable case, the relations between differentiability, continuity, and the existence of partial derivatives are more subtle.

THEOREM 13.11. (Properties of Differentiable Functions).

If f is differentiable at a point \mathbf{r}_0 , then it is continuous at \mathbf{r}_0 and its partial derivatives exist at \mathbf{r}_0 .

PROOF. The property (13.6) requires that the error of the linear approximation decrease faster than the distance $\|\mathbf{r} - \mathbf{r}_0\|$:

$$(13.7) \quad f(\mathbf{r}) = L(\mathbf{r}) + \varepsilon(\mathbf{r})\|\mathbf{r} - \mathbf{r}_0\|, \quad \text{where } \varepsilon(\mathbf{r}) \rightarrow 0 \text{ as } \mathbf{r} \rightarrow \mathbf{r}_0.$$

A linear function is continuous (it is a polynomial of degree 1). Therefore, $L(\mathbf{r}) \rightarrow L(\mathbf{r}_0) = f(\mathbf{r}_0)$ as $\mathbf{r} \rightarrow \mathbf{r}_0$. By taking the limit $\mathbf{r} \rightarrow \mathbf{r}_0$ in (13.7), it is concluded that $f(\mathbf{r}) \rightarrow f(\mathbf{r}_0)$. Hence, f is continuous at \mathbf{r}_0 . If the multivariable limit (13.6) exists, then it does not depend on a path approaching the limit point. In particular, take a straight line parallel to the j th coordinate axis. If $\hat{\mathbf{e}}_j$ is the unit vector parallel to this axis, then vector equation of the line is $\mathbf{r} = \mathbf{r}(t) = \mathbf{r}_0 + t\hat{\mathbf{e}}_j$. Then $\|\mathbf{r} - \mathbf{r}_0\| = |t| \rightarrow 0$ as $t \rightarrow 0$ along the line, and

$$\begin{aligned} f(\mathbf{r}(t)) - L(\mathbf{r}(t)) &= f(\mathbf{r}_0 + t\hat{\mathbf{e}}_j) - f(\mathbf{r}_0) - \mathbf{n} \cdot \hat{\mathbf{e}}_j t \\ &= f(\mathbf{r}_0 + t\hat{\mathbf{e}}_j) - f(\mathbf{r}_0) - n_j t, \end{aligned}$$

where n_j is the j th component of the vector \mathbf{n} . By the same reasoning as in the one-variable case with $\Delta x = t$ (given after Definition 13.17), the condition (13.6) implies

$$\lim_{t \rightarrow 0} \frac{f(\mathbf{r}_0 + t\hat{\mathbf{e}}_j) - f(\mathbf{r}_0)}{t} - n_j = 0 \quad \Leftrightarrow \quad n_j = f'_{x_j}(\mathbf{r}_0)$$

according to Definition 13.16 of partial derivatives at a point. The existence of partial derivatives is guaranteed by the existence of the limit (13.6). \square

The following important remarks are in order. In contrast to the one-variable case, *the existence of partial derivatives at a point does not generally imply continuity at that point.*

EXAMPLE 13.24. *Consider the function*

$$f(x, y) = \begin{cases} \frac{xy}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}.$$

Show that this function is not continuous at $(0, 0)$, but that the partial derivatives $f'_x(0, 0)$ and $f'_y(0, 0)$ exist.

SOLUTION: In order to check the continuity, one has to calculate the limit $\lim_{(x,y) \rightarrow (0,0)} f(x, y)$. If it exists and equals $f(0, 0) = 0$, then the function is continuous at $(0, 0)$. This limit does not exist. Along lines $(x, y) = (t, at)$, the function has constant value $f(t, at) = at^2/(t^2 + a^2t^2) = a/(1 + a^2)$ and hence does not approach $f(0, 0) = 0$ as $t \rightarrow 0^+$. To find the partial derivatives in question, note 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 $f'_y(0, y) = 0$. In particular, the partial derivatives exist at the origin, $f'_x(0, 0) = f'_y(0, 0) = 0$. \square

This example shows that *both the continuity of a function and the existence of its partial derivatives at a point are necessary conditions for differentiability of the function at that point*. In contrast to the one-variable case, they are not sufficient; that is, the converse of Theorem 13.11 is false. A good linear approximation in the sense of (13.6) (or (13.7)) may not exist even if a function is continuous and has partial derivatives at a point.

EXAMPLE 13.25. Let

$$f(x, y) = \begin{cases} \frac{xy}{\sqrt{x^2+y^2}} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}.$$

Show that f is continuous at $(0, 0)$ and has the partial derivatives $f'_x(0, 0)$ and $f'_y(0, 0)$, but it is not differentiable at $(0, 0)$.

SOLUTION: The continuity is verified by the squeeze principle. Put $r = \sqrt{x^2 + y^2}$. Then $|xy| = |x||y| \leq r^2$. Therefore, $|f(x, y)| \leq r^2/r = r \rightarrow 0$ as $r \rightarrow 0$, which means that $\lim_{(x,y) \rightarrow (0,0)} f(x, y) = 0 = f(0, 0)$, and hence f is continuous at $(0, 0)$. The partial derivatives are found in the same fashion as in Example 13.24. Since $f(x, 0) = 0$ for all x , $f'_x(x, 0) = 0$. Similarly, $f'_y(0, y) = 0$ follows from $f(0, y) = 0$ for all y . In particular, $f'_x(0, 0) = f'_y(0, 0) = 0$. The continuity of f and the existence of its partial derivatives at $(0, 0)$ suggest that if a linear approximation with the property (13.6) exists, then $L(x, y) = 0$ (Theorem 13.11). However, $L(x, y) = 0$ does not satisfy (13.6). Indeed, in this case, $[f(x, y) - L(x, y)]/\|\mathbf{r}\| = f(x, y)/\|\mathbf{r}\|$ is the function from Example 13.24, which is not continuous at $(0, 0)$, and the limit (13.6) does not even exist. So the function is not differentiable at $(0, 0)$. \square

The following theorem establishes a *sufficient* condition for differentiability (its proof is omitted).

THEOREM 13.12. (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 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 in a *neighborhood* of a point, then the function is differentiable at that point and there exists a good linear approximation in the sense of (13.6).

EXAMPLE 13.26. Find the region in which the function $e^{xz} \cos(yz)$ is differentiable.

SOLUTION: The function e^{xz} is the composition of the exponential e^u and the polynomial $u = xz$. So its partial derivatives are continuous everywhere. Similarly, the partial derivatives of $\cos(yz)$ are also continuous everywhere. By the product rule for partial derivatives, the partial derivatives of $e^{xz} \cos(yz)$ are continuous everywhere. By Theorem 13.12, the function is differentiable everywhere and a good linear approximation exists everywhere. \square

Remark. Theorem 13.12 provides only a sufficient condition for differentiability. There are differentiable functions at a point whose partial derivatives exist but are not continuous at that point. An example is discussed in Study Problem 13.7.

90.3. Tangent Plane Approximation. The concept of differentiability is important for approximations. Only differentiable functions have a good linear approximation. Owing to the uniqueness of a good linear approximation, it is convenient to give it a name.

DEFINITION 13.18. (Linearization of a Multivariable Function).

Let f be a function of m variables $\mathbf{r} = (x_1, x_2, \dots, x_m)$ on D that is differentiable at an interior point $\mathbf{r}_0 = (a_1, a_2, \dots, a_m)$ of D . Put $n_i = f'_{x_i}(\mathbf{r}_0)$, $i = 1, 2, \dots, m$. The linear 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 .

If Δx_i denotes the deviation of x_i from a_i , then

$$(13.8) \quad L(\mathbf{r}) = f(\mathbf{r}_0) + n_1 \Delta x_1 + n_2 \Delta x_2 + \cdots + n_m \Delta x_m, \quad n_i = f'_{x_i}(\mathbf{r}_0).$$

Geometrical Significance of Linearization. Consider the graph $z = f(x, y)$ of a continuous two-variable function. Suppose that f has continuous partial derivatives at a point (x_0, y_0) . Consider the curve of intersection

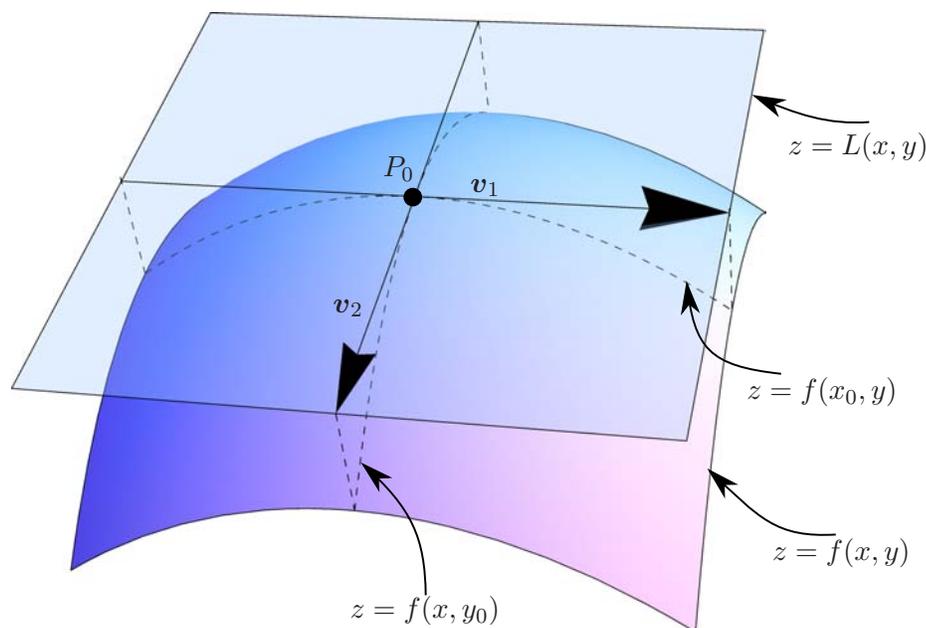


FIGURE 13.7. The tangent plane to the graph $z = f(x, y)$ at the point $P_0 = (x_0, y_0, f(x_0, y_0))$. The curves $z = f(x_0, y)$ and $z = f(x, y_0)$ are the cross sections of the graph by the coordinate planes $x = x_0$ and $y = y_0$, respectively. The vectors \mathbf{v}_1 and \mathbf{v}_2 are tangent to the curves of the cross sections at the point P_0 . The plane through P_0 and parallel to these vectors is the tangent plane to the graph. Its normal is $\mathbf{n} = \mathbf{v}_1 \times \mathbf{v}_2$.

of the graph with the coordinate plane $x = x_0$. Its equation is $z = f(x_0, y)$. Then 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 $\mathbf{v}_1 = \mathbf{r}'(y_0) = (0, 1, f'_y(x_0, y_0))$ (see Figure 13.7). 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

$$z = z_0 + n_1(x - x_0) + n_2(y - y_0), \quad n_1 = f'_x(x_0, y_0), \quad n_2 = f'_y(x_0, y_0).$$

Thus, the graph $z = L(x, y)$ of the linearization L of a differentiable function f at (x_0, y_0) defines the tangent plane to the graph of f at (x_0, y_0) . For this reason, the linearization of f at a point is also called the *tangent plane approximation*; it is a good linear approximation in the sense of (13.6). The tangent plane approximation (or the linearization in general) is a multivariable analog of the tangent line approximation for single-variable functions.

EXAMPLE 13.27. 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 polynomial function $f(x, y) = x^2 + 3y^2$ that is continuous and has continuous partial derivatives at any point, and hence f is differentiable. The components of a normal of the tangent plane 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 $4x + 6y - z = 7$. \square

EXAMPLE 13.28. Use the linearization to estimate the number $[(2.03)^2 + (1.97)^2 + (0.94)^2]^{1/2}$.

SOLUTION: Let $f(x, y, z) = [x^2 + y^2 + z^2]^{1/2}$. It is continuous and has continuous partial derivatives everywhere except the origin because it is a composition of the polynomial $g = x^2 + y^2 + z^2$ and the power function: $f = (g)^{1/2}$. 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 $f(\mathbf{r}_0) = 3$. Since f is differentiable at \mathbf{r}_0 , its linearization can be used to approximate values of f near \mathbf{r}_0 . 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 (see (13.8)) gives

$$f(x, y, z) \approx L(x, y, z) = 3 + (2/3)\Delta x + (2/3)\Delta y + (1/3)\Delta z = 2.98.$$

\square

90.4. Study Problems.

Problem 13.7. *Let*

$$f(x, y) = \begin{cases} (x^2 + y^2) \sin\left(\frac{1}{x^2 + y^2}\right) & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}.$$

“Show that f is differentiable at $(0, 0)$ (and hence that $f'_x(0, 0)$ and $f'_y(0, 0)$ exist), but that f'_x and f'_y are not continuous at $(0, 0)$.”

SOLUTION: By the definition of partial derivatives,

$$f'_x(0, 0) = \lim_{h \rightarrow 0} \frac{f(h, 0) - f(0, 0)}{h} = \lim_{h \rightarrow 0} h \sin(1/h^2) = 0,$$

which follows from the squeeze principle: $0 \leq |h \sin(1/h^2)| \leq |h| \rightarrow 0$. Similarly, $f'_y(0, 0) = 0$. Since $f(0, 0) = 0$ and $f'_x(0, 0) = f'_y(0, 0) = 0$, one has to verify whether the linear function $L(x, y) = 0$ satisfies the condition (13.6):

$$\lim_{(x, y) \rightarrow (0, 0)} \frac{f(x, y) - L(x, y)}{\sqrt{x^2 + y^2}} = \lim_{r \rightarrow 0^+} r \sin(1/r^2) = 0$$

by the squeeze principle, $0 \leq |r \sin(1/r^2)| \leq r \rightarrow 0$ as $r \rightarrow 0^+$. Thus, $L(x, y) = 0$ is a good linear approximation, and the function f is differentiable at the origin.

For $(x, y) \neq (0, 0)$,

$$f'_x(x, y) = 2x \sin\left(\frac{1}{x^2 + y^2}\right) - \frac{2x}{x^2 + y^2} \cos\left(\frac{1}{x^2 + y^2}\right).$$

The first term in this expression converges to 0, $0 \leq |2x \sin(1/r^2)| \leq 2|x| \rightarrow 0$ as $(x, y) \rightarrow (0, 0)$, whereas the second term can take arbitrarily large values in any neighborhood of the origin. To see this, consider a sequence of points that converges to the origin: $(x, y) = (t_n, 0)$, $n = 0, 1, \dots$, where $1/t_n^2 = \pi/2 + \pi n$ or $t_n = 1/(\pi/2 + \pi n)^{1/2} \rightarrow 0$ as $n \rightarrow \infty$. Then $\cos(1/t_n^2) = (-1)^n$, $\sin(1/t_n^2) = 0$, and $f'_x(t_n, 0) = 2t_n \sin(1/t_n^2) + (2/t_n) \cos(1/t_n^2) = 2(-1)^n(\pi/2 + \pi n)^{1/2}$. So $f'_x(t_n, 0)$ can take arbitrarily large positive and negative values as $n \rightarrow \infty$, and the limit $\lim_{(x, y) \rightarrow (0, 0)} f'_x(x, y)$ does not exist, which means that the partial derivative $f'_x(x, y)$ is not continuous at the origin. Owing to the symmetry $f(x, y) = f(y, x)$, the same conclusion holds for $f'_y(x, y)$. \square

90.5. Exercises.

(1) Let $f(x, y) = xy^2$ if $(x, y) \neq (0, 0)$ and $f(0, 0) = 1$ and let $g(x, y) = \sqrt{|x|} + \sqrt{|y|}$. Are the functions f and g differentiable at $(0, 0)$?

(2) Let $f(x, y) = xy^2/(x^2 + y^2)$ if $x^2 + y^2 \neq 0$ and $f(0, 0) = 0$. Show that f is continuous and has bounded partial derivatives f'_x and f'_y , but it is not differentiable at $(0, 0)$. Investigate the continuity of the partial derivatives near $(0, 0)$.

(3) Show that the function $f(x, y) = \sqrt{|xy|}$ is continuous at $(0, 0)$ and has the partial derivatives $f'_x(0, 0)$ and $f'_y(0, 0)$, but it is not differentiable at $(0, 0)$. Investigate the continuity of the partial derivatives f'_x and f'_y near the origin.

(4) Let $f(x, y) = x^3/(x^2 + y^2)$ if $(x, y) \neq (0, 0)$ and $f(0, 0) = 0$. Show that f is continuous, has partial derivatives at $(0, 0)$, but is not differentiable at $(0, 0)$.

(5) Find the domain in which the following functions are differentiable:

(i) $f(x, y) = y\sqrt{x}$

(ii) $f(x, y) = \frac{xy}{2x+y}$

(iii) $f(x, y, z) = \sin(xy + z)e^{zy}$

(iv) $f(x, y, z) = \sqrt{x^2 + y^2 - z^2}$

(v) $f(\mathbf{r}) = \ln(1 - \|\mathbf{r}\|)$, where $\mathbf{r} = (x_1, x_2, \dots, x_m)$

(vi) $f(x, y) = \sqrt[3]{x^3 + y^3}$

(vii) $f(\mathbf{r}) = e^{-1/\|\mathbf{r}\|^2}$ if $\mathbf{r} \neq \mathbf{0}$ and $f(\mathbf{0}) = 0$, where $\mathbf{r} = (x_1, x_2, \dots, x_m)$

Hint: Show $f'_{x_j}(\mathbf{0}) = 0$. Investigate the continuity of the partial derivatives.

(6) The line through a point P_0 of a surface perpendicular to the tangent plane at P_0 is called the *normal line*. Find an equation of the tangent plane and symmetric equations of the normal line to each of the following surfaces at the specified point:

(i) $z = x^2 + 3y - y^4x$, $(1, 2, -1)$

(ii) $z = \sqrt{x^3y}$, $(1, 4, 2)$

(iii) $z = y \ln(x^2 - 3y)$, $(2, 1, 0)$

(iv) $y = \tan^{-1}(xz^2)$, $(1, \frac{\pi}{4}, -1)$

(v) $x = z \cos(y - z)$, $(1, 1, 1)$

(vi) $z = y + \ln(x/z)$, $(1, 1, 1)$

(7) Find the linearization of each of the following functions at the specified point:

(i) $f(x, y) = \frac{2y+3}{4x+1}$, $(0, 0)$

(ii) $f(x, y, z) = z^{1/3}\sqrt{x + \cos^2(y)}$, $(0, 0, 1)$

(iii) $f(\mathbf{r}) = \sin(\mathbf{n} \cdot \mathbf{r})$, $\mathbf{r} = \mathbf{r}_0$, where \mathbf{n} is a fixed vector orthogonal to \mathbf{r}_0 and $\mathbf{r} = (x_1, x_2, \dots, x_m)$

(8) Use the linearization to approximate the following numbers. Then use a calculator to find the numbers. Compare the results.

- (i) $\sqrt{20 - 7x^2 - x^2}$, where $(x, y) = (1.08, 1.95)$
(ii) xy^2z^3 , where $(x, y, z) = (1.002, 2.003, 3.004)$
(iii) $\frac{(1.03)^2}{\sqrt[3]{0.98} \sqrt[4]{(1.05)^2}}$
(iv) $(0.97)^{1.05}$

(9) Consider the equation $f(x, y, z) = 0$ that has a root $z = z(x, y)$ for every fixed pair (x, y) . Suppose that $f(x_0, y_0, z_0) = 0$ and f is differentiable at (x_0, y_0, z_0) so that $f'_z(x_0, y_0, z_0) \neq 0$. If $L(x, y, z)$ is the linearization of f at (x_0, y_0, z_0) , the equation $L(x, y, z) = 0$ is called a linearization of the equation $f(x, y, z) = 0$. Its solution determines an approximation to the root $z = z(x, y)$ near (x_0, y_0) . Find this approximation, and use the result to solve the equation $yz \ln(1 + xz) - x \ln(1 + zy) = 0$ for $z = z(x, y)$ near the point $(1, 1, 1)$. In particular, estimate the root z at $(x, y) = (0.8, 1.1)$.

(10) Suppose that a function $f(x, y)$ is continuous with respect to x at each fixed y and has a bounded partial derivative $f'_y(x, y)$, that is, $|f'_y(x, y)| \leq M$ for some $M > 0$ and all (x, y) . Prove that f is continuous.

91. Chain Rules and Implicit Differentiation

91.1. 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 $f(r \cos \theta, r \sin \theta) = r^3 \sin^2 \theta = F(r, \theta)$. One can compute the rates of change of f with respect to the new variables:

$$\frac{\partial f}{\partial r} = \frac{\partial F}{\partial r} = 3r^2 \sin^2 \theta, \quad \frac{\partial f}{\partial \theta} = \frac{\partial F}{\partial \theta} = r^3 \cos \theta.$$

Alternatively, these rates can be computed as

$$\begin{aligned} \frac{\partial f}{\partial r} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial r} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial r} = (3x^2 + y^2) \cos \theta + 2xy \sin \theta = 3r^2 \sin^2 \theta, \\ \frac{\partial f}{\partial \theta} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial \theta} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial \theta} = -(3x^2 + y^2)r \sin \theta + 2xyr \cos \theta = r^3 \cos \theta, \end{aligned}$$

where x and y have been expressed in the polar coordinates to obtain the final expressions. The latter relations are an example of a *chain rule* for functions of two variables. Suppose that the rates $f'_x(x_0, y_0)$ and $f'_y(x_0, y_0)$ are known at a particular point (x_0, y_0) . Then, by using

the chain rule, *an explicit form of the function f in the new variables is not required to find its rates with respect to the new variables* because the rates x'_r , x'_θ , y'_r , and y'_θ at (r_0, θ_0) corresponding to (x_0, y_0) can easily be computed.

On the other hand, consider the function $f(x, y) = y^3/(x^2 + y^2)$ if $(x, y) \neq (0, 0)$ and $f(0, 0) = 0$. This function is continuous at the origin (if $R^2 = x^2 + y^2$, then $|f(x, y) - f(0, 0)| \leq |y^3|/R^2 \leq R^3/R^2 = R \rightarrow 0$ as $R \rightarrow 0$). It has partial derivatives at the origin. Indeed, $f(x, 0) = 0$, and hence $f'_x(x, 0) = 0$ for any x and, in particular, $f'_x(0, 0) = 0$. Similarly, $f(0, y) = y$, and hence $f'_y(0, y) = 1$ so that $f'_y(0, 0) = 1$. Let $x = t \cos \theta$ and $y = t \sin \theta$, where θ is a numerical parameter. Then $F(t) = f(t \cos \theta, t \sin \theta) = t^3 \sin^3 \theta / t^2 = t \sin^3 \theta$. Therefore, $F'(t) = \sin^3 \theta$. This implies that $F'(0) = \sin^3 \theta$. However, the chain rule fails: $F'(0) = df/dt|_{t=0} = f'_x(0, 0)x'(0) + f'_y(0, 0)y'(0) = \sin \theta$. It is not difficult to verify that the chain rule $df/dt = f'_x x'(t) + f'_y y'(t)$ is true for all $t \neq 0$. The reader is advised to verify that the function is not differentiable at $(0, 0)$ (see Study Problem 13.12), which is the reason for the chain rule is not valid at that point. It appears that, in contrast to the one-variable case, the mere existence of partial derivatives is not sufficient to validate the chain rule in the multi-variable case, and a stronger condition of f is required.

THEOREM 13.13. (Chain Rule).

Let f be a function of n variables $\mathbf{r} = (x_1, x_2, \dots, x_n)$. Suppose that each variable x_i is, in turn, a function of m variables $\mathbf{u} = (u_1, u_2, \dots, u_m)$. The composition of $x_i = x_i(\mathbf{u})$ with $f(\mathbf{r})$ defines f as a function of \mathbf{u} . If the functions x_i are differentiable at a point \mathbf{u} and f is differentiable at the point $\mathbf{r} = (x_1(\mathbf{u}), x_2(\mathbf{u}), \dots, x_n(\mathbf{u}))$, then the rate of change of f with respect to u_j , $j = 1, 2, \dots, m$, reads

$$\frac{\partial f}{\partial u_j} = \frac{\partial f}{\partial x_1} \frac{\partial x_1}{\partial u_j} + \frac{\partial f}{\partial x_2} \frac{\partial x_2}{\partial u_j} + \cdots + \frac{\partial f}{\partial x_n} \frac{\partial x_n}{\partial u_j} = \sum_{i=1}^n \frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial u_j}.$$

PROOF. Since the functions $x_i(\mathbf{u})$ are differentiable, the partial derivatives $\partial x_i / \partial u_j$ exist and define a good linear approximation in the sense (13.7). In particular, for a fixed value of \mathbf{u} and for every i ,

$$\Delta x_i = x_i(\mathbf{u} + \hat{\mathbf{e}}_j h) - x_i(\mathbf{u}) = \frac{\partial x_i}{\partial u_j} h + \varepsilon_i(h) |h|, \quad \varepsilon_i(h) \rightarrow 0 \quad \text{as } h \rightarrow 0.$$

Define the vector $\Delta \mathbf{r}_h = (\Delta x_1, \Delta x_2, \dots, \Delta x_n)$. It has the property that $\Delta \mathbf{r}_h \rightarrow \mathbf{0}$ as $h \rightarrow 0$. If $F(\mathbf{u}) = f(x_1(\mathbf{u}), x_2(\mathbf{u}), \dots, x_n(\mathbf{u}))$, then, by the

definition of the partial derivatives at a point \mathbf{u} ,

$$\frac{\partial f}{\partial u_j} = \lim_{h \rightarrow 0} \frac{F(\mathbf{u} + \hat{\mathbf{e}}_j h) - F(\mathbf{u})}{h} = \lim_{h \rightarrow 0} \frac{f(\mathbf{r} + \Delta \mathbf{r}_h) - f(\mathbf{r})}{h}$$

if the limit exists. By the hypothesis, the function f is differentiable and hence has partial derivatives $\partial f / \partial x_i$ at the point $\mathbf{r} = (x_1(\mathbf{u}), x_2(\mathbf{u}), \dots, x_n(\mathbf{u}))$ that determine a good linear approximation (13.8) in the sense of (13.7):

$$f(\mathbf{r} + \Delta \mathbf{r}_h) - f(\mathbf{r}) = \frac{\partial f}{\partial x_1} \Delta x_1 + \frac{\partial f}{\partial x_2} \Delta x_2 + \cdots + \frac{\partial f}{\partial x_n} \Delta x_n + \varepsilon(\Delta \mathbf{r}_h) \|\Delta \mathbf{r}_h\|,$$

where $\varepsilon(\Delta \mathbf{r}_h) \rightarrow 0$ as $\Delta \mathbf{r}_h \rightarrow \mathbf{0}$ or as $h \rightarrow 0$. The substitution of this relation into the limit shows that the limit exists, and the conclusion of the theorem follows. Indeed, the first n terms contain the limits

$$\lim_{h \rightarrow 0} \frac{\Delta x_i}{h} = \frac{\partial x_i}{\partial u_j} + \lim_{h \rightarrow 0} \varepsilon_i(h) \frac{|h|}{h} = \frac{\partial x_i}{\partial u_j}$$

because $|h|/h = \pm 1$ for all $h \neq 0$ and $\varepsilon_i(h) \rightarrow 0$ as $h \rightarrow 0$. The ratio $\|\Delta \mathbf{r}_h\|/|h| = [(\Delta x_1/h)^2 + (\Delta x_2/h)^2 + \cdots + (\Delta x_n/h)^2]^{1/2} \rightarrow M < \infty$ as $h \rightarrow 0$, where M is determined by the partial derivatives $\partial x_i / \partial u_j$. Therefore, the limit of the last term vanishes:

$$\begin{aligned} \lim_{h \rightarrow 0} \varepsilon(\Delta \mathbf{r}_h) \frac{\|\Delta \mathbf{r}_h\|}{h} &= \lim_{h \rightarrow 0} \varepsilon(\Delta \mathbf{r}_h) \frac{|h|}{h} \frac{\|\Delta \mathbf{r}_h\|}{|h|} \\ &= \lim_{h \rightarrow 0} \varepsilon(\Delta \mathbf{r}_h) \frac{|h|}{h} \cdot \lim_{h \rightarrow 0} \frac{\|\Delta \mathbf{r}_h\|}{|h|} = 0 \cdot M = 0 \end{aligned}$$

because $\varepsilon(\Delta \mathbf{r}_h)|h|/h = \pm \varepsilon(\Delta \mathbf{r}_h)$ if $h \neq 0$ and $\varepsilon(\Delta \mathbf{r}_h) \rightarrow 0$ as $h \rightarrow 0$. \square

It is clear from the proof that the partial derivatives $\partial f / \partial x_i$ in the chain rule are taken at the point $(x_1(\mathbf{u}), x_2(\mathbf{u}), \dots, x_n(\mathbf{u}))$. 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 $\mathbf{u} = (r, \theta)$.

EXAMPLE 13.29. *Let a function $f(x, y, z)$ be differentiable at $\mathbf{r}_0 = (1, 2, 3)$ and have the following rates of change: $f'_x(\mathbf{r}_0) = 1$, $f'_y(\mathbf{r}_0) = 2$, 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 $z'_s = 3$. They are continuous functions and hence $x(t, s)$, $y(t, s)$, and $z(t, s)$ are differentiable by Theorem 13.12. By the chain rule,

$$\begin{aligned} f'_t(\mathbf{r}_0) &= f'_x(\mathbf{r}_0)x'_t(\mathbf{u}_0) + f'_y(\mathbf{r}_0)y'_t(\mathbf{u}_0) + f'_z(\mathbf{r}_0)z'_t(\mathbf{u}_0) \\ &= 1 \cdot 2 + 2 \cdot 1 + (-2) \cdot 0 = 4, \\ f'_s(\mathbf{r}_0) &= f'_x(\mathbf{r}_0)x'_s(\mathbf{u}_0) + f'_y(\mathbf{r}_0)y'_s(\mathbf{u}_0) + f'_z(\mathbf{r}_0)z'_s(\mathbf{u}_0) \\ &= 1 \cdot 1 + 2 \cdot 1 + (-2) \cdot 3 = -3. \end{aligned}$$

□

EXAMPLE 13.30. Let $f(x, y, z) = z^2(1 + x^2 + 2y^2)^{-1}$. Find the rate of change of f along the curve $\mathbf{r}(t) = (\sin t, \cos t, e^t)$ in the direction of increasing t .

SOLUTION: The function f is differentiable as the ratio of two polynomials (its partial derivatives are continuous):

$$f'_x = -\frac{2xz^2}{(1 + x^2 + 2y^2)^2}, \quad f'_y = -\frac{4yz^2}{(1 + x^2 + 2y^2)^2}, \quad f'_z = \frac{2z}{1 + x^2 + 4y^2}.$$

The components of $\mathbf{r}(t)$ are also differentiable: $x'(t) = \cos t$, $y'(t) = -\sin t$, $z'(t) = e^t$. By the chain rule for $n = 3$ and $m = 1$,

$$\begin{aligned} \frac{df}{dt} &= f'_x(\mathbf{r}(t))x'(t) + f'_y(\mathbf{r}(t))y'(t) + f'_z(\mathbf{r}(t))z'(t) \\ &= -\frac{2e^{2t} \sin t}{(2 + \cos^2 t)^2} \cdot (\cos t) - \frac{4e^{2t} \cos t}{(2 + \cos^2 t)^2} \cdot (-\sin t) + \frac{2e^t}{2 + \cos^2 t} \cdot e^t \\ &= \frac{e^{2t}(5 + \sin(2t) + \cos(2t))}{(2 + \cos^2 t)^2}, \end{aligned}$$

where $2 \sin t \cos t = \sin(2t)$ and $2 \cos^2 t = 1 + \cos(2t)$ have been used. □

The chain rule can be used to calculate higher-order partial derivatives.

EXAMPLE 13.31. If $g(u, v) = f(x, y)$, where $x = (u^2 - v^2)/2$ and $y = uv$, find g''_{uv} . Assume that f has continuous second partial derivatives. If $f'_y(1, 2) = 1$, $f''_{xx}(1, 2) = f''_{yy}(1, 2) = 2$, and $f''_{xy}(1, 2) = 3$, find the value of g''_{uv} at $(x, y) = (1, 2)$.

SOLUTION: One has $x'_u = u$, $x'_v = -v$, $y'_u = v$, and $y'_v = u$. Then

$$g'_u = f'_x x'_u + f'_y y'_u = f'_x u + f'_y v.$$

The derivative $g''_{uv} = (g'_u)'_v$ is calculated by applying the chain rule to the function g'_u :

$$\begin{aligned} g''_{uv} &= u(f'_x)'_v + v(f'_y)'_v + f'_y \\ &= u(f''_{xx}x'_v + f''_{xy}y'_v) + v(f''_{yx}x'_v + f''_{yy}y'_v) + f'_y \\ &= u(-vf''_{xx} + uf''_{xy}) + v(-vf''_{yx} + uf''_{yy}) + f'_y \\ &= uv(f''_{yy} - f''_{xx}) + (u^2 - v^2)f''_{xy} + f'_y = y(f''_{yy} - f''_{xx}) + 2xf''_{xy} + f'_y, \end{aligned}$$

where $f''_{xy} = f''_{yx}$ has been used. The value of g''_{uv} at the point in question is $2 \cdot (2 - 2) + 2 \cdot 3 + 1 = 7$. \square

91.2. 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:

$$F(x, y, z) = 0 \implies z = z(x, y) = x^2 + y^4;$$

that is, the function $z(x, y)$ is defined as a root of $F(x, y, z)$ and has the characteristic property that

$$(13.9) \quad F(x, y, z(x, y)) = 0 \quad \text{for all } (x, y).$$

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 (13.9) 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 F is differentiable. Furthermore, the root $z(x, y)$ is also assumed to be differentiable on an open disk D in the plane. Since relation (13.9) holds for all $(x, y) \in D$, the partial derivatives of its left side must also vanish in D . They 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 $x'_u = 1$, $x'_v = 0$, $y'_u = 0$, $y'_v = 1$, and $z'_u(u, v) = z'_x(x, y)$ and $z'_v(u, v) = z'_y(x, y)$ because $x = u$

and $y = v$. Therefore,

$$\frac{\partial}{\partial u} F(x, y, z(x, y)) = \frac{\partial F}{\partial x} + \frac{\partial F}{\partial z} \frac{\partial z}{\partial x} = 0 \quad \Longrightarrow \quad z'_x = -\frac{F'_x}{F'_z},$$

$$\frac{\partial}{\partial v} F(x, y, z(x, y)) = \frac{\partial F}{\partial y} + \frac{\partial F}{\partial z} \frac{\partial z}{\partial y} = 0 \quad \Longrightarrow \quad z'_y = -\frac{F'_y}{F'_z}.$$

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 differentiability of $z(x, y)$ have been left unanswered in the above analysis. The following theorem addresses them all.

THEOREM 13.14. (Implicit Function Theorem).

Let F be a function of $n+1$ variables, $F(\mathbf{r}, z)$, where $\mathbf{r} = (x_1, x_2, \dots, x_n)$ and z is real such that F and F'_z 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 $F'_z(\mathbf{r}_0, z_0) \neq 0$. There exists an open neighborhood D of \mathbf{r}_0 , an open interval I , and a unique function $z : D \rightarrow I$ such that for $(\mathbf{r}, y) \in D \times I$, $F(\mathbf{r}, y) = 0$ if and only if $y = z(\mathbf{r})$. Moreover, the function z is continuous. If, in addition, F is differentiable in B , then the function $z = z(\mathbf{r})$ is differentiable in D and

$$z'_{x_i}(\mathbf{r}) = -\frac{F'_{x_i}(\mathbf{r}, z(\mathbf{r}))}{F'_z(\mathbf{r}, z(\mathbf{r}))}$$

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 its differentiability. Once these facts are established, a derivation of the implicit differentiation formula follows the same way as in the $n = 2$ case:

$$\frac{\partial F}{\partial x_i} + \frac{\partial F}{\partial z} \frac{\partial z}{\partial x_i} = 0 \quad \Longrightarrow \quad z'_{x_i}(\mathbf{r}) = -\frac{F'_{x_i}(\mathbf{r}, z(\mathbf{r}))}{F'_z(\mathbf{r}, z(\mathbf{r}))}.$$

Remark. If the function F has sufficiently many continuous higher-order partial derivatives, then higher order partial derivatives of $z(\mathbf{r})$ can be obtained by differentiation of these relations. An example is given in Study Problem 13.9.

EXAMPLE 13.32. 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 partial derivatives

$$\begin{aligned} F'_x &= \pi yz \cos(xyz) - 3z, & F'_y &= \pi xz \cos(xyz) + z, \\ F'_z &= \pi xy \cos(xyz) - 3x + y \end{aligned}$$

are continuous for all (x, y, z) ; hence, F is differentiable everywhere. 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,

$$z'_x(1, 1) = -\frac{F'_x(1, 1, \pi/2)}{F'_z(1, 1, \pi/2)} = -\frac{3\pi}{4}, \quad z'_y(1, 1) = -\frac{F'_y(1, 1, \pi/2)}{F'_z(1, 1, \pi/2)} = \frac{\pi}{4}.$$

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 linearizing the function $z(x, y)$ near $(1, 1)$. The continuity of partial derivatives ensures that $z(x, y)$ is differentiable at $(1, 1)$ and has a good linear approximation in the sense of (13.6) (see Study Problem 13.8). \square

91.3. Study Problems.

Problem 13.8. 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

$$\begin{aligned} z(1 + \Delta x, 1 + \Delta y) &\approx z(1, 1) + z'_x(1, 1) \Delta x + z'_y(1, 1) \Delta y \\ &= \frac{\pi}{2} \left(1 - \frac{3\Delta x}{2} + \frac{\Delta y}{2} \right). \end{aligned}$$

Putting $\Delta x = 0.04$ and $\Delta y = -0.04$, this equation yields the estimate $z(1.04, 0.96) \approx 0.45\pi$. \square

Problem 13.9. Let the function $z(x, y)$ be defined implicitly by $z^5 + zx - y = 0$ in a neighborhood of $(1, 2, 1)$. Find all its first and second partial derivatives. In particular, give the values of these partial derivatives at $(x, y) = (1, 2)$.

SOLUTION: Let $F(x, y, z) = z^5 + zx - y$. Then $F'_z = 5z^4 + x$. The function $z(x, y)$ exists in a neighborhood of $(1, 2)$ by the implicit function theorem because $F(1, 2, 1) = 0$ and $F'_z(1, 2, 1) = 6 \neq 0$. The first and second partial derivatives of F are continuous everywhere:

$$\begin{aligned} F'_x &= z, & F'_y &= -1, & F'_z &= 5z^4 + x, \\ F''_{xx} &= 0, & F''_{xy} &= 0, & F''_{xz} &= 1, \\ F''_{yy} &= 0, & F''_{yz} &= 0, & F''_{zz} &= 20z^3. \end{aligned}$$

By implicit differentiation,

$$z'_x = -\frac{F'_x}{F'_z} = -\frac{z}{5z^4 + x}, \quad z'_y = -\frac{F'_y}{F'_z} = \frac{1}{5z^4 + x}.$$

Taking the partial derivatives of these relations with respect to x and y and using the quotient rule for differentiation, the second partial derivatives are obtained:

$$\begin{aligned} z''_{xx} &= -\frac{z'_x(5z^4 + x) - z(20z^3z'_x + 1)}{(5z^4 + x)^2} = \frac{(15z^4 - x)z'_x + z}{(5z^4 + x)^2}, \\ z''_{xy} &= z''_{yx} = (z'_y)'_x = -\frac{20z^3z'_x + 1}{(5z^4 + x)^2}, \quad z''_{yy} = -\frac{20z^3z'_y}{(5z^4 + x)^2}. \end{aligned}$$

The explicit form of z'_x and z'_y may be substituted into these relations to express the second partial derivatives via x , y , and z . At the point $(1, 2)$, the values of the first partial derivatives are $z'_x(1, 2) = -1/6$ and $z'_y(1, 2) = 1/6$. Using these values, the values of the second partial derivatives are evaluated: $z''_{xx}(1, 2) = -1/27$, $z''_{xy}(1, 2) = 7/108$, and $z''_{yy}(1, 2) = -5/54$. \square

91.4. 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 $x = 2/u - v + w$, $y = vuv$, $z = e^w$.
- (6) If $z(u, v) = f(x, y)$, where $x = e^u \cos v$ and $y = e^u \sin v$, show that $z''_{xx} + z''_{yy} = e^{-2s}(z''_{uu} + z''_{vv})$.
- (7) If $z(u, v) = f(x, y)$, where $x = u^2 + v^2$ and $y = 2uv$, find all the second-order partial derivatives of $z(u, v)$.
- (8) If $z(u, v) = f(x, y)$, where $x = u + v$ and $y = u - v$, show that $(z'_x)^2 + (z'_y)^2 = z'_u z'_v$.
- (9) Find all the first and second partial derivatives of the following functions:

- (i) $g(x, y, z) = f(x^2 + y^2 + z^2)$
 (ii) $g(x, y) = f(x, x/y)$
 (iii) $g(x, y, z) = f(x, xy, xyz)$
 (iv) $g(x, y) = f(x/y, y/x)$
 (v) $g(x, y, z) = f(x + y + z, x^2 + y^2 + z^2)$
 (vi) $g(x, y) = f(x + y, xy)$

- (10) Find $g''_{xx} + g''_{yy} + g''_{zz}$ if $g(x, y, z) = f(x + y + z, x^2 + y^2 + z^2)$.
- (11) Let $x = r \cos \theta$ and $y = r \sin \theta$. Show that

$$f''_{xx} + f''_{yy} = \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial f}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2}.$$

- (12) Let $x = \rho \sin \phi \cos \theta$, $y = \rho \sin \phi \sin \theta$, $z = \rho \cos \phi$. The variables (ρ, ϕ, θ) are called *spherical coordinates* and discussed in Section 104.3. Show that

$$f''_{xx} + f''_{yy} + f''_{zz} = \frac{1}{\rho^2} \frac{\partial}{\partial \rho} \left(\rho^2 \frac{\partial f}{\partial \rho} \right) + \frac{1}{\rho^2 \sin \phi} \frac{\partial}{\partial \phi} \left(\sin \phi \frac{\partial f}{\partial \phi} \right) + \frac{1}{\rho^2 \sin^2 \phi} \frac{\partial^2 f}{\partial \theta^2}.$$

- (13) Prove that if a function $f(x, y)$ satisfies the Laplace equation $f''_{xx} + f''_{yy} = 0$, then the function $g(x, y) = f(x/(x^2 + y^2), y/(x^2 + y^2))$, $x^2 + y^2 > 0$, also satisfies the Laplace equation.

- (14) Prove that if a function $f(x, t)$ satisfies the diffusion equation $f'_t = a^2 f''_{xx}$, then the function

$$g(x, t) = \frac{1}{a\sqrt{t}} e^{-x^2/(4a^2t)} f\left(\frac{x}{a^2t}, -\frac{x}{a^4t}\right), \quad t > 0,$$

also satisfies the diffusion equation.

(15) Prove that if $f(x, y, z)$ satisfies the Laplace equation $f''_{xx} + f''_{yy} + f''_{zz} = 0$, then the function

$$g(x, y, z) = \frac{1}{r} f\left(\frac{a^2x}{r^2}, \frac{a^2y}{r^2}, \frac{a^2z}{r^2}\right), \quad r = \sqrt{x^2 + y^2 + z^2} \neq 0,$$

also satisfies the Laplace equation.

(16) Show that the function $g(x, y) = x^n f(y/x^2)$, where f is a differentiable function, satisfies the equation $xg'_x + 2yg'_y = ng$.

(17) Show that the function $g(x, y, z) = x^n f(y/x^a, z/x^b)$, where f is a differentiable function, satisfies the equation $xg'_x + ayg'_y + bzg'_z = ng$.

(18) Let the function $z = f(x, y)$ be defined implicitly. Find its first and second partial derivatives if

- (i) $x + 2y + 3z = e^z$
- (ii) $x - z = \tan^{-1}(yz)$
- (iii) $x/z = \ln(z/y) + 1$

(19) Let $z(x, y)$ be a solution of the equation $z^3 - xz + y = 0$ such that $z(3, -2) = 2$. Find the linearization of $z(x, y)$ near $(3, -2)$ and use it to estimate $z(2.8, -2.3)$.

(20) Find f'_x and f'_y , where $f = (x + z)/(y + z)$ and z is defined by the equation $ze^z = xe^x + ye^y$.

(21) Show that the function $z(x, y)$ defined by the equation $F(x - az, y - bz) = 0$, where F is a differentiable function of two variables and a and b are constants, satisfies the equation $az'_x + bz'_y = 1$.

(22) Let the temperature of the air at a point (x, y, z) be $T(x, y, z)$ degrees Celsius. Suppose that T is a differentiable function. 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 $T'_z(2, 6, 8) = 1$, how fast is the temperature rising (or decreasing) on the insect's path as it flies through the point $(2, 6, 8)$?

(23) Consider a function $f = f(x, y, z)$ and the change of variables: $x = 2uv$, $y = u^2 - v^2 + w$, $z = u^3vw$. Find the partial derivatives f'_u , f'_v , and f'_w at the point $u = v = w = 1$, if $f'_x = a$, $f'_y = b$, and $f'_z = c$ at $(x, y, z) = (2, 1, 1)$.

(24) 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 instant, find the rates at which the volume, the surface area, and the largest diagonal are changing.

(25) 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 $f'_x(tx, ty) = t^{n-1}f'_x(x, y)$.

(26) 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 $(\partial z/\partial x)(\partial x/\partial y)(\partial y/\partial z) = -1$.

(27) Let $x^2 = vw$, $y^2 = uw$, $z^2 = uv$, and $f(x, y, z) = F(u, v, w)$. Show that $xf'_x + yf'_y + zf'_z = uF'_u + vF'_v + wF'_w$.

(28) Simplify $z'_x \sec x + z'_y \sec y$ if $z = \sin y + f(\sin x - \sin y)$, where f is a differentiable function.

92. The Differential and Taylor Polynomials

Just like in the one-variable case, given variables $\mathbf{r} = (x_1, x_2, \dots, x_m)$, one can introduce independent variables $d\mathbf{r} = (dx_1, dx_2, \dots, dx_m)$ that are infinitesimal variations of \mathbf{r} and also called *differentials* of \mathbf{r} .

DEFINITION 13.19. (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 .

The differential is a function of $2m$ independent variables \mathbf{r} and $d\mathbf{r}$. Consider the graph $y = f(x)$ of a function f of a single variable x (see Figure 13.8, left panel). The differential $df(x_0) = f'(x_0) dx$ at a point x_0 determines the increment of y along the tangent line $y = L(x) = f(x_0) + f'(x_0)(x - x_0)$ as x changes from x_0 to $x_0 + \Delta x$, where $\Delta x = dx$. Similarly, the differential $df(x_0, y_0)$ of a function of two variables at a point $P_0 = (x_0, y_0)$ determines the increment of $z = L(x, y)$ along the tangent plane to the graph $z = f(x, y)$ at the point $(x_0, y_0, f(x_0, y_0))$ when (x, y) changes from (x_0, y_0) to $(x_0 + \Delta x, y_0 + \Delta y)$, where $dx = \Delta x$ and $dy = \Delta y$, as depicted in the right panel of Figure 13.8. In general, the differential $df(\mathbf{r}_0)$ and the linearization of f at a point \mathbf{r}_0 are related as

$$L(\mathbf{r}) = f(\mathbf{r}_0) + df(\mathbf{r}_0), \quad dx_i = \Delta x_i, \quad i = 1, 2, \dots, m;$$

that is, if the infinitesimal variations (or differentials) $d\mathbf{r}$ are replaced by 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 . According to (13.6), the difference $f(\mathbf{r}) - f(\mathbf{r}_0) - df(\mathbf{r}_0)$ tends to 0 faster than $\|\Delta\mathbf{r}\|$ as $\Delta\mathbf{r} \rightarrow \mathbf{0}$, and hence the differential can be used to study variations of a differentiable function f under small variations of its arguments.

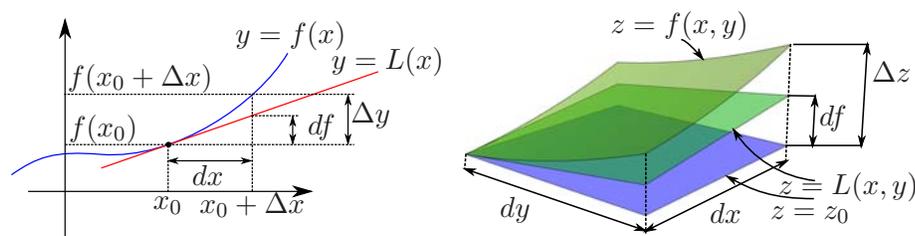


FIGURE 13.8. Geometrical significance of the differential.

Left: The differential of a function of one variable. It defines the increment of y along the tangent line $y = L(x)$ to the graph $y = f(x)$ at (x_0, y_0) , $y_0 = f(x_0)$, when x changes from x_0 to $x_0 + \Delta x$, where $dx = \Delta x$. As $\Delta x \rightarrow 0$, $(\Delta y - df)/\Delta x = \Delta y/\Delta x - f'(x_0) \rightarrow 0$; that is, the difference $\Delta y - df$ tends to 0 faster than Δx . **Right:** The differential of a function of two variables. It defines the increment of z along the tangent plane $z = L(x, y)$ to the graph $z = f(x, y)$ at (x_0, y_0, z_0) , $z_0 = f(x_0, y_0)$, when (x, y) changes from (x_0, y_0) to $(x_0 + \Delta x, y_0 + \Delta y)$, where $dx = \Delta x$ and $dy = \Delta y$. The difference $\Delta z - df$ tends to 0 faster than $\|\Delta \mathbf{r}\| = \sqrt{(\Delta x)^2 + (\Delta y)^2}$ as $\|\Delta \mathbf{r}\| \rightarrow 0$.

EXAMPLE 13.33. Find $df(x, y)$ if $f(x, y) = \sqrt{1 + x^2 y}$. In particular, evaluate $df(1, 3)$ for $(dx, dy) = (0.1, -0.2)$. What is the significance of this number?

SOLUTION: The function has continuous partial derivatives in a neighborhood of $(1, 3)$ and hence is differentiable at $(1, 3)$. One has

$$df(x, y) = f'_x(x, y) dx + f'_y(x, y) dy = \frac{xy dx}{\sqrt{1 + x^2 y}} + \frac{x^2 dy}{2\sqrt{1 + x^2 y}}.$$

Then

$$df(1, 3) = \frac{3}{2} dx + \frac{1}{4} dy = 0.15 - 0.05 = 0.1.$$

The number $f(1, 3) + df(1, 3)$ defines the value of the linearization $L(x, y)$ of f at $(1, 3)$ for $(x, y) = (1 + dx, 3 + dy)$. It can be used to approximate $f(1 + dx, 3 + dy) - f(1, 3) \approx df(1, 3)$ when $|dx|$ and $|dy|$ are small enough. In particular, $f(1 + 0.1, 3 - 0.2) - f(1, 3) = 0.09476$ (a calculator value), which is to be compared with $df(1, 3) = 0.1$. \square

92.1. Error Analysis. Suppose a quantity f depends on several other quantities, say, x , y , and z , for definiteness, that is, f is a function $f(x, y, z)$. Suppose measurements show that $x = x_0$, $y = y_0$, and $z = z_0$. Since, in practice, all measurements contain errors, the value $f(x_0, y_0, z_0)$ does not have much practical significance until its error is determined.

For example, the volume of a rectangle with dimensions x , y , and z is the function of three variables $V(x, y, z) = xyz$. In practice, repetitive measurements give the values of x , y , and z from 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 upper bounds of the absolute errors or the maximal uncertainties of the measurements. To indicate the maximal uncertainty in the measured quantities, one writes $x = x_0 \pm \delta x$ and similarly for y and z . Different methods of the length measurement would have different absolute error bounds. In other words, the dimensions x , y , and z and the bounds δx , δy , and δz are all independent variables. Since the error bounds 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 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 error bounds $\delta \mathbf{r}$, what is the absolute error bound of the volume value calculated at \mathbf{r}_0 ? For each particular measurement, the error is $V(\mathbf{r}_0 + d\mathbf{r}) - V(\mathbf{r}_0) = dV(\mathbf{r}_0)$ if terms tending to 0 faster than $\|d\mathbf{r}\|$ can be neglected. 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 absolute error bound 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 bound 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 13.20. (Absolute and Relative Error Bounds).

Let f be a quantity that depends on other quantities $\mathbf{r} = (x_1, x_2, \dots, x_m)$ so that $f = f(\mathbf{r})$ is a differentiable function. Suppose that the values $x_i = a_i$ are known with the absolute error bounds δx_i . Put $\mathbf{r}_0 = (a_1, a_2, \dots, a_m)$ 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, if $f(\mathbf{r}_0) \neq 0$, $\delta f/|f(\mathbf{r}_0)|$ are called, respectively, the absolute and relative error bounds of the value of f at $\mathbf{r} = \mathbf{r}_0$.

In the above example, the relative error bound 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 , and \mathbf{r}_0 is fixed, the maximum is attained by setting dx_i equal to δx_i for all i for which the coefficient $f'_{x_i}(\mathbf{r}_0)$ is positive, and equal to $-\delta x_i$ for all i for which the coefficient $f'_{x_i}(\mathbf{r}_0)$ is negative. So the absolute error bound 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 a Linear Approximation. If a function $f(x)$ is differentiable sufficiently many times, then its linear approximation can be systematically improved by using Taylor polynomials (see Calculus I and Calculus II). The Taylor theorem asserts that if $f(x)$ has continuous derivatives up to order n on an interval I containing x_0 and $f^{(n+1)}$ exists and is bounded on I , $|f^{(n+1)}(x)| < M_{n+1}$ for some constant M_{n+1} , then

$$\begin{aligned} f(x) &= T_n(x) + \varepsilon_{n+1}(x), \\ T_n(x) &= f(x_0) + \frac{f'(x_0)}{1!} \Delta + \frac{f''(x_0)}{2!} \Delta^2 + \cdots + \frac{f^{(n)}(x_0)}{n!} \Delta^n, \\ (13.10) \quad |\varepsilon_{n+1}(x)| &\leq \frac{M_{n+1}}{(n+1)!} |x - x_0|^{n+1}, \end{aligned}$$

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 $f(x) \approx T_n(x)$. 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:

$$(13.11) \quad |\varepsilon_2(x)| = |f(x) - L(x)| \leq \frac{M_2}{2} \Delta^2.$$

The differential df can be viewed as the result of the action of the operator $d = dx(d/dx)$ on f : $df = dx f'$. If the variation Δ of x is viewed as an independent variable, like the differential dx , then it is convenient to introduce *higher-order differentials* of f by the rule

$$d^n f(x) = f^{(n)}(x)(dx)^n = \left(dx \frac{d}{dx} \right)^n f(x),$$

where the action of the powers d^n on f is understood as successive actions of the operator d , $d^n f = d^{n-1}(df)$, in which the variables dx and x are independent. For example, $d^2 f = dx(d/dx)(f' dx) = (dx)^2(d/dx)f' = (dx)^2 f''$ (when differentiating, the variable dx is viewed

as a constant). Then the Taylor polynomial $T_n(x)$ about x_0 is

$$T_n(x) = f(x_0) + \frac{1}{1!}df(x_0) + \frac{1}{2!}d^2f(x_0) + \cdots + \frac{1}{n!}d^n f(x_0),$$

where $dx = x - x_0$. It represents an expansion of $f(x_0 + dx)$ in powers of the differential dx . The Taylor polynomial approximation $f(x_0 + dx) \approx T_n(x)$ is an approximation in which the contributions of higher powers $(dx)^k$, $k > n$, are neglected, provided f is differentiable sufficiently many times. The Taylor theorem ensures that this approximation is better than a linear approximation in the sense that the approximation error decreases faster than $(dx)^n = (x - x_0)^n$ as $x \rightarrow x_0$. It also provides more information about a local behavior of the function near a particular point x_0 (e.g., the concavity of f near x_0). Naturally, this concept should be quite useful in the multivariable case.

92.3. Taylor Polynomials of Two Variables. Let $f(x, y)$ be a function of two variables. The differentials dx and dy are another two independent variables. By analogy with the one-variable case, the differential df is viewed as the result of the action of the operator d on f :

$$df(x, y) = \left(dx \frac{\partial}{\partial x} + dy \frac{\partial}{\partial y} \right) f(x, y) = dx f'_x(x, y) + dy f'_y(x, y).$$

DEFINITION 13.21. *Suppose that f has continuous partial derivatives up to order n . The quantity*

$$d^n f(x, y) = \left(dx \frac{\partial}{\partial x} + dy \frac{\partial}{\partial y} \right)^n f(x, y)$$

is called the n -th order differential of f , where the action of powers d^n on f is defined successively $d^n f = d^{n-1}(df)$ and the variables dx , dy , x , and y are viewed as independent when differentiating.

The differential $d^n f$ is a function of four variables dx , dy , x , and y . For example,

$$\begin{aligned} d^2 f &= \left(dx \frac{\partial}{\partial x} + dy \frac{\partial}{\partial y} \right)^2 f = \left(dx \frac{\partial}{\partial x} + dy \frac{\partial}{\partial y} \right) (dx f'_x + dy f'_y) \\ &= \left((dx)^2 \frac{\partial}{\partial x} + dx dy \frac{\partial}{\partial y} \right) f'_x + \left(dx dy \frac{\partial}{\partial x} + (dy)^2 \frac{\partial}{\partial y} \right) f'_y \\ &= f''_{xx}(dx)^2 + 2f''_{xy} dx dy + f''_{yy}(dy)^2. \end{aligned}$$

By the continuity of partial derivatives, the order of differentiation is irrelevant, $f''_{xy} = f''_{yx}$. The numerical coefficients at each of the terms are binomial coefficients: $(a + b)^2 = a^2 + 2ab + b^2$. Since the order of

differentiation is irrelevant (Clairaut's theorem), this observation holds in general:

$$d^n f = \sum_{k=0}^n B_k^n \frac{\partial^n f}{\partial x^{n-k} \partial y^k} (dx)^{n-k} (dy)^k, \quad B_k^n = \frac{n!}{k!(n-k)!},$$

where B_k^n are the binomial coefficients: $(a+b)^n = \sum_{k=0}^n B_k^n a^{n-k} b^k$.

EXAMPLE 13.34. Find $d^n f$ if $f(x, y) = e^{ax+by}$, where a and b are constants.

SOLUTION: Since $f'_x = ae^{ax+by}$ and $f'_y = be^{ax+by}$,

$$df = ae^{ax+by} dx + be^{ax+by} dy = e^{ax+by}(adx + bdy).$$

Since $f''_{xx} = a^2 e^{ax+by}$, $f''_{xy} = abe^{ax+by}$, and $f''_{yy} = b^2 e^{ax+by}$,

$$\begin{aligned} d^2 f &= a^2 e^{ax+by} (dx)^2 + 2abe^{ax+by} dx dy + b^2 e^{ax+by} (dy)^2 \\ &= e^{ax+by} (a^2 (dx)^2 + 2ab dx dy + b^2 (dy)^2) = e^{ax+by} (a dx + b dy)^2. \end{aligned}$$

Furthermore, by noting that each differentiation with respect to x brings down a factor a , while the partial derivative with respect to y brings down a factor b , it is concluded that $\partial^n f / \partial x^{n-k} \partial y^k = a^{n-k} b^k e^{ax+by}$. Using the binomial expansion, one infers

$$d^n f = e^{ax+by} (a dx + b dy)^n$$

for all $n = 1, 2, \dots$ □

DEFINITION 13.22. (Taylor Polynomials of Two Variables)

Let f have continuous partial derivatives up to order n . The Taylor polynomial of order n about a point (x_0, y_0) is

$$T_n(x, y) = f(x_0, y_0) + \frac{1}{1!} df(x_0, y_0) + \frac{1}{2!} d^2 f(x_0, y_0) + \dots + \frac{1}{n!} d^n f(x_0, y_0),$$

where $dx = x - x_0$ and $dy = y - y_0$.

For example, put $\mathbf{r} = (x, y)$, $\mathbf{r}_0 = (x_0, y_0)$, $dx = x - x_0$, and $dy = y - y_0$. The first four Taylor polynomials are

$$T_0(\mathbf{r}) = f(\mathbf{r}_0),$$

$$T_1(\mathbf{r}) = f(\mathbf{r}_0) + f'_x(\mathbf{r}_0) dx + f'_y(\mathbf{r}_0) dy = L(\mathbf{r}),$$

$$T_2(\mathbf{r}) = T_1(\mathbf{r}) + \frac{f''_{xx}(\mathbf{r}_0)}{2} (dx)^2 + f''_{xy}(\mathbf{r}_0) dx dy + \frac{f''_{yy}(\mathbf{r}_0)}{2} (dy)^2,$$

$$T_3(\mathbf{r}) = T_2(\mathbf{r}) + \frac{f'''_{xxx}(\mathbf{r}_0)}{6} (dx)^3 + \frac{f'''_{xxy}(\mathbf{r}_0)}{2} (dx)^2 dy$$

$$+ \frac{f'''_{xyy}(\mathbf{r}_0)}{2} dx(dy)^2 + \frac{f'''_{yyy}(\mathbf{r}_0)}{6} (dy)^3.$$

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.

EXAMPLE 13.35. Let $P_n(x, y)$ be a polynomial of degree n . Find its Taylor polynomials about $(0, 0)$. In particular, find Taylor polynomials for $P_3(x, y) = 1 + 2x - xy + y^2 + 4x^3 - y^2x$.

SOLUTION: All partial derivatives of P_n of order higher than n vanish. Therefore, $d^k P_n = 0$ for $k > n$, and hence for any polynomial of degree n , $T_n = P_n$ and also $T_k = P_n$ if $k > n$. Any polynomial can be uniquely decomposed into the sum $P_n = Q_0 + Q_1 + \cdots + Q_n$, where Q_k is a homogeneous polynomial of degree k ; it contains only monomials of degree k . The differential $d^k f$ is a homogeneous polynomial of degree k in the variables dx and dy . Therefore, Definition 13.22 defines T_k as the sum of homogeneous polynomials in x and y if $(x_0, y_0) = (0, 0)$. Two polynomials are equal only if the coefficients at the corresponding monomials match. It follows from $T_n = P_n$ that $T_n = T_{n-1} + (T_n - T_{n-1}) = Q_0 + Q_1 + \cdots + Q_{n-1} + Q_n$. Since Q_n and $T_n - T_{n-1}$ contains only monomials of degree n , the equality is possible only if $Q_n = T_n - T_{n-1}$, and hence $T_{n-1} = Q_0 + Q_1 + \cdots + Q_{n-1}$. Continuing the process recursively backward, it is concluded that

$$T_k = Q_0 + Q_1 + \cdots + Q_k, \quad k = 0, 1, \dots, n.$$

In particular, for the given polynomial P_3 , one has $Q_0 = 1$, $Q_1 = 2x$, $Q_2 = -xy + y^2$, and $Q_3 = 4x^3 - y^2x$. Therefore, its Taylor polynomials about the origin are $T_0 = 1$, $T_1 = T_0 + 2x$, $T_2 = T_1 - xy + y^2$, and $T_k = P_3$ for $k \geq 3$. \square

THEOREM 13.15. (Taylor Theorem).

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 $n - 1$ on D . Then $f(\mathbf{r}) = T_{n-1}(\mathbf{r}) + \varepsilon_n(\mathbf{r})$, where the remainder $\varepsilon(\mathbf{r})$ satisfies the condition

$$|\varepsilon_n(\mathbf{r})| \leq h_n(\mathbf{r}) \|\mathbf{r} - \mathbf{r}_0\|^{n-1}, \quad \text{where } h_n(\mathbf{r}) \rightarrow 0 \text{ as } \mathbf{r} \rightarrow \mathbf{r}_0.$$

In Section 92.4, Taylor polynomials for functions of any number of variables will be defined. Theorem 13.15 is true, just as written, no matter how many variables there are; that is $\mathbf{r} = (x_1, x_2, \dots, x_m)$ for any number of variables m . For $n = 2$, this theorem is nothing but Theorem 13.12. The continuity of partial derivatives ensures the existence of a good linear approximation $L(\mathbf{r}) = T_1(\mathbf{r})$ in the sense that the difference $f(\mathbf{r}) - T_1(\mathbf{r})$ decreases to 0 faster than $\|\mathbf{r} - \mathbf{r}_0\|$ as $\mathbf{r} \rightarrow \mathbf{r}_0$. For $n > 2$, it states that the approximation of f by the Taylor

polynomial T_{n-1} is a good approximation in the sense that the error decreases faster than $\|\mathbf{r} - \mathbf{r}_0\|^{n-1}$. A practical significance of the Taylor theorem is that higher-order differentials of a function can be used to obtain successively better approximations of values of a function near a point if the function has continuous partial derivatives of higher orders in a neighborhood of that point.

EXAMPLE 13.36. Let $f(x, y) = \sqrt{1 + x^2y}$. Find $df(1, 3)$ and $d^2f(1, 3)$ and use them to approximate $f(1 + 0.1, 3 - 0.2)$.

SOLUTION: Put $(dx, dy) = (0.1, -0.2)$. It was found in Example 13.33 that $df(1, 3) = 0.1$. The second partial derivatives are obtained by the quotient rule (see f'_x and f'_y in Example 13.33):

$$\begin{aligned} f''_{xx}(1, 3) &= \left. \frac{y(1 + x^2y)^{1/2} - x^2y^2(1 + x^2y)^{-1/2}}{1 + x^2y} \right|_{(1,3)} = \frac{3}{8}, \\ f''_{xy}(1, 3) &= \left. \frac{2x(1 + x^2y)^{1/2} - x^3y(1 + x^2y)^{-1/2}}{2(1 + x^2y)} \right|_{(1,3)} = \frac{5}{16}, \\ f''_{yy}(1, 3) &= \left. -\frac{x^4}{4}(1 + x^2y)^{-3/2} \right|_{(1,3)} = -\frac{1}{8}. \end{aligned}$$

Therefore,

$$\begin{aligned} d^2f(1, 3) &= f''_{xx}(1, 3)(dx)^2 + 2f''_{xy}(1, 3) dx dy + f''_{yy}(1, 3)(dy)^2 \\ &= \frac{3}{8}(dx)^2 + \frac{5}{8} dx dy - \frac{1}{8}(dy)^2 = -0.01375. \end{aligned}$$

The linear approximation is $f(1 + dx, 3 + dy) \approx f(1, 3) + df(1, 3) = 2 + 0.1 = 2.1$. The quadratic approximation is $f(1 + dx, 3 + dy) \approx f(1, 3) + df(1, 3) + \frac{1}{2}d^2f(1, 3) = 2.1 - 0.01375/2 = 2.093125$, while a calculator value of $f(1 + dx, 3 + dy)$ is 2.094755 (rounded to the same significant digit). Evidently, the quadratic approximation (the approximation by the second-degree Taylor polynomial) is better than the linear approximation. \square

Yet, the Taylor theorem does not allow us to estimate the accuracy of the approximation because the function h_n remains unknown. What is the order of approximation needed to obtain an error smaller than some prescribed value?

COROLLARY 13.3. (Accuracy of Taylor Polynomial Approximations). *If, in addition to the hypotheses of Theorem 13.15, the function f has partial derivatives of order n that are bounded on D , that is, there exist numbers M_{nk} , $k = 1, 2, \dots, n$, such that $|\partial^n f(\mathbf{r})/\partial^{n-k}x\partial^ky| \leq M_{nk}$ for*

all $\mathbf{r} \in D$, then the remainder satisfies

$$|\varepsilon_n(\mathbf{r})| \leq \sum_{k=0}^n \frac{B_k^n M_{nk}}{n!} |x - x_0|^{n-k} |y - y_0|^k$$

for all $(x, y) \in D$, where $B_k^n = n!/(k!(n-k)!)$ are binomial coefficients.

Next, note $|x - x_0| \leq \|\mathbf{r} - \mathbf{r}_0\|$ and $|y - y_0| \leq \|\mathbf{r} - \mathbf{r}_0\|$ and hence $|x - x_0|^{n-k} |y - y_0|^k \leq \|\mathbf{r} - \mathbf{r}_0\|^n$. Making use of this inequality, one infers that

$$(13.12) \quad |\varepsilon_n(\mathbf{r})| \leq \frac{M_n}{n!} \|\mathbf{r} - \mathbf{r}_0\|^n,$$

where the constant $M_n = \sum_{k=0}^n B_k^n M_{nk}$. In particular, for the linear approximation $n = 2$,

$$(13.13) \quad |f(\mathbf{r}) - L(\mathbf{r})| \leq \frac{M_2}{2} \|\mathbf{r} - \mathbf{r}_0\|^2,$$

where $M_2 = M_{20} + 2M_{11} + M_{02}$. The results (13.12) and (13.13) are to be compared with the similar results (13.10) and (13.11) in the one-variable case. If the second partial derivatives are continuous and bounded near \mathbf{r}_0 , then variations of their values may be neglected in a sufficiently small neighborhood of \mathbf{r}_0 , and the numbers M_{20} , M_{11} , and M_{02} may be approximated by the absolute values of the corresponding partial derivatives at \mathbf{r}_0 so that

$$|\varepsilon_2| \approx \frac{1}{2} \left(|f''_{xx}(\mathbf{r}_0)|(dx)^2 + 2|f''_{xy}(\mathbf{r}_0)||dx dy| + |f''_{yy}(\mathbf{r}_0)|(dy)^2 \right)$$

for sufficiently small variations $dx = x - x_0$ and $dy = y - y_0$. Such an estimate is often sufficient for practical purposes to assess the accuracy of the linear approximation. This estimate works even better if f has continuous partial derivatives of the third order because the second partial derivatives would have a good linear approximation and variations of their values near \mathbf{r}_0 are of order $\|d\mathbf{r}\|$. Consequently, they can only produce variations of ε_2 of order $\|d\mathbf{r}\|^3$, which can be neglected as compared to $\|d\mathbf{r}\|^2$ for sufficiently small $\|d\mathbf{r}\|$.

EXAMPLE 13.37. Find the linear approximation near $(0, 0)$ of $f(x, y) = \sqrt{1 + x + 2y}$ and assess its accuracy in the square $|x| < 1/4$, $|y| < 1/4$.

SOLUTION: One has $f'_x = \frac{1}{2}(1 + x + 2y)^{-1/2}$ and $f'_y = (1 + x + 2y)^{-1/2}$ so that $f'_x(0, 0) = 1/2$ and $f'_y(0, 0) = 1$. The linear approximation is $T_1(x, y) = L(x, y) = 1 + x/2 + y$. The second partial derivatives are $f''_{xx} = -\frac{1}{4}(1 + x + 2y)^{-3/2}$, $f''_{xy} = -\frac{1}{2}(1 + x + 2y)^{-3/2}$, and $f''_{yy} = -(1 + x +$

$2y)^{-3/2}$. Their absolute values are maximal if the combination $1+x+2y$ is minimal on the square. Setting $x = y = -1/4$, $1/4 < 1+x+2y$ in the square. Therefore, $|f''_{xx}| < M_{20} = 2$, $|f''_{xy}| < M_{11} = 4$, and $|f''_{yy}| < M_{02} = 8$. Thus, $|f(x, y) - L(x, y)| < x^2 + |xy| + 2y^2$ in the square. \square

EXAMPLE 13.38. 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 $f(h, r) = \pi hr^2$. 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. Put $dh = 2\delta = 0.1$ and $dr = \delta = 0.05$. Then $V \approx df(10, 5)$. One has $f'_h = \pi r^2$ and $f'_r = 2\pi hr$; hence, $df(10, 5) = f'_h(10, 5) dh + f'_r(10, 5) dr = 25\pi dh + 100\pi dr = 7.5\pi \text{ cm}^3$.

To assess the accuracy, note that f is a polynomial, and therefore all its partial derivatives of any order are continuous. In particular, $f''_{hh} = 0$, $f''_{hr} = 2\pi r$, and $f''_{rr} = 2\pi h$. Since dh and dr are small compared to $r = 5$ and $h = 10$, the variations of second derivatives in the rectangle $[5 - dr, 5 + dr] \times [10 - dh, 10 + dh]$ may be neglected. Then $|\varepsilon_2| = \frac{1}{2}(M_{20}(dh)^2 + 2M_{11}|dh dr| + M_{02}(dr)^2)$, where the estimates $M_{20} = |f''_{hh}(10, 5)| = 0$, $M_{11} = |f''_{hr}(10, 5)| = 10\pi$, and $M_{02} = |f''_{rr}(10, 5)| = 20\pi$ can be used. So $|\varepsilon_2| = 0.075\pi$. The relative error is $|\varepsilon_2|/V = 0.01$, or 1%. \square

92.4. Multivariable Taylor Polynomials. For more than two variables, Taylor polynomials are defined similarly. Let $\mathbf{r} = (x_1, x_2, \dots, x_m)$ and let $d\mathbf{r} = (dx_1, dx_2, \dots, dx_m)$. Suppose that a function f has continuous partial derivatives up to order n . The n -th order differential of $f(\mathbf{r})$ is defined by

$$d^n f(\mathbf{r}) = \left(dx_1 \frac{\partial}{\partial x_1} + dx_2 \frac{\partial}{\partial x_2} + \cdots + dx_m \frac{\partial}{\partial x_m} \right)^n f(\mathbf{r}),$$

where the variables \mathbf{r} and $d\mathbf{r}$ are viewed as independent when differentiating. The Taylor polynomial of degree n about a point \mathbf{r}_0 is

$$T_n(\mathbf{r}) = f(\mathbf{r}_0) + \frac{1}{1!} df(\mathbf{r}_0) + \frac{1}{2!} d^2 f(\mathbf{r}_0) + \cdots + \frac{1}{n!} d^n f(\mathbf{r}_0),$$

where $d\mathbf{r} = \mathbf{r} - \mathbf{r}_0$. The Taylor theorem has a natural extension to the multivariable case: $f(\mathbf{r}) = T_{n-1}(\mathbf{r}) + \varepsilon_n(\mathbf{r})$, where the remainder $\varepsilon_n(\mathbf{r})$ satisfies the condition (13.12). Taylor polynomials obey the recurrence relation

$$T_n(\mathbf{r}) = T_{n-1}(\mathbf{r}) + \frac{1}{n!} d^n f(\mathbf{r}_0).$$

So, for practical purposes, the error $|\varepsilon_n|$ of the approximation $f \approx T_{n-1}$ may be estimated by $d^n f(\mathbf{r}_0)/n!$, where dx_j are replaced by their absolute values $|dx_j|$ and the values of partial derivatives are also replaced by their absolute values (just like it has been done when estimating $|\varepsilon_2|$ in the case of two variables), provided the partial derivatives of f of order n or higher are continuous in a neighborhood of \mathbf{r}_0 .

Calculation of higher-order derivatives to find Taylor polynomials might be a technically tedious problem. In some special cases, however, it can be avoided. The concept is illustrated by the following example.

EXAMPLE 13.39. Find T_3 for the function $f(x, y, z) = \sin(xy + z)$ about the origin.

SOLUTION: The Taylor polynomial T_3 in question is a polynomial of degree 3 in x , y , and z , which is uniquely determined by the coefficients of monomials of degree less than or equal to 3. Put $u = xy + z$. The variable u is small near the origin. So the Taylor polynomial approximation for f near the origin is determined by the Taylor polynomials for $\sin u$ about $u = 0$. The latter is obtained from the Maclaurin series $\sin u = u - \frac{1}{6}u^3 + \varepsilon_5(u)$, where ε_5 contains only monomials of degree 5 and higher. Since the polynomial u vanishes at the origin, its powers u^n may contain only monomials of degree n and higher. Therefore, T_3 is obtained from

$$\begin{aligned} u - \frac{1}{6}u^3 &= (xy + z) - \frac{1}{6}(xy + z)^3 \\ &= z + xy - \frac{1}{6}\left(z^3 + 3(xy)z^2 + 3(xy)^2z + (xy)^3\right) \end{aligned}$$

by retaining in the latter all monomials up to degree 3, which yields $T_3(\mathbf{r}) = z + xy - \frac{1}{6}z^3$. Evidently, the procedure is far simpler than calculating 19 partial derivatives (up to the third order)! \square

92.5. Study Problems.

Problem 13.10. Find T_1 , T_2 , and T_3 for $f(x, y, z) = (1 + xy)/(1 + x + y^2 + z^3)$ about the origin.

SOLUTION: The function f is a rational function. It is therefore sufficient to find a suitable Taylor polynomial for the function $(1 + x + y^2 + z^3)^{-1}$ and then multiply it by the polynomial $1 + xy$, retaining only monomials up to degree 3. Put $u = x + y^2 + z^3$. Then $(1 + u)^{-1} = 1 - u + u^2 - u^3 + \dots$ (as a geometric series). Note that, for $n \geq 4$, the terms u^n contain only monomials of degree 4 and higher and

hence can be omitted. Up to degree 3, one has $u^2 = x^2 + 2xy^2 + \dots$ and $u^3 = x^3 + \dots$. Therefore,

$$(1+xy)(1-u+u^2-u^3) = (1+xy)(1-x-y^2-z^3+x^2+2xy^2-x^3+\dots).$$

Carrying out the multiplication and arranging the monomials in the order of increasing degrees, one infers:

$$\begin{aligned} T_1(x, y, z) &= 1 - x, \\ T_2(x, y, z) &= T_1(x, y, z) + x^2 + xy - y^2, \\ T_3(x, y, z) &= T_2(x, y, z) - x^3 - x^2y + 2xy^2 - z^3. \end{aligned}$$

□

Problem 13.11. (Multivariable Taylor and Maclaurin Series)

Suppose that a function f has continuous partial derivatives of any order and the remainder in the Taylor polynomial approximation $f = T_{n-1} + \varepsilon_n$ near \mathbf{r}_0 converges to 0 as $n \rightarrow \infty$ (i.e., $\varepsilon_n \rightarrow 0$). Then the function can be represented by the Taylor series about a point \mathbf{r}_0 :

$$f(\mathbf{r}) = f(\mathbf{r}_0) + \sum_{n=1}^{\infty} \frac{1}{n!} d^n f(\mathbf{r}_0),$$

where $d\mathbf{r} = \mathbf{r} - \mathbf{r}_0$. The Taylor series about $\mathbf{r}_0 = \mathbf{0}$ is called the Maclaurin series. Find the Maclaurin series of $\sin(xy^2)$.

SOLUTION: Since the argument of the sine is the polynomial xy^2 , the Maclaurin series of f can be obtained from the Maclaurin series of $\sin u$ by setting $u = xy^2$ in it. From Calculus II,

$$f(\mathbf{r}) = \sin u = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{(2n-1)!} u^{2n-1} = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{(2n-1)!} x^{2n-1} y^{4n-2}.$$

□

92.6. Exercises.

(1) Find the differential df of each of the following functions:

- (i) $f(x, y) = x^3 + y^3 - 3xy(x - y)$
- (ii) $f(x, y) = y \cos(x^2y)$
- (iii) $f(x, y) = \sin(x^2 + y^2)$
- (iv) $f(x, y, z) = x + yz + ye^{xyz}$
- (v) $f(x, y, z) = \ln(x^x y^y z^z)$
- (vi) $f(x, y, z) = y/(1 + xyz)$
- (vii) $f(\mathbf{r}) = \sqrt{a^2 - \|\mathbf{r}\|^2}$, where a is a constant and $\mathbf{r} = (x_1, x_2, \dots, x_m)$

(2) Four positive numbers, each less than 100, are rounded and then multiplied together. Use differentials to estimate the maximum possible error in the computed product that might result from the rounding.

(3) A boundary stripe 10 cm wide is painted around a rectangle whose dimensions are 50 m by 100 m. Use differentials to approximate the number of square meters of paint in the stripe. Assess the accuracy of the approximation.

(4) A rectangle has sides of $x = 6$ m and $y = 8$ m. Use differentials to estimate the change of the length of the diagonal and the area of the rectangle if x is increased by 2 cm and y is decreased by 5 cm. Assess the accuracy of the estimates.

(5) Consider a sector of a disk with radius $R = 20$ cm and the angle $\theta = \pi/3$. Use the differential to determine how much the radius should be decreased in order for the area of the sector to remain the same when the angle is increased by 1° . Assess the accuracy of the estimate.

(6) Let the quantities f and g be measured with relative errors R_f and R_g . Show that the relative error of the product fg is the sum $R_f + R_g$.

(7) Measurements of the radius r and the height h of a cylinder are $r = 2.2 \pm 0.1$ and $h = 3.1 \pm 0.2$, in meters. Find the absolute and relative errors of the volume of the cylinder calculated from these data.

(8) The adjacent sides of a triangle have lengths $a = 100 \pm 2$ and $b = 200 \pm 5$, in meters, and the angle between them is $\theta = 60^\circ \pm 1^\circ$. Find the relative and absolute errors in calculation of the length of the third side of the triangle.

(9) If R is the total resistance of n resistors, connected in parallel, with resistances R_j , $j = 1, 2, \dots, n$, then $R^{-1} = R_1^{-1} + R_2^{-1} + \dots + R_n^{-1}$. If each resistance R_j is known with a relative error of 0.5%, what is the relative error of R ?

(10) Use the Taylor theorem to assess the maximal error of the linear approximation of the following functions about the origin in the ball of radius R (i.e., for $\|\mathbf{r}\| \leq R$):

(i) $f(x, y) = \sqrt{1 + \sin(x + y)}$, $R = \frac{1}{2}$

(ii) $f(x, y) = \frac{1+3x}{2+y}$, $R = 1$

(iii) $f(x, y, z) = \ln(1 + x + 2y - 3z)$, $R = 0.1$

(11) Find the indicated differentials of the following functions:

(i) $f(x, y) = x - y + x^2y$, $d^n f$, $n = 1, 2, \dots$

(ii) $f(x, y) = \ln(x + y)$, $d^n f$, $n = 1, 2, \dots$

(iii) $f(x, y) = \sin(x) \cosh(y)$, $d^3 f$

(iv) $f(x, y, z) = xyz$, $d^n f$, $n = 1, 2, \dots$

(v) $f(x, y, z) = 1/(1 + xyz)$, $d^2 f$

(vi) $f(\mathbf{r}) = \|\mathbf{r}\|$, df and $d^2 f$, $\mathbf{r} = (x_1, x_2, \dots, x_m)$

(12) Let $f(\mathbf{r}) = g(u)$, where u is a linear function of \mathbf{r} , $u = c + \mathbf{n} \cdot \mathbf{r}$, where c is a constant and \mathbf{n} is a constant vector. Show that $d^n f = g^{(n)}(u)(\mathbf{n} \cdot d\mathbf{r})^n$.

(13) Let $Q_n(x, y, z)$ be a homogeneous polynomial of degree n (it contains only monomials of degree n). Show that $d^n Q_n(x, y, z) = n! Q_n(dx, dy, dz)$.

(14) Find the Taylor polynomial T_2 about a specified point and a given function:

(i) $f(x, y) = y + x^3 + 2xy^2 - x^2y^2, (1, 1)$

(ii) $f(x, y) = \sin(xy), (\pi/2, 1)$

(iii) $f(x, y) = x^y, (1, 1)$

(15) Let $f(x, y) = x^y$. Use Taylor polynomials T_1 , T_2 , and T_3 to approximate $f(1.2, 0.7)$. Compare the results of the three approximations with a calculator value of $f(1.2, 0.7)$.

(16) Use the method of Example 13.39 and Study Problem 13.10 to find the indicated Taylor polynomials about the origin for each of the following functions:

(i) $f(x, y) = \sqrt{1 + x + 2y}, T_n(x, y), n \leq 2$

(ii) $f(x, y) = \frac{xy}{1 - x^2 - y^2}, T_n(x, y), n \leq 4$

(iii) $f(x, y, z) = \sin(x + 2y + z^2), T_n(x, y, z), n \leq 3$

(iv) $f(x, y, z) = e^{xy} \cos(zy), T_n(x, y, z), n \leq 4$

(v) $f(x, y) = \ln(1 + x + 2y)/(1 + x^2 + y^2), T_n(x, y), n \leq 2$

(17) Find polynomials of degree 2 to calculate approximate values of the following functions in the region in which $x^2 + y^2$ is small as compared with 1:

(i) $f(x, y) = \cos y / \cos x$

(ii) $\tan^{-1}\left(\frac{1+x+y}{1-x+y}\right)$

(18) Find a nonzero polynomial of the smallest degree to approximate a local behavior of the function $\cos(x + y + z) - \cos(x) \cos(y) \cos(z)$ near the origin.

(19) Let

$$g(r) = \frac{1}{2\pi} \int_0^{2\pi} f(x_0 + r \cos \theta, y_0 + r \sin \theta) d\theta,$$

where f has continuous partial derivatives up to order 4 and x_0 and y_0 are constants. Find T_4 about $r = 0$ for $g(r)$.

(20) Consider the roots $z = z(x, y)$ of the equation $F(x, y, z) = z^5 + xz - y = 0$ near $(1, 2, 1)$. Use Taylor polynomials $T_1(x, y)$ and $T_2(x, y)$ about $(1, 2)$ to approximate $z(x, y)$. In particular, calculate the approximations $z_1 = T_1(0.7, 2.5)$ and $z_2 = T_2(0.7, 2.5)$ of $z(0.7, 2.5)$. Use a calculator to find $F(0.7, 2.5, z_1)$ and $F(0.7, 2.5, z_2)$. Their deviation from 0 determines an error of the approximations z_1 and z_2 . Which

of the approximations is more accurate? *Hint:* Use the result of Study Problem 13.9.

93. Directional Derivative and the Gradient

93.1. Directional Derivative. Let f be a function of several variables $\mathbf{r} = (x_1, x_2, \dots, x_m)$. The partial derivative $f'_{x_i}(\mathbf{r}_0)$ is the rate of change in the direction of the i th 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 southeast or northwest directions. Naturally, these slopes generally differ from the slopes f'_x and f'_y .

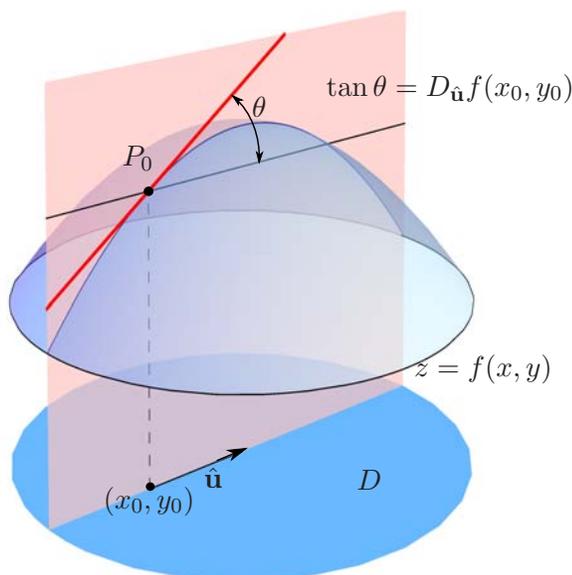


FIGURE 13.9. Geometrical significance of the directional derivative of a function of two variables. Consider the graph $z = f(x, y)$ over a region D . The vertical plane (i.e., parallel to the z axis) through (x_0, y_0) and parallel to the unit vector $\hat{\mathbf{u}}$ intersects the graph along a curve. The slope of the tangent line to the curve of intersection at the point $P_0 = (x_0, y_0, f(x_0, y_0))$ is determined by the directional derivative $D_{\hat{\mathbf{u}}}f(x_0, y_0)$. So the directional derivative defines the rate of change of f at (x_0, y_0) in the direction $\hat{\mathbf{u}}$.

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 $F(h) = f(\mathbf{r}(h))$. The numbers $F(0)$ and $F(h)$ 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 $F'(0)$. Therefore, the following definition is natural.

DEFINITION 13.23. (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

$$D_{\hat{\mathbf{u}}}f(\mathbf{r}_0) = \lim_{h \rightarrow 0} \frac{f(\mathbf{r}_0 + h\hat{\mathbf{u}}) - f(\mathbf{r}_0)}{h}$$

if the limit exists.

The number $D_{\hat{\mathbf{u}}}f(\mathbf{r}_0)$ is the rate of change of f at \mathbf{r}_0 in the direction of $\hat{\mathbf{u}}$. Suppose that f is a differentiable function. By definition, $D_{\hat{\mathbf{u}}}f(\mathbf{r}_0) = df(\mathbf{r}(h))/dh$ taken at $h = 0$, where $\mathbf{r}(h) = \mathbf{r}_0 + h\hat{\mathbf{u}}$. So, by the chain rule,

$$\frac{df(\mathbf{r}(h))}{dh} = f'_{x_1}(\mathbf{r}(h))x'_1(h) + f'_{x_2}(\mathbf{r}(h))x'_2(h) + \cdots + f'_{x_m}(\mathbf{r}(h))x'_m(h).$$

Setting $h = 0$ in this relation and taking into account that $\mathbf{r}'(h) = \hat{\mathbf{u}}$ or $x'_i(h) = u_i$, where $\hat{\mathbf{u}} = (u_1, u_2, \dots, u_m)$, one infers that

$$(13.14) \quad D_{\hat{\mathbf{u}}}f(\mathbf{r}_0) = f'_{x_1}(\mathbf{r}_0)u_1 + f'_{x_2}(\mathbf{r}_0)u_2 + \cdots + f'_{x_m}(\mathbf{r}_0)u_m.$$

Remark. If f has partial derivatives at \mathbf{r}_0 , but is not differentiable at \mathbf{r}_0 , then the relation (13.14) is false. An example is given in Study Problem 13.12. Note that (13.14) follows from the chain rule, but the mere existence of partial derivatives is not sufficient for the chain rule to hold. Furthermore, even if a function has directional derivatives at a point in every direction, it may not be differentiable at that point (no good linear approximation exists at that point).

Equation (13.14) provides a convenient way to compute the directional derivative if f is differentiable. 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 13.40. 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 northwest 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 northwest, he has to climb at an angle of 45° relative to the horizon. \square

EXAMPLE 13.41. Find the directional derivative of $f(x, y, z) = x^2 + 3xz + z^2y$ 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). \square

93.2. The Gradient and Its Geometrical Significance.

DEFINITION 13.24. (The Gradient).

Let f be a differentiable function of several variables $\mathbf{r} = (x_1, x_2, \dots, x_m)$ 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 called 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

$$(13.15) \quad D_{\mathbf{u}}f(\mathbf{r}_0) = \nabla f(\mathbf{r}_0) \cdot \hat{\mathbf{u}}.$$

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 a vector in either a plane or space, respectively. In Example 13.40, the gradient at $(1, 2)$ is $\nabla f(1, 2) = (-3/\sqrt{2}, -2/\sqrt{2})$. In Example 13.41, 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 nonzero 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 $\theta = \pi$. Assume that $\nabla f(\mathbf{r}_0) \neq \mathbf{0}$. Let θ be the angle between the gradient $\nabla f(\mathbf{r}_0)$ and the unit vector $\hat{\mathbf{u}}$. Then

$$(13.16) \quad D_{\mathbf{u}}f(\mathbf{r}_0) = \nabla f(\mathbf{r}_0) \cdot \hat{\mathbf{u}} = \|\nabla f(\mathbf{r}_0)\| \|\hat{\mathbf{u}}\| \cos \theta = \|\nabla f(\mathbf{r}_0)\| \cos \theta$$

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 as depicted in the left and right panels of Figure 13.10, respectively. It is then concluded that the

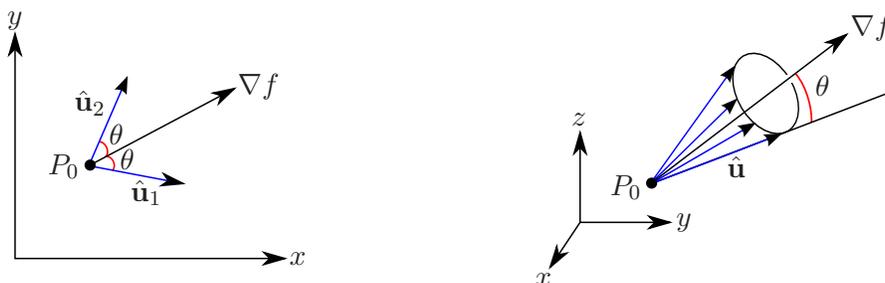


FIGURE 13.10. **Left:** The same rate of change of a function of two variables at a point P_0 occurs in two directions that have the same angle with the gradient $\nabla f(P_0)$. **Right:** The same rate of change of a function of three variables at a point P_0 occurs in infinitely many directions that have the same angle with the gradient $\nabla f(P_0)$.

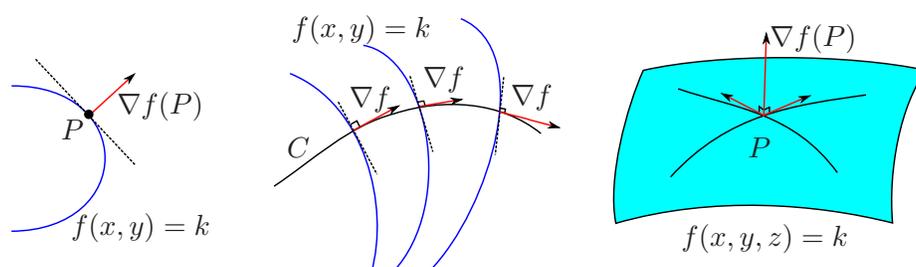


FIGURE 13.11. **Left:** The gradient at a point P is normal to a level curve $f(x, y) = k$ through P of a function f of two variables. **Middle:** A curve C of steepest descent or ascent for a function f has the characteristic property that the gradient ∇f is tangent to it. The level curves (surfaces) of f are normal to C . The function f increases most rapidly along C in the direction of ∇f , and f decreases most rapidly along C in the opposite direction $-\nabla f$. **Right:** The gradient of a function of three variables is normal to any curve through P in the level surface $f(x, y, z) = k$. So $\nabla f(P)$ is a normal to the tangent plane through P to the level surface.

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 13.40, 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 $\varphi \approx 146^\circ$. If the hiker goes in this direction, he has to climb up at an angle of $\tan^{-1}(\sqrt{13/2}) \approx 69^\circ$ with the horizon. Also, note the hiker's original direction was $\varphi = 135^\circ$, which makes the angle 11° with the direction of the steepest ascent. So the slope in the direction $\varphi = 146^\circ + 11^\circ = 157^\circ$ 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 differentiable function of two variables. Suppose that there is a differentiable vector function $\mathbf{r}(t) = (x(t), y(t))$ that traces out the level curve. This vector function

should satisfy 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,

$$\frac{d}{dt}f(x(t), y(t)) = 0 \implies \frac{df}{dt} = f'_x x'(t) + f'_y y'(t) = \nabla f(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = 0$$

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. This is often expressed by saying that the gradient of f is always normal to smooth level curves of f .*

The existence of a smooth vector function that traverses a level curve of f can be established by the implicit function theorem. Suppose that f has continuous partial derivatives and $\nabla f(\mathbf{r}_0) \neq \mathbf{0}$ at a point $\mathbf{r}_0 = (x_0, y_0)$ on the level surface $f(x, y) = k$. The level surface can also be defined as the set of roots of the function $F(x, y) = f(x, y) - k$, that is, the set of solution of $F(x, y) = 0$. In particular, the point \mathbf{r}_0 is a root, $F(x_0, y_0) = 0$. The function F has continuous partial derivatives and $\nabla F(\mathbf{r}_0) = \nabla f(\mathbf{r}_0) \neq \mathbf{0}$. The components of the gradient do not vanish simultaneously, and without loss of generality, one can assume that $F'_y(\mathbf{r}_0) = f'_y(\mathbf{r}_0) \neq 0$. Then, by the implicit function theorem, there is a function $y = g(x)$ such that $F(x, g(x)) = 0$ in some open interval containing x_0 ; that is, the graph $y = g(x)$ coincides with the level curve $f(x, y) = k$ in a neighborhood of (x_0, y_0) , where $y_0 = g(x_0)$. Furthermore, the derivative $g'(x) = -F'_x/F'_y = -f'_x/f'_y$ exists in that interval. Hence, the vector function $\mathbf{r}(t) = (t, g(t))$ traverses the graph $y = g(x)$ and the level curve near $\mathbf{r}_0 = \mathbf{r}(t_0)$, where $t_0 = x_0$. It is smooth because $\mathbf{r}'(t) = (1, g'(t))$ so that $\mathbf{r}'(t) \neq \mathbf{0}$.

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* through that point. 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.

The case of functions of three variables can be analyzed along similar lines. Let a function $f(x, y, z)$ have continuous partial derivatives

in a neighborhood of $\mathbf{r}_0 = (x_0, y_0, z_0)$ such that $\nabla f(\mathbf{r}_0) \neq \mathbf{0}$. Consider a level surface of f through \mathbf{r}_0 , $f(x, y, z) = k$, which is a set of roots of the functions $F(x, y, z) = f(x, y, z) - k$. Since the components of the gradient $\nabla F(\mathbf{r}_0) = \nabla f(\mathbf{r}_0)$ do not vanish simultaneously, one can assume that, say, $F'_z(\mathbf{r}_0) = f'_z(\mathbf{r}_0) \neq 0$. By the implicit function theorem, there is a function $g(x, y)$ such that the graph $z = g(x, y)$ coincides with the level surface near \mathbf{r}_0 ; that is, $F(x, y, g(x, y)) = 0$ for all (x, y) near (x_0, y_0) . The function g has continuous partial derivatives, and its linearization at (x_0, y_0) defines a plane tangent to the graph $z = g(x, y)$ at \mathbf{r}_0 , where $z_0 = g(x_0, y_0)$. A normal of the tangent plane is $\mathbf{n} = (g'_x, g'_y, -1)$, where the derivatives are taken at the point (x_0, y_0) . Using $g'_x = -F'_x/F'_z = -f'_x/f'_z$ and $g'_y = -F'_y/F'_z = -f'_y/f'_z$, where the derivatives of f are taken at \mathbf{r}_0 , it follows that

$$\mathbf{n} = -\frac{1}{f'_z(\mathbf{r}_0)} \left(f'_x(\mathbf{r}_0), f'_y(\mathbf{r}_0), f'_z(\mathbf{r}_0) \right) = -\frac{1}{f'_z(\mathbf{r}_0)} \nabla f(\mathbf{r}_0).$$

Thus, the gradient $\nabla f(\mathbf{r}_0)$ is proportional to \mathbf{n} and hence *is also normal to the tangent plane*. Furthermore, if $\mathbf{r}(t) = (x(t), y(t), z(t))$ is a smooth curve on the level surface, that is, $f(\mathbf{r}(t)) = k$ for all values of t , then $df/dt = 0$ (the values of f do not change along the curve), and, by the chain rule,

$$df/dt = f'_x x' + f'_y y' + f'_z z' = \nabla f(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = 0.$$

So the gradient $\nabla f(\mathbf{r}_0)$ is orthogonal to a tangent vector to *any* such curve through \mathbf{r}_0 , and all tangent vectors to curves through \mathbf{r}_0 lie in the tangent plane to the level surface through \mathbf{r}_0 .

All these findings are summarized in the following theorem.

THEOREM 13.16. (Geometrical Properties of the Gradient).

Let f be differentiable at \mathbf{r}_0 . Let S be the level set through the point \mathbf{r}_0 , and assume $\nabla f(\mathbf{r}_0) \neq \mathbf{0}$. 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) If f has continuous partial derivatives on an open ball D containing \mathbf{r}_0 , then the portion of S inside D is a smooth surface (or curve), and ∇f is normal to S at \mathbf{r}_0 .

EXAMPLE 13.42. 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 because the components of $\nabla f = (2x, 4y, 6z)$ are continuous. 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 $2x + 2y + 3z = 9$. \square

Theorem 13.16 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. To this end, it is only noted that if $\mathbf{r}(t)$ is a smooth curve in the level set $f(\mathbf{r}) = k$ of a differentiable function, then f has a constant value along any such curve, and, by the chain rule, it follows that $df(\mathbf{r}(t))/dt = \nabla f(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = 0$ for any t . At any particular point $\mathbf{r}_0 = \mathbf{r}(t_0)$, all tangent vectors $\mathbf{r}'(t_0)$ to such curves through \mathbf{r}_0 are orthogonal to a *single* vector $\nabla f(\mathbf{r}_0)$. Intuitively, these vectors should form an $(m - 1)$ -dimensional Euclidean space (called a *tangent space* to the level surface at \mathbf{r}_0), just like all vectors in a plane in three-dimensional Euclidean space are orthogonal to a normal of the plane.

Remark. The gradient can be viewed as the result of the action of the operator $\nabla = (\partial/\partial x_1, \partial/\partial x_2, \dots, \partial/\partial x_m)$ on a function f . ∇f is understood in the sense of multiplication of the “vector” ∇ by a scalar f . With this notation, the differential operator d has a compact form $d = d\mathbf{r} \cdot \nabla$. The linearization $L(\mathbf{r})$ of $f(\mathbf{r})$ at \mathbf{r}_0 and the differentials of f also have a simple form for any number of variables:

$$L(\mathbf{r}) = f(\mathbf{r}_0) + \nabla f(\mathbf{r}_0) \cdot (\mathbf{r} - \mathbf{r}_0), \quad d^n f(\mathbf{r}) = (d\mathbf{r} \cdot \nabla)^n f(\mathbf{r}).$$

93.3. Study Problems.

Problem 13.12. (Differentiability and Directional Derivative).

Let $f(x, y) = y^3/(x^2 + y^2)$ if $(x, y) \neq (0, 0)$ and $f(0, 0) = 0$. Show that $D_{\mathbf{u}}f(0, 0)$ exists for any $\hat{\mathbf{u}}$, but it is not given by the relation (13.14). Show that this function is not differentiable at $(0, 0)$. Thus, the existence of all directional derivatives at a point does not imply differentiability at that point. In other words, despite that the function has a rate of change at a point in every direction, a good linear approximation may not exist at that point.

SOLUTION: Put $\hat{\mathbf{u}} = (\cos \theta, \sin \theta)$ for $0 \leq \theta < 2\pi$. By the definition of the directional derivative,

$$D_{\mathbf{u}}f(0, 0) = \lim_{h \rightarrow 0} \frac{f(h \cos \theta, h \sin \theta) - f(0, 0)}{h} = \lim_{h \rightarrow 0} \frac{h^3 \sin^3 \theta}{h^3} = \sin^3 \theta.$$

In particular, for $\theta = 0$, $\hat{\mathbf{u}} = (1, 0)$, and $D_{\mathbf{u}}f(0, 0) = f'_x(0, 0) = 0$; similarly for $\theta = \pi/2$, $\hat{\mathbf{u}} = (0, 1)$, and $D_{\mathbf{u}}f(0, 0) = f'_y(0, 0) = 1$. If the relation (13.14) were used, one would have found that $D_{\mathbf{u}}f(0, 0) = \sin \theta$, which contradicts the above result. If a good linear approximation exists, then it should be $L(x, y) = f'_x(0, 0)x + f'_y(0, 0)y = y$. But $(f(x, y) - L(x, y))/(x^2 + y^2)^{1/2} = -yx^2/(x^2 + y^2)^{3/2}$ does not vanish as $(x, y) \rightarrow (0, 0)$ because it has a nonzero constant value along any straight line $(x, y) = (t, at)$, $a \neq 0$. So f is not differentiable at the origin. \square

Problem 13.13. *Suppose that three level surfaces $f(x, y, z) = 1$, $g(x, y, z) = 2$, and $h(x, y, z) = 3$ are intersecting along a smooth curve C . Let P be a point on C in whose neighborhood f , g , and h have continuous partial derivatives and their gradients do not vanish at P . 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 . \square

Problem 13.14. (Energy Conservation in Mechanics).

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. \square

93.4. Exercises.

(1) Let $f(x, y)$ be a differentiable function. How would you specify the directions at a particular point in which the function does not change at all? How many such directions exist if the first partial derivatives do not vanish at that point? Answer the same questions for a function $f(x, y, z)$.

(2) For each of the following functions, find the gradient and the directional derivative at a specified point in the direction parallel to a given vector \mathbf{v} . Indicate whether the function increases or decreases in that direction.

(i) $f(x, y) = x^2y$, $(1, 2)$, $\mathbf{v} = (4, 5)$

(ii) $f(x, y) = x/(1 + xy)$, $(1, 1)$, $\mathbf{v} = (2, 1)$

(iii) $f(x, y, z) = x^2y - zy^2 + xz^2$, $(1, 2, -1)$, $\mathbf{v} = (1, -2, 2)$

(iv) $f(x, y, z) = \tan^{-1}(1 + x + y^2 + z^3)$, $(1, -1, 1)$, $\mathbf{v} = (1, 1, 1)$

(v) $f(x, y, z) = \sqrt{x + yz}$, $(1, 1, 3)$, $\mathbf{v} = (2, 6, 3)$

(vi) $f(x, y, z) = (x + y)/z$, $(2, 1, 1)$, $\mathbf{v} = (2, -1, -2)$

(3) Find the maximal and minimal rates of change of each of the following functions at a specified point and the directions in which they occur. Find the directions in which the function does not change.

(i) $f(x, y) = x/y^2$, $(2, 1)$

(ii) $f(x, y) = x^y$, $(2, 1)$

(iii) $f(x, y, z) = xz/(1 + yz)$, $(1, 2, 3)$

(iv) $f(x, y, z) = x \sin(yz)$, $(1, 2, \pi/3)$

(v) $f(x, y, z) = x^{y^z}$, $(2, 2, 1)$

(4) Let $f(x, y) = y/(1 + x^2 + y)$. Find all unit vectors $\hat{\mathbf{u}}$ along which the rate of change of f at $(2, -3)$ is a number $-1 \leq p \leq 1$ times the maximal rate of change of f at $(2, -3)$.

(5) For the function $f(x, y, z) = \frac{1}{2}x^2 - \frac{1}{2}y^2x + z^3y$ at the point $P_0(1, 2, -1)$ find:

(i) The maximal rate of change of f and the direction in which it occurs;

(ii) A direction in which the rate of change is half of the maximal rate of change. How many such directions exist?

(iii) The rate of change in the direction to the point $P_1(3, 1, 1)$

(6) If f and u are differentiable functions, prove that $\nabla f(u) = f'(u)\nabla u$.

(7) Find $\nabla \|\mathbf{c} \times \mathbf{r}\|^2$, where \mathbf{c} is a constant vector.

(8) If f , u , and v are differentiable functions, prove that $\nabla f(u, v) = f'_u \nabla u + f'_v \nabla v$.

(9) Find the directional derivative of $f(\mathbf{r}) = (x/a)^2 + (y/b)^2 + (z/c)^2$ at a point \mathbf{r} in the direction of \mathbf{r} . Find the points at which this derivative is equal to $\|\nabla f\|$.

(10) Find the angle between the gradients of $f = x/(x^2 + y^2 + z^2)$ at the points $(1, 2, 2)$ and $(-3, 1, 0)$.

(11) Let $f = z/\sqrt{x^2 + y^2 + z^2}$. Sketch the level surfaces of f and $\|\nabla f\|$. What is the significance of the level surfaces of $\|\nabla f\|$? Find the maximal and minimal values of f and $\|\nabla f\|$ in the region $1 \leq z \leq 2$.

(12) Let a curve C be defined as the intersection of the plane $\sin\theta(x - x_0) - \cos\theta(y - y_0) = 0$, where θ is a parameter, and the graph $z = f(x, y)$, where f is differentiable. Find $\tan\alpha$, where α is the angle between the tangent line to C at $(x_0, y_0, f(x_0, y_0))$ and the xy plane.

(13) Consider the function $f(x, y, z) = 2\sqrt{z + xy}$ and three points $P_0(1, 2, 2)$, $P_1(-1, 4, 1)$, and $P_2(-2, -2, 2)$. In which direction does f change faster at P_0 , toward P_1 or toward P_2 ? What is the direction in which f increases most rapidly at P_0 ?

(14) For the function $f(x, y, z) = xy + zy + zx$ at the point $P_0(1, -1, 0)$, find:

- (i) The maximal rate of change
- (ii) The rate of change in the direction $\mathbf{v} = (-1, 2, -2)$
- (iii) The angle θ between \mathbf{v} and the direction in which the maximal rate of f occurs

(15) Let $f(x, y, z) = x/(x^2 + y^2 + z^2)^{1/2}$. Find the rate of change of f in the direction of the tangent vector to the curve $\mathbf{r}(t) = (t, 2t^2, -2t^2)$ at the point $(1, 2, -2)$.

(16) Find the points at which the gradient of $f = x^3 + y^3 + z^3 - 3xyz$ is

- (i) Orthogonal to the z axis
- (ii) Parallel to the z axis
- (iii) Zero

(17) Let $f = \ln\|\mathbf{r} - \mathbf{r}_0\|$, where \mathbf{r}_0 is a fixed vector. Find points in space where $\|\nabla f\| = 1$.

(18) For each of the following surfaces, find the tangent plane and the normal line at a specified point:

- (i) $x^2 + y^2 + z^2 = 169$, $(3, 4, 12)$
- (ii) $x^2 - 2y^2 + z^2 + yz = 2$, $(2, 1, -1)$
- (iii) $x = \tan^{-1}(y/z)$, $(\pi/4, 1, 1)$
- (iv) $z = y + \ln(x/z)$, $(1, 1, 1)$
- (v) $2^{x/z} + 2^{y/z} = 8$, $(2, 2, 1)$
- (vi) $x^2 + 4y^2 + 3z^2 = 5$, $(1, -1/2, -1)$

(19) Find the points of the surface $x^2 + 2y^2 + 3z^2 + 2xy + 2zx + 4yz = 8$ at which the tangent planes are parallel to the coordinate planes.

(20) Find the tangent planes to the surface $x^2 + 2y^2 + 3z^2 = 21$ that are parallel to the plane $x + 4y + 6z = 0$.

(21) Find the points on the ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$ at which the normal line makes equal angles with the coordinate axes.

(22) Consider the paraboloid $z = x^2 + y^2$.

(i) Give the parametric equations of the normal line through a point $P_0(x_0, y_0, z_0)$ on the paraboloid;

(ii) Consider all normal lines through points with a fixed value of z_0 (say, $z_0 = 2$). Show that all such lines intersect at one single point that lies on the z axis and find the coordinates of this point.

(23) Find the points on the hyperboloid $x^2 - y^2 + 2z^2 = 5$, where the normal line is parallel to the line that joins the points $(3, -1, 0)$ and $(5, 3, 8)$.

(24) Find an equation of the plane tangent to the surface $x^2 + y^2 - 4z^2 = 1$ at a generic point (x_0, y_0, z_0) of the surface.

(25) Find the rate of change of the function $h(x, y) = \sqrt{10 - x^2 y^2}$ at the point $P_0(1, 1)$ in the direction toward the point $P(-2, 5)$. Let $h(x, y)$ be the height in a neighborhood of P_0 . Would you be climbing up or getting down when you go from P_0 toward P ?

(26) Your Mars rover is caught on the slope of a mountains by a dust storm. The visibility is 0. Your current position is $P_0(1, 2)$. You can escape in the direction of a cave located at $P_1(4, -2)$ or in the direction of the base located at $P_2(17, 14)$. Which way would you drive to avoid steep climbing or descending if the height in the area can be approximated by the function $h(x, y) = xy + x^2$?

(27) You are flying a small aircraft on the planet Weirdo. You have disturbed a nest of nasty everything-eating bugs. The onboard radar indicates that the concentration of the bugs is $C(x, y, z) = 100 - x^2 - 2y^2 - 3z^2$ and $C(x, y, z) = 0$ if $x^2 + 2y^2 + 3z^2 > 100$. If your current position is $(2, 3, 1)$, in which direction would you fire a mass-destruction microwave laser to kill as many poor bugs as possible near you? Find the optimal escape trajectory.

(28) Let two level curves $f(x, y) = 0$ and $g(x, y) = 0$ of functions f and g , whose partial derivatives are continuous, intersect at some point P_0 . The rate of change of the function f at P_0 along the curve $g(x, y) = 0$ is half of its maximal rate of change at P_0 . What is the angle at which the curves intersect (the angle between the tangent lines)?

(29) Suppose that the directional derivatives $D_{\mathbf{u}}f = a$ and $D_{\mathbf{v}}f = b$ of a differentiable function $f(x, y)$ are known at a particular point P_0 for two unit nonparallel vectors $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ that make the angles θ and ϕ with the x axis counted counterclockwise from the latter, respectively. Find the gradient of f at P_0 .

(30) Three tests of drilling into rock along the directions $\mathbf{u} = (1, 2, 2)$, $\mathbf{v} = (0, 4, 3)$, and $\mathbf{w} = (0, 0, 1)$ showed that the gold concentration

increases at the rates 3 g/m, 3 g/m, and 1 g/m, respectively. Assume that the concentration is a differentiable function. In what direction would you drill to maximize the gold yield and at what rate does the gold concentration increase in that direction? If the concentration is not differentiable, would you follow your previous finding about the drilling direction? Explain.

(31) A level surface of a differentiable function $f(x, y, z)$ contains the curves $\mathbf{r}_1(t) = (2+3t, 1-t^2, 3-4t+t^2)$ and $\mathbf{r}_2(t) = (1+t^2, 2t^3-1, 2t+1)$. Can this information be used to find the tangent plane to the surface at $(2, 1, 3)$? If so, find an equation of the plane.

(32) Prove that tangent planes to the surface $xyz = a^3 > 0$ and the coordinate planes form tetrahedrons of equal volumes.

(33) Prove the total length of intervals from the origin to the points of intersection of tangent planes to the surface $\sqrt{x} + \sqrt{y} + \sqrt{z} = \sqrt{a}$, $a > 0$, with the coordinate axes is constant.

(34) Two surfaces are called *orthogonal* at a point of intersection if the normal lines to the surfaces at that point are orthogonal. Show that the surfaces $x^2 + y^2 + z^2 = r^2$, $x^2 + y^2 = z^2 \tan^2 \phi$, and $y \cos \theta = x \sin \theta$ are pairwise orthogonal at their points of intersection for any values of the constants $r > 0$, $0 < \phi < \pi$, and $0 \leq \theta < 2\pi$.

(35) Find the directional derivative of $f(x, y, z)$ in the direction of the gradient of $g(x, y, z)$. What is the geometrical significance of this derivative?

(36) Find the angle at which the cylinder $x^2 + y^2 = a^2$ intersects the surface $bz = xy$ at a generic point of intersection (x_0, y_0, z_0) .

(37) A ray of light reflects from a mirrored surface at a point P just as it would reflect from the mirrored plane tangent to the surface at P (if the light travels along a vector \mathbf{u} , then the reflected light travels along a vector obtained from \mathbf{u} by reversing the direction of the component parallel to the normal to the surface). Show that the light coming from the top of the z axis and parallel to it will be focused by the parabolic mirror $az = x^2 + y^2$, $a > 0$, to a single point. Find its coordinates. This property of parabolic mirrors is used to design telescopes.

(38) Let $f(x, y) = y$ if $y \neq x^2$ and $f(x, y) = 0$ if $y = x^2$. Find $D_{\mathbf{n}}f(0, 0)$ for all unit vectors $\hat{\mathbf{n}}$. Show that $f(x, y)$ is not differentiable at $(0, 0)$. Is the function continuous at $(0, 0)$?

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 $B_\delta = \{\mathbf{r} \mid \|\mathbf{r} - \mathbf{r}_0\| < \delta\}$; 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 13.25. (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.

Minimal and maximal values are also called *extremum values*. In the one-variable case, Fermat's theorem asserts that if a differentiable function has a local extremum at x_0 , then its derivative vanishes at x_0 . The tangent line to the graph of f at x_0 is horizontal: $y = f(x_0) + df(x_0) = f(x_0) + f'(x_0)dx = f(x_0)$. There is an extension of Fermat's theorem.

THEOREM 13.17. (Necessary Condition for a Local Extremum)

If a differentiable function f has a local extremum at an interior point \mathbf{r}_0 of its domain D , then $df(\mathbf{r}_0) = 0$ or $\nabla f(\mathbf{r}_0) = 0$ (all partial derivatives of f vanish at \mathbf{r}_0).

PROOF. Consider a smooth curve $\mathbf{r}(t)$ through the point \mathbf{r}_0 such that $\mathbf{r}(t_0) = \mathbf{r}_0$. Then $d\mathbf{r}(t_0) = \mathbf{r}'(t_0)dt \neq \mathbf{0}$ (the curve is smooth and hence has a nonzero tangent vector). The function $F(t) = f(\mathbf{r}(t))$ defines the values of f along the curve. Therefore, $F(t)$ must have a local extremum at $t = t_0$. Since f is differentiable, the differential $dF(t_0) = F'(t_0)dt$ exists by the chain rule: $dF(t_0) = d\mathbf{r}(t_0) \cdot \nabla f(\mathbf{r}_0)$. By Fermat's theorem $dF(t_0) = 0$ and hence $d\mathbf{r}(t_0) \cdot \nabla f(\mathbf{r}_0) = 0$. This relation means that the vectors $d\mathbf{r}(t_0)$ and $\nabla f(\mathbf{r}_0)$ are orthogonal for all smooth curves through \mathbf{r}_0 . The only vector that is orthogonal to any vector is the zero vector, and the conclusion of the theorem follows: $\nabla f(\mathbf{r}_0) = \mathbf{0}$. \square

In particular, for a differentiable function f of two variables, this theorem states that the tangent plane to the graph of f at a local extremum is horizontal.

The converse of this theorem is not true. Let $f(x, y) = xy$. It is differentiable everywhere, and its partial derivatives are $f'_x = y$ and $f'_y = x$. They vanish at the origin, $\nabla f(0, 0) = \mathbf{0}$. However, the function has neither a local maximum nor a local minimum. Indeed, consider a straight line through the origin, $x = at$, $y = bt$. Then the values of f along the line are $F(t) = f(x(t), y(t)) = abt^2$. So $F(t)$ has a minimum at $t = 0$ if $ab > 0$ or a maximum if $ab < 0$. Each case is possible. For example, if $a = b = 1$, then $ab = 1 > 0$; if $a = -b = 1$, then $ab = -1 < 0$. Thus, f cannot have a local extremum at $(0, 0)$. The graph $z = xy$ is a hyperbolic paraboloid rotated through an angle $\pi/4$ about the z axis (see Example 11.29). It looks like a saddle. If the graph of $f(x, y)$ has a horizontal tangent plane at (x_0, y_0) and looks like a hyperbolic paraboloid in a small neighborhood of (x_0, y_0) , then the point (x_0, y_0) is called a *saddle point* (see Figure 13.12, right panel).

Remark. The above analysis might make the impression that if $\nabla f(\mathbf{r}_0) = \mathbf{0}$ and the values of f along any straight line through \mathbf{r}_0 have a local extremum (i.e., $F(t) = f(\mathbf{r}_0 + \mathbf{v}t)$ has either a local maximum or a local minimum at $t = 0$ for all vectors \mathbf{v}), then f has a local extremum at \mathbf{r}_0 . This conjecture is *false*! An example is given in Study Problem 13.16.

A local extremum may occur at a point at which the function is not differentiable. For example, $f(x, y) = |x| + |y|$ is continuous everywhere

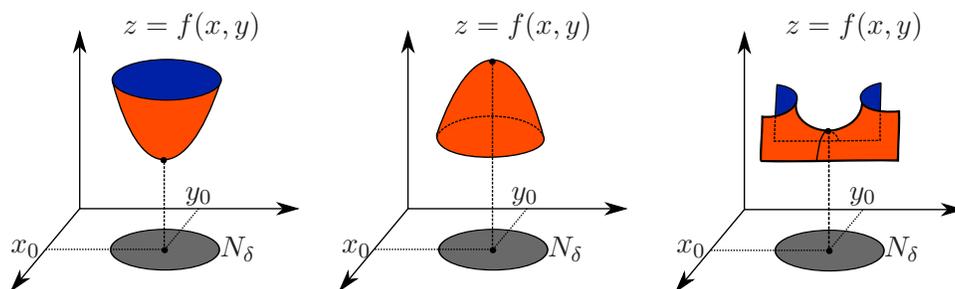


FIGURE 13.12. **Left:** The graph $z = f(x, y)$ near a local minimum of f . The values of f are no less than $f(x_0, y_0)$ for all (x, y) in a sufficiently small neighborhood N_δ of (x_0, y_0) . **Middle:** The graph $z = f(x, y)$ near a local maximum of f . The values of f do not exceed $f(x_0, y_0)$ for all (x, y) in a sufficiently small neighborhood N_δ of (x_0, y_0) . **Right:** The graph $z = f(x, y)$ near a saddle point of f . In a sufficiently small neighborhood N_δ of (x_0, y_0) , the values of f have a local maximum along some lines through (x_0, y_0) and a local minimum along the other lines through (x_0, y_0) .

and has an absolute minimum at $(0, 0)$. However, the partial derivatives $f'_x(0, 0)$ and $f'_y(0, 0)$ do not exist (e.g., $f'_x(x, y) = (|x|)'$, which is 1 if $x > 0$ and is -1 if $x < 0$, so $f'_x(0, y)$ does not exist and neither does $f'_x(0, 0)$).

DEFINITION 13.26. (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. Concavity. Recall from Calculus I that if the graph of a function $f(x)$ lies above all its tangent lines in an interval I , then f is concave upward on I . If the graph lies below all its tangent lines in I , then f is concave downward on I . Furthermore, if $f'(x_0) = 0$ (the tangent line is horizontal at x_0) and f is concave upward in small open interval I containing x_0 , then $f(x) > f(x_0)$ for all $x \neq x_0$ in I , and hence f has a local maximum. Similarly, f has a local minimum at x_0 , where $f'(x_0) = 0$, if it is concave downward in a neighborhood of x_0 . If the function f is twice differentiable on I , then it is concave upward if $f''(x) > 0$ on I and it is concave downward if $f''(x) < 0$ on I . The concavity test can be restated in the form of the second-order differential $d^2f(x) = f''(x)(dx)^2$, which is a function of two independent variables x and dx . If $d^2f(x) > 0$ for $dx \neq 0$, f is concave upward; if $d^2f(x) < 0$ for $dx \neq 0$, f is concave downward. Suppose that $f'(x_0) = 0$, $f''(x_0) \neq 0$, and f'' is continuous at x_0 . The continuity of f'' ensures that $d^2f(x)$ has the same sign as $d^2f(x_0)$ for all x near x_0 and all $dx \neq 0$. Hence, the graph of f has a fixed concavity in a neighborhood of x_0 . Thus, if $d^2f(x_0) < 0$ ($dx \neq 0$), then f has a local maximum at x_0 ; if $d^2f(x_0) > 0$ ($dx \neq 0$), then f has a local minimum at x_0 . It turns out that this sufficient condition for a function to have a local extremum has a natural extension to functions of several variables.

THEOREM 13.18. (Sufficient Condition for a Local Extremum).

Suppose that a function f has continuous second partial derivatives in an open ball containing a point \mathbf{r}_0 and $\nabla f(\mathbf{r}_0) = \mathbf{0}$. Then

$$\begin{array}{ll} f \text{ has a local maximum at } \mathbf{r}_0 & \text{if } d^2f(\mathbf{r}_0) < 0, \\ f \text{ has a local minimum at } \mathbf{r}_0 & \text{if } d^2f(\mathbf{r}_0) > 0 \end{array}$$

for all $d\mathbf{r}$ such that $\|d\mathbf{r}\| \neq 0$.

The proof of this theorem is omitted. However, an analogy can be made with the one-variable case. By the Taylor theorem (Theorem

13.15), values of a function f in a sufficiently small neighborhood of a point \mathbf{r}_0 are well approximated as $f(\mathbf{r}) = f(\mathbf{r}_0) + df(\mathbf{r}_0) + \frac{1}{2}d^2f(\mathbf{r}_0)$, where $d\mathbf{r} = \mathbf{r} - \mathbf{r}_0$. The first two terms define a linearization $L(\mathbf{r}) = f(\mathbf{r}_0) + df(\mathbf{r}_0)$ (or tangent plane) of f at \mathbf{r}_0 . Therefore,

$$f(\mathbf{r}) - L(\mathbf{r}) = \frac{1}{2}d^2f(\mathbf{r}_0),$$

where the contributions of terms smaller than $\|d\mathbf{r}\|^2$ have been neglected according to Theorem 13.15. This equation shows that if $d^2f(\mathbf{r}_0) < 0$ for all $\|d\mathbf{r}\| \neq 0$ (as a function of independent variables $d\mathbf{r}$), then the values of f are strictly less than the values of its linearization in a neighborhood $0 < \|\mathbf{r} - \mathbf{r}_0\| < \delta$ if $\delta > 0$ is small enough. The continuity of the second partial derivatives in a neighborhood of \mathbf{r}_0 ensures that, for all \mathbf{r} near \mathbf{r}_0 and all $d\mathbf{r}$, $d^2f(\mathbf{r})$ is a continuous function of two independent variables \mathbf{r} and $d\mathbf{r}$. Therefore, if $d^2f(\mathbf{r}_0) < 0$, $\|d\mathbf{r}\| \neq 0$, then $d^2f(\mathbf{r}) < 0$ for all \mathbf{r} near \mathbf{r}_0 . The values of f along *any* smooth curve through \mathbf{r}_0 have a local maximum at \mathbf{r}_0 if, in addition, $df(\mathbf{r}_0) = 0$. In the two-variable case, one can say that the graph of f is concave downward near \mathbf{r}_0 ; it looks like a paraboloid concave downward (see Figure 13.12, middle panel). The function has a local maximum. Similarly, if $d^2f(\mathbf{r}_0) > 0$ for all $\|d\mathbf{r}\| \neq 0$, then the graph of f lies *above* its tangent planes for all $0 < \|\mathbf{r} - \mathbf{r}_0\| < \delta$. The function has a local minimum at \mathbf{r}_0 . The graph of f near \mathbf{r}_0 looks like a paraboloid concave upward (see Figure 13.12, left panel).

94.3. Second-Derivative Test. The differential $d^2f(\mathbf{r}_0)$ is a homogeneous quadratic polynomial in the variables $d\mathbf{r}$. Its sign is determined by its coefficients, which are the second-order partial derivatives of f at \mathbf{r}_0 . The case of functions of two variables is discussed first.

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 13.19. (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 disk containing \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

- (1) If the roots are strictly positive, $\lambda_i > 0$, then f has a local minimum at \mathbf{r}_0 .
- (2) If the roots are strictly negative, $\lambda_i < 0$, then f has a local maximum at \mathbf{r}_0 .
- (3) If the roots do not vanish but have different signs, then f has neither a local maximum nor a local minimum at \mathbf{r}_0 .
- (4) If at least one of the roots vanishes, then f may have a local maximum, a local minimum, or none of the above (the second-derivative test is inconclusive).

In case (3), the critical point is said to be a *saddle point* of f and the graph of f crosses its tangent plane at \mathbf{r}_0 .

PROOF. Consider a rotation

$$(dx, dy) = (dx' \cos \phi - dy' \sin \phi, dy' \cos \phi + dx' \sin \phi).$$

Following the proof of Theorem 11.8 (classification of quadric cylinders), the second-order differential is written in the new variables (dx', dy') as

$$d^2 f(\mathbf{r}_0) = a(dx)^2 + 2c dx dy + b(dy)^2 = a'(dx')^2 + 2c' dx' dy' + b'(dy')^2,$$

$$a' = \frac{1}{2} \left(a + b + (a - b) \cos(2\phi) + 2c \sin(2\phi) \right),$$

$$b' = \frac{1}{2} \left(a + b - (a - b) \cos(2\phi) - 2c \sin(2\phi) \right),$$

$$2c' = 2c \cos(2\phi) - (a - b) \sin(2\phi).$$

The rotation angle is chosen so that $c' = 0$. Put $A^2 = (a - b)^2 + 4c^2$. If $\cos(2\phi) = (a - b)/A$ and $\sin(2\phi) = 2c/A$, then $c' = 0$. With this choice,

$$a' = \frac{1}{2}(a + b + A), \quad b' = \frac{1}{2}(a + b - A).$$

Next note that $a' + b' = a + b$ and $a'b' = \frac{1}{4}((a + b)^2 - A^2) = ab - c^2$. On the other hand, the roots of the quadratic equation $P_2(\lambda) = 0$ also satisfy the same conditions $\lambda_1 + \lambda_2 = a + b$ and $\lambda_1 \lambda_2 = ab - c^2$. Thus, $a' = \lambda_1$, $b' = \lambda_2$, and

$$d^2 f(\mathbf{r}_0) = \lambda_1(dx')^2 + \lambda_2(dy')^2.$$

If λ_1 and λ_2 are strictly positive, then $d^2 f(\mathbf{r}_0) > 0$ for all $(dx, dy) \neq (0, 0)$, and by Theorem 13.18 the function has a local minimum at \mathbf{r}_0 . If λ_1 and λ_2 are strictly negative, then $d^2 f(\mathbf{r}_0) < 0$ for all $(dx, dy) \neq (0, 0)$ and by Theorem 13.18 the function has a local maximum at \mathbf{r}_0 . If λ_1 and λ_2 do not vanish but have opposite signs, $\lambda_1 \lambda_2 < 0$, then in a neighborhood of \mathbf{r}_0 , the graph of f looks like

$$z = f(\mathbf{r}_0) + \lambda_1(x' - x'_0)^2 + \lambda_2(y' - y'_0)^2,$$

where the coordinates (x', y') are obtained from (x, y) by rotation through an angle ϕ . When λ_1 and λ_2 have different signs, this surface is a hyperbolic paraboloid (a saddle), and f has neither a local minimum nor a local maximum. Case (4) is easily proved by examples (see Study Problem 13.17). \square

COROLLARY 13.4. *Let the function f satisfy the hypotheses of Theorem 13.19. Put $D = ab - c^2$, where $a = f''_{xx}(\mathbf{r}_0)$, $b = f''_{yy}(\mathbf{r}_0)$, and $c = f''_{xy}(\mathbf{r}_0)$. Then*

- (1) *If $D > 0$ and $a > 0$ ($b > 0$), then $f(\mathbf{r}_0)$ is a local minimum.*
- (2) *If $D > 0$ and $a < 0$ ($b < 0$), then $f(\mathbf{r}_0)$ is a local maximum.*
- (3) *If $D < 0$, then $f(\mathbf{r}_0)$ is not a local extremum.*

This corollary is a simple consequence of the second-derivative test. Note that $\lambda_1\lambda_2 = D$. So $D < 0$ if $\lambda_1\lambda_2 < 0$ or \mathbf{r}_0 is a saddle point. Similarly, the conditions (1) and (2) are equivalent to the cases when λ_1 and λ_2 are strictly positive or negative, respectively.

EXAMPLE 13.43. *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 continuous partial derivatives everywhere of any order. So its critical points are solutions of the system of equations

$$\begin{cases} f'_x = x^2 + y^2 - 2x = 0 \\ f'_y = 2xy - 2y = 0 \end{cases}.$$

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.

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 $(2, 0)$. \square

EXAMPLE 13.44. Investigate the function $f(x, y) = e^{x^2-y}(5-2x+y)$ for extreme values.

SOLUTION: The function is defined on the whole plane, and, as the product of an exponential and a polynomial, it has continuous partial derivatives of any order. So its extreme values, if any, can be investigated by the second-derivative test.

Critical Points. Using the product rule for derivatives,

$$f'_x = e^{x^2-y} \left(2x(5-2x+y) - 2 \right) = 0 \quad \Rightarrow \quad x(5-2x+y) = 1,$$

$$f'_y = e^{x^2-y} \left((-1)(5-2x+y) + 1 \right) = 0 \quad \Rightarrow \quad 5-2x+y = 1.$$

The substitution of the second equation into the first one yields $x = 1$. Then it follows from the second equation that $y = -2$. So the function has just one critical point $(1, -2)$.

Second-Derivative Test. Using the product rule for derivatives,

$$f''_{xx} = (f'_x)'_x = e^{x^2-y} \left[2x \left(2x(5-2x+y) - 2 \right) + 2(5-2x+y) - 4 \right],$$

$$f''_{yy} = (f'_y)'_y = e^{x^2-y} \left[(-1) \left((-1)(5-2x+y) + 1 \right) - 1 \right],$$

$$f''_{xy} = (f'_y)'_x = e^{x^2-y} \left[2x \left((-1)(5-2x+y) + 1 \right) + 2 \right].$$

Therefore, $a = f''_{xx}(1, -2) = -2e^3$, $b = f''_{yy}(1, -2) = -e^3$, and $c = f''_{xy}(1, -2) = 2e^3$. Therefore, $D = ab - c^2 = -2e^6 < 0$. By Corollary 13.4, the only critical point is a saddle point. The function has no extreme values. \square

94.4. Study Problems.

Problem 13.15. Find all critical points of the function $f(x, y) = \sin(x) \sin(y)$ and determine whether they are a local maximum, a local minimum, a saddle point.

SOLUTION: The function has continuous partial derivatives of any order on the whole plane. So the second-derivative test applies to study critical points.

Critical Points. If n and m are integers, then

$$\begin{aligned} f'_x = \cos(x) \sin(y) = 0 &\Rightarrow x = \frac{\pi}{2} + \pi n \quad \text{or} \quad y = \pi m, \\ f'_y = \sin(x) \cos(y) = 0. & \end{aligned}$$

If $x = \frac{\pi}{2} + \pi n$, then it follows from the second equation that $y = \frac{\pi}{2} + \pi m$. If $y = \pi m$, then it follows from the second equation that $x = \pi n$. Thus, for any pair of integers n and m , the points $\mathbf{r}_{nm} = (\frac{\pi}{2} + \pi n, \frac{\pi}{2} + \pi m)$ and $\mathbf{r}'_{nm} = (\pi n, \pi m)$ are critical points of the function.

Second-Derivative Test has to be applied to each critical point. The second partial derivatives are

$$f''_{xx} = -\sin(x) \sin(y), \quad f''_{yy} = -\sin(x) \sin(y), \quad f''_{xy} = \cos(x) \cos(y)$$

For the critical points \mathbf{r}_{nm} , one has $a = f''_{xx}(\mathbf{r}_{nm}) = -(-1)^{n+m}$, $b = f''_{yy}(\mathbf{r}_{nm}) = -(-1)^{n+m} = a$, and $c = f''_{xy}(\mathbf{r}_{nm}) = 0$. The characteristic equation is $(a - \lambda)^2 = 0$ and hence $\lambda_1 = \lambda_2 = -(-1)^{n+m}$. If $n + m$ is even, then the roots are negative and $f(\mathbf{r}_{nm}) = 1$ is a local maximum. If $n + m$ is odd, then the roots are positive and $f(\mathbf{r}_{nm}) = -1$ is a local minimum. For the critical points \mathbf{r}'_{nm} , one has $a = f''_{xx}(\mathbf{r}'_{nm}) = 0$, $b = f''_{yy}(\mathbf{r}'_{nm}) = 0$, and $c = f''_{xy}(\mathbf{r}'_{nm}) = (-1)^{n+m}$. The characteristic equation $\lambda^2 - c^2 = \lambda^2 - 1 = 0$ has two roots $\lambda = \pm 1$ of opposite signs. Thus, \mathbf{r}'_{nm} are saddle points of f . In fact, the local extrema of this function are also its absolute extrema. \square

Problem 13.16. 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 $0 \leq \varphi \leq 2\pi$. 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 $0 \geq (a - b)^2 = a^2 - 2ab + b^2$ and hence $2ab \leq a^2 + b^2$ for any numbers a and b . Therefore, $4ab = 2ab + 2ab \leq 2ab + a^2 + b^2 = (a + b)^2$. 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,

$$\frac{4x^6y^2}{(x^4 + y^2)^2} \leq \frac{4x^6y^2}{4x^4y^2} = x^2 \rightarrow 0 \quad \text{as } (x, y) \rightarrow (0, 0).$$

Thus, $f(x, y) \rightarrow f(0, 0) = 0$ as $(x, y) \rightarrow (0, 0)$, 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 $\varphi \neq \pm\pi/2$ so that $\sin \varphi \neq 0$, 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. \square

Problem 13.17. *Suppose that $\lambda_1 = 0$ or $\lambda_2 = 0$ (or both) in the second-derivative test for a function f . Give examples of f when f has a local maximum, or a local minimum, or its graph looks like a saddle, or none of the above.*

SOLUTION: 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 $f(x, y) \geq 0$ for all (x, y) and f has a minimum at $(0, 0)$. Let $s = -1$. Then the curves $x = \pm y^2$ divide the plane into four sectors with the vertex at the critical point (the origin). In the sectors containing the x axis, $f(x, y) \geq 0$, whereas in the sectors containing the y axis, $f(x, y) \leq 0$. Thus, the graph of f has the shape of a saddle. The function $f(x, y) = -(x^2 + sy^4)$ has a maximum at $(0, 0)$ if $s > 0$.

If $s < 0$, the graph of f has the shape of a saddle. 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 function vanishes along the coordinate axes. So the plane is divided into four sectors (quadrants) in each of which the function has a fixed sign. The function is positive in the first and fourth quadrants ($x > 0$) and negative in the second and third quadrants ($x < 0$). It then follows that f has no maximum or minimum, and its graph does not have the shape of a saddle. Next, put $f(x, y) = x^2 - y^3$. It has a critical point $(0, 0)$ and $a = 2$, $b = 0$, and $c = 0$, that is, $\lambda_1 = 2$ and $\lambda_2 = 0$. The zeros of f form the curve $y = x^{2/3}$, which divides the plane into two parts so that f is negative above this curve and f is positive below it. Therefore, the graph of f does not have the shape of a saddle, and f does not have a minimum or maximum at $(0, 0)$. The behavior of the functions xy^2 and $x^2 - y^3$ near their critical point resembles the behavior of a function of one variable near its critical point that is also an *inflection point*. \square

94.5. Exercises.

(1) For each of the following functions, find all critical points and determine if they are a relative maximum, a relative minimum, or a saddle point:

- (i) $f(x, y) = x^2 + (y - 2)^2$
- (ii) $f(x, y) = x^2 - (y - 2)^2$
- (iii) $f(x, y) = (x - y + 1)^2$
- (iv) $f(x, y) = x^2 - xy + y^2 - 2x + y$
- (v) $f(x, y) = \frac{1}{3}x^3 + y^2 - x^2 - 3x - y + 1$
- (vi) $f(x, y) = x^2y^3(6 - x - y)$
- (vii) $f(x, y) = x^3 + y^3 - 3xy$
- (viii) $f(x, y) = x^4 + y^4 - x^2 - 2xy - 2y^2$
- (ix) $f(x, y) = xy + 50/x + 20/y, x > 0, y > 0$
- (x) $f(x, y) = x^2 + y^2 + \frac{1}{x^2y^2}$
- (xi) $f(x, y) = \cos(x) \cos(y)$
- (xii) $f(x, y) = \cos x + y^2$
- (xiii) $f(x, y) = y^3 + 6xy + 8x^3$
- (xiv) $f(x, y) = x^3 - 2xy + y^2$
- (xv) $f(x, y) = xy(1 - x - y)$

- (xvi) $f(x, y) = x \cos y$
 (xvii) $f(x, y) = xy\sqrt{1 - x^2/a^2 - y^2/b^2}$
 (xviii) $f(x, y) = (ax + by + c)/\sqrt{1 + x^2 + y^2}$, $a^2 + b^2 + c^2 \neq 0$
 (xix) $f(x, y) = (5x + 7y - 25)e^{-x^2 - xy - y^2}$
 (xx) $f(x, y) = \sin x + \sin y + \cos(x + y)$
 (xxi) $f(x, y) = \frac{1}{3}x^3 + xy^2 - x^2 - y^2$
 (xxii) $f(x, y) = \frac{1}{3}y^3 + xy + \frac{8}{3}x^3$
 (xxiii) $f(x, y) = x^2 + xy + y^2 - 4 \ln x - 10 \ln y$
 (xxiv) $f(x, y) = xy \ln(x^2 + y^2)$
 (xxv) $f(x, y) = x + y + \sin(x) \sin(y)$
 (xxvi) $f(x, y) = \sin(x) + \cos(y) + \cos(x - y)$
 (xxvii) $f(x, y) = x - 2y + \ln(\sqrt{x^2 + y^2}) + 3 \tan^{-1}(y/x)$

(2) Let the function $z = z(x, y)$ be defined implicitly by the given equation. Use the implicit differentiation to find extreme values of $z(x, y)$:

- (i) $x^2 + y^2 + z^2 - 2x + 2y - 4z - 10 = 0$
 (ii) $x^2 + y^2 + z^2 - xz - yz + 2x + 2y + 2z - 2 = 0$
 (iii) $(x^2 + y^2 + z^2)^2 = a^2(x^2 + y^2 - z^2)$

95. Maximum and Minimum Values (Continued)

95.1. Second-Derivative Test for Multivariable Functions. Theorem 13.18 holds for any number of variables, and there is a multivariable analog of the second-derivative test (Theorem 13.19). As in the two-variable case, the numbers $f''_{x_i x_j}(\mathbf{r}_0) = D_{ij}$ can be arranged into an $m \times m$ matrix. By Clairaut's theorem, this matrix is symmetric $D_{ij} = D_{ji}$. 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. The following facts are established by methods of linear algebra:

- (1) The characteristic polynomial of a real symmetric $m \times m$ matrix has m real roots $\lambda_1, \lambda_2, \dots, \lambda_m$ (a root of multiplicity k counted k times).

(2) There exists a rotation

$$d\mathbf{r} = (dx_1, dx_2, \dots, dx_m) \rightarrow d\mathbf{r}' = (dx'_1, dx_2, \dots, dx'_m),$$

which is a linear homogeneous transformation that preserves the length $\|d\mathbf{r}\| = \|d\mathbf{r}'\|$, such that

$$\begin{aligned} d^2f(\mathbf{r}_0) &= \sum_{i=1}^m \sum_{j=1}^m f''_{x_i x_j}(\mathbf{r}_0) dx_i dx_j = \sum_{i=1}^m \sum_{j=1}^m D_{ij} dx_i dx_j \\ &= \lambda_1(dx'_1)^2 + \lambda_2(dx'_2)^2 + \cdots + \lambda_n(dx'_m)^2. \end{aligned}$$

(3) The roots of the characteristic polynomial satisfy the conditions:

$$\begin{aligned} \lambda_1 + \lambda_2 + \cdots + \lambda_m &= D_{11} + D_{22} + \cdots + D_{mm}, \\ \lambda_1 \lambda_2 \cdots \lambda_m &= \det D. \end{aligned}$$

Fact (2) implies that if all roots of the characteristic polynomial are strictly positive, then $d^2f(\mathbf{r}_0) > 0$ for all $\|d\mathbf{r}\| = \|d\mathbf{r}'\| \neq 0$, and hence $f(\mathbf{r}_0)$ is a local minimum by Theorem 13.18. Similarly, if all the roots are strictly negative, then $f(\mathbf{r}_0)$ is a local maximum. Corollary 13.4 follows from fact (2) for $m = 2$. For $m > 2$, these properties of the roots are insufficient to establish a multivariable analog of Corollary 13.4. Fact (3) also implies that if $\det D = 0$, then one or more roots are 0. Hence, $d^2f(\mathbf{r}_0) = 0$ for some $d\mathbf{r} \neq \mathbf{0}$, and the hypotheses of Theorem 13.18 are not fulfilled.

THEOREM 13.20. (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 in some open ball centered at \mathbf{r}_0 . Let λ_i , $i = 1, 2, \dots, m$, be roots of the characteristic polynomial $P_m(\lambda)$ of the matrix of second derivatives $D_{ij} = f''_{x_i x_j}(\mathbf{r}_0)$.

- (1) If all the roots are strictly positive, $\lambda_i > 0$, then f has a local minimum.
- (2) If all the roots are strictly negative, $\lambda_i < 0$, then f has a local maximum.
- (3) If all the roots do not vanish but have different signs, then f has no local minimum or maximum at \mathbf{r}_0 .
- (4) If some of the roots vanish, then f may have a local maximum, or a local minimum, or none of the above (the test is inconclusive).

In case (3), the difference $f(\mathbf{r}) - f(\mathbf{r}_0)$ changes its sign in a neighborhood of \mathbf{r}_0 . It is an m -dimensional analog of a *saddle point*. Case (4) holds if $\det D = 0$, which is easy to verify. In general, roots of

$P_m(\lambda)$ are found numerically. If some of the roots are guessed, then a synthetic division can be used to reduce the order of the equation. If $P_m(\lambda_1) = 0$, then there is a polynomial Q_{m-1} of degree $m - 1$ such that $P_m(\lambda) = (\lambda - \lambda_1)Q_{m-1}(\lambda)$ so that the other roots satisfy $Q_{m-1}(\lambda) = 0$. The signs of the roots can also be established by a graphical method (an example is given Study Problem 13.18).

EXAMPLE 13.45. Investigate the function $f(x, y, z) = \frac{1}{3}x^3 + \frac{1}{2}y^2 + z^2 + xy + 2z$ for extreme values.

SOLUTION: The function is a polynomial so it has continuous partial derivatives of any order everywhere. So its critical points satisfy the equations:

$$\begin{cases} f'_x = x^2 + y = 0 \\ f'_y = y + x = 0 \\ f'_z = 2z + 2 = 0 \end{cases} \Leftrightarrow \begin{cases} x^2 = x \\ y = -x \\ z = -1 \end{cases}.$$

The first equation has two solutions $x = 0$ and $x = 1$. So the function has two critical points $\mathbf{r}_1 = (0, 0, -1)$ and $\mathbf{r}_2 = (1, -1, -1)$. The second-order partial derivatives are

$$f''_{xx} = 2x, \quad f''_{xy} = f''_{yy} = 1, \quad f''_{xz} = f''_{yz} = 0, \quad f''_{zz} = 2.$$

For the critical point \mathbf{r}_1 , the characteristic polynomial

$$\det \begin{pmatrix} -\lambda & 1 & 0 \\ 1 & 1 - \lambda & 0 \\ 0 & 0 & 2 - \lambda \end{pmatrix} = (2 - \lambda)(\lambda^2 - \lambda - 1)$$

has the roots 2 and $(1 \pm \sqrt{5})/2$. They do not vanish but have different signs. So \mathbf{r}_1 is a saddle point of f (no extreme value). For the critical point \mathbf{r}_2 , the characteristic polynomial

$$\det \begin{pmatrix} 2 - \lambda & 1 & 0 \\ 1 & 1 - \lambda & 0 \\ 0 & 0 & 2 - \lambda \end{pmatrix} = (2 - \lambda)(\lambda^2 - 3\lambda + 1)$$

has positive roots $2 > 0$ and $(3 \pm \sqrt{5})/2 > 0$. So $f(1, -1, -1) = -7/6$ is a local minimum. \square

95.2. When the Second-Derivative Test Is Inconclusive. 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 has continuous partial derivative of sufficiently high orders in a neighborhood of a critical point \mathbf{r}_0 , then the Taylor theorem provides a useful

technique for answering this question. The local behavior of a function near \mathbf{r}_0 is determined by higher-order differentials $d^n f(\mathbf{r})$, where $d\mathbf{r} = \mathbf{r} - \mathbf{r}_0$. It is generally easier to study the concavity of a polynomial than that of a general function. The concept is illustrated by the following example.

EXAMPLE 13.46. Investigate a local behavior of the function $f(x, y) = \sin(xy)/(xy)$ if $x \neq 0$ and $y \neq 0$ and $f(0, y) = f(x, 0) = 1$.

SOLUTION: Since $u = xy$ is small near the origin, by the Taylor theorem $\sin u = u - u^3/6 + \varepsilon(u)u^3$, where $\varepsilon(u) \rightarrow 0$ as $u \rightarrow 0$. Therefore,

$$f(x, y) = 1 - \frac{u^2}{6} (1 - 6\varepsilon(u)) = 1 - \frac{x^2 y^2}{6} (1 - 6\varepsilon(xy)).$$

In a disk $x^2 + y^2 < \delta^2$ of a sufficiently small radius $\delta > 0$, $1 - 6\varepsilon(xy) > 0$ because $\varepsilon(xy) \rightarrow 0$ as $(x, y) \rightarrow (0, 0)$. So the function attains a local maximum at $(0, 0)$ because $x^2 y^2 \geq 0$ in the disk. The inequality $|\sin u| \leq |u|$ suggests that f attains the maximum value 1 also along the coordinate axes (critical points are not isolated). The established local behavior of the function implies that its second partial derivatives vanish at $(0, 0)$, and hence both roots of the characteristic polynomial $P_2(\lambda) = \lambda^2$ vanish (the second-derivative test is inconclusive). \square

95.3. 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 (see Calculus I). 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 $x = 2$. The function is differentiable for all x , and therefore its critical points are determined by $f'(x) = 2x = 0$. 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 and compared not only at the critical points but also at the boundaries of the interval.

The situation for multivariable functions is similar. For example, the function $f(x, y) = x^2 + y^2$ whose arguments are restricted to the square $D = [0, 1] \times [0, 1]$ attains its absolute maximum and minimum values on the boundary of D as shown in the left panel of Figure 13.13.

DEFINITION 13.27. (Closed Set).

A set D in a Euclidean space is said to be closed if it contains all its limit points.

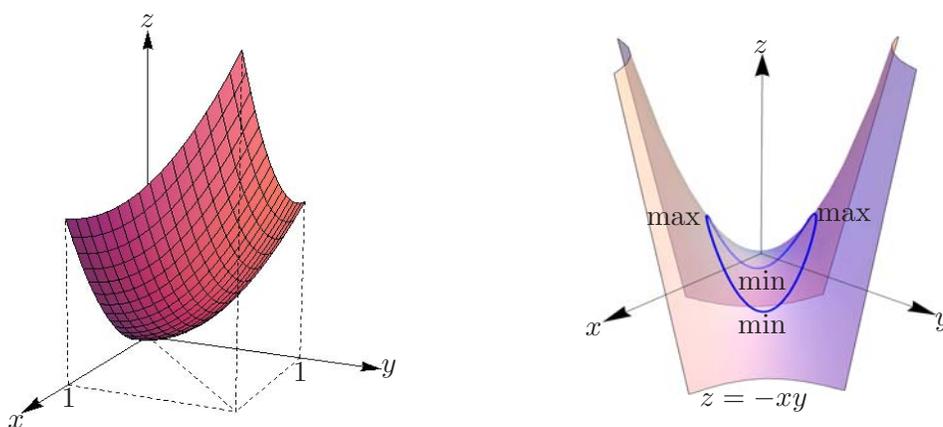


FIGURE 13.13. **Left:** The graph $z = x^2 + y^2$ (a circular paraboloid) over the square $D = [0, 1] \times [0, 1]$. The function $f(x, y) = x^2 + y^2$ attains its absolute maximum and minimum values on the boundary of D : $f(0, 0) \leq f(x, y) \leq f(1, 1)$ for all points in D . **Right:** The graph $z = -xy$. The values of $f(x, y) = -xy$ along the circle $x^2 + y^2 = 4$ are shown by the curve on the graph. The function has two local maxima and minima on the disk $x^2 + y^2 \leq 4$, while it has no maximum and minimum values on the entire plane.

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 13.28. (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 13.21. (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 $\mathbf{r}_2 \in D$.

The closedness of D is essential. For example, if the function $f(x, y) = x^2 + y^2$ is restricted to the *open* square $D = (0, 1) \times (0, 1)$, then it has no extreme values on D . For all (x, y) in D , $f(0, 0) < f(x, y) < f(1, 1)$ and there are points in D arbitrarily close to $(0, 0)$ and $(1, 1)$, but the points $(0, 0)$ and $(1, 1)$ are not in D . So f takes values on D arbitrarily close to 0 and 2, never reaching them. The boundedness of D is also crucial. For example, if the function $f(x, y) = x^2 + y^2$ is restricted to the first quadrant, $x \geq 0$ and $y \geq 0$, then f has no maximum value on D . It should be noted that the continuity of f and the closedness and boundedness of D are *sufficient* (not necessary) conditions for f to attain its absolute extreme values on D . There are noncontinuous or continuous functions on an unbounded or non-closed region D (or both) that attain their extreme values on D . Such examples are given in Calculus I, and functions of a single variable may always be viewed as a particular case of functions of two or more variables: $f(x, y) = g(x)$.

By the extreme value theorem, it follows that the points \mathbf{r}_1 and \mathbf{r}_2 are either critical points of f (because local extrema always occur at critical points) 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 13.47. 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 has continuous partial derivatives of any order on the whole plane. The disk $x^2 + y^2 \leq 4$ is a closed set in the plane. So the hypotheses of the extreme value theorem are fulfilled.

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 $t \in [0, 2\pi]$. The values of the

function on the boundary are $F(t) = f(x(t), y(t)) = 4 + 4 \cos t \sin t = 4 + 2 \sin(2t)$. 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 maximum 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 $(0, 0)$. \square

EXAMPLE 13.48. 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 has continuous partial derivatives of any order on the whole space.

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 $S_1 = \{(x, y, z) \mid z = 4, x^2 + y^2 \leq 4\}$ in the plane $z = 4$ and the portion of the paraboloid $S_2 = \{(x, y, z) \mid z = x^2 + y^2, x^2 + y^2 \leq 4\}$. 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, $x^2 + y^2 \leq 4$. 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 $r^2 = x^2 + y^2 \leq 4$. 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 $r^2 = x^2 + y^2 \leq 4$ 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 extreme values of f 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 $z = x^2 + y^2$. \square

95.4. Study Problems.

Problem 13.18. Investigate the extreme values of the function $f(x, y, z) = x + y^2/(4x) + z^2/y + 2/z$ if $x > 0$, $y > 0$, and $z > 0$.

SOLUTION: The critical points are determined by the equations

$$f'_x = 1 - \frac{y^2}{4x^2} = 0, \quad f'_y = \frac{y}{2x} - \frac{z^2}{y^2} = 0, \quad f'_z = \frac{2z}{y} - \frac{2}{z^2} = 0.$$

The first equation is equivalent to $y = 2x$ (since $x > 0$ and $y > 0$). The substitution of this relation into the second equation gives $z = y$ because $y > 0$ and $z > 0$. The substitution of this relation into the third equation yields $z = 1$ as $z > 0$. There is only one critical point $\mathbf{r}_0 = (\frac{1}{2}, 1, 1)$ in the positive octant. The second partial derivatives at \mathbf{r}_0 are

$$\begin{aligned} f''_{xx}(\mathbf{r}_0) &= \frac{y^2}{2x^3} \Big|_{\mathbf{r}_0} = 4, & f''_{xy}(\mathbf{r}_0) &= -\frac{y}{2x^2} \Big|_{\mathbf{r}_0} = -2, \\ f''_{xz}(\mathbf{r}_0) &= 0, & f''_{yy}(\mathbf{r}_0) &= \frac{1}{2x} + \frac{2z^2}{y^3} \Big|_{\mathbf{r}_0} = 3, \\ f''_{yz}(\mathbf{r}_0) &= -\frac{2z}{y^2} \Big|_{\mathbf{r}_0} = -2, & f''_{zz}(\mathbf{r}_0) &= \frac{2}{y} + \frac{4}{z^3} \Big|_{\mathbf{r}_0} = 6. \end{aligned}$$

The characteristic equation of the second derivative matrix is

$$\begin{aligned} \det \begin{pmatrix} 4 - \lambda & -2 & 0 \\ -2 & 3 - \lambda & -2 \\ 0 & -2 & 6 - \lambda \end{pmatrix} &= (4 - \lambda)[(3 - \lambda)(6 - \lambda) - 4] - 4(6 - \lambda) \\ &= -\lambda^3 + 13\lambda^2 - 46\lambda + 32 = 0. \end{aligned}$$

First of all, $\lambda = 0$ is not a root. To analyze the signs of the roots, the following method is employed. The characteristic equation is written in the form

$$\lambda(\lambda^2 - 13\lambda + 46) = 32.$$

This equation determines the points of intersection of the graph $y = g(\lambda) = \lambda(\lambda^2 - 13\lambda + 46)$ with the horizontal line $y = 32$. The polynomial $g(\lambda)$ has one simple root $g(0) = 0$ because the quadratic equation $\lambda^2 - 13\lambda + 46 = 0$ has no real roots. Therefore, $g(\lambda) > 0$ if $\lambda > 0$ and $g(\lambda) < 0$ if $\lambda < 0$. This implies that the intersection of the horizontal line $y = 32 > 0$ with the graph $y = g(\lambda)$ occurs only for $\lambda > 0$. Thus,

all three roots of the characteristic polynomial $P_3(\lambda)$ are positive, and hence $f(1/2, 1, 1) = 4$ is a minimum. \square

Problem 13.19. (The Least Squares Method).

Suppose that a scientist has a reason to believe that two quantities x and y are related linearly, $y = mx + b$, where m and b are unknown constants. The scientist performs an experiment and collects data as points on the plane (x_i, y_i) , $i = 1, 2, \dots, N$. Since the data contain errors, the points do not lie on a straight line. Let $d_i = y_i - (mx_i + b)$ be the vertical deviation of the point (x_i, y_i) from the line $y = mb + x$. The method of least squares determines the constants m and b by demanding that the sum of squares $\sum d_i^2$ attains its minimal value, thus providing the "best" fit to the data points. Find m and b .

SOLUTION: Consider the function $f(m, b) = \sum_{i=1}^N d_i^2$. Its critical points satisfy the equations $f'_b = -2 \sum_{i=1}^N d_i = 0$ and $f'_m = -2 \sum_{i=1}^N x_i d_i = 0$ because $(d_i)'_b = -1$ and $(d_i)'_m = -x_i$. The substitution of the explicit form of d_i into these equation, yields the following system:

$$m \sum_{i=1}^N x_i + bN = \sum_{i=1}^N y_i, \quad m \sum_{i=1}^N x_i^2 + b \sum_{i=1}^N x_i = \sum_{i=1}^N x_i y_i$$

whose solution determines the slope m and the constant b of the least squares linear fit to the data points. Note that the second-derivative test here is not really necessary to conclude that f has a minimum at the critical point. Explain why! \square

95.5. Exercises.

(1) For each of the following functions, find all critical points and determine if they are a relative maximum, a relative minimum, or a saddle point:

- (i) $f(x, y, z) = x^2 + y^2 + z^2 + 2x + 4y - 8z$
- (ii) $f(x, y, z) = x^2 + y^3 + z^2 + 12xy - 2z$
- (iii) $f(x, y, z) = x^2 + y^3 + z^2 + 12xy + 2z$
- (iv) $f(x, y, z) = \sin x + z \sin y$
- (v) $f(x, y, z) = x^2 + \frac{5}{3}y^3 + z^2 - 2xy - 4zy$
- (vi) $f(x, y, z) = x + y^2/(4x) + z^2/y + 2/z$
- (vii) $f(x, y, z) = a^2/x + x^2/y + y^2/z + z^2/b$, $x > 0$, $y > 0$, $z > 0$, $b > 0$
- (viii) $f(x, y, z) = \sin x + \sin y + \sin z + \sin(x+y+z)$, where $(x, y, z) \in [0, \pi] \times [0, \pi] \times [0, \pi]$
- (ix) $f(x_1, \dots, x_m) = \sum_{k=1}^m \sin x_k$
- (x) $f(\mathbf{r}) = (R^2 - \|\mathbf{r}\|^2)^2$, where $\mathbf{r} = (x_1, \dots, x_m)$ and R is a constant

$$(xi) \quad f(x_1, \dots, x_m) = x_1 + x_2/x_1 + x_3/x_2 + \cdots + x_m/x_{m-1} + 2/x_m, \\ x_i > 0, \quad i = 1, 2, \dots, m$$

(2) Given two positive numbers a and b , find m numbers x_i , $i = 1, 2, \dots, m$, between a and b so that the ratio

$$\frac{x_1 x_2 \cdots x_m}{(a + x_1)(x_1 + x_2) \cdots (x_m + b)}$$

is maximal.

(3) Use multivariable Taylor polynomials to show that the origin is a critical point of each of the following functions. Determine if there is a local maximum, a local minimum, or none of the above.

- (i) $f(x, y) = x^2 + xy^2 + y^4$
- (ii) $f(x, y) = \ln(1 + x^2 y^2)$
- (iii) $f(x, y) = x^2 \ln(1 + x^2 y^2)$
- (iv) $f(x, y) = xy(\cos(x^2 y) - 1)$
- (v) $f(x, y) = (x^2 + 2y^2) \tan^{-1}(x + y)$
- (vi) $f(x, y) = \cos(e^{xy} - 1)$
- (vii) $f(x, y) = \ln(y^2 \sin^2 x + 1)$
- (viii) $f(x, y) = e^{x+y^2} - 1 - \sin(x - y^2)$
- (ix) $f(x, y, z) = \sin(xy + z^2)/(xy + z^2)$
- (x) $f(x, y, z) = 2 - 2 \cos(x + y + z) - x^2 - y^2 - z^2$

(4) Let $f(x, y, z) = xy^2 z^3(a - x - 2y - 3z)$, $a > 0$. Find all its critical points and determine if they are a relative maximum, a relative minimum, or a saddle point.

(5) Give examples of a function $f(x, y)$ of two variables on a region D that attains its extreme values and has the following properties:

- (i) f is continuous on D , and D is not closed.
- (ii) f is not continuous on D , and D is bounded and closed.
- (iii) f is not continuous on D , and D is not bounded and not closed.

Do such examples contradict the extreme value theorem? Explain.

(6) For each of the following functions, find the extreme values on the specified set D :

- (i) $f(x, y) = 1 + 2x - 3y$, D is the closed triangle with vertices $(0, 0)$, $(1, 2)$, and $(2, 4)$
- (ii) $f(x, y) = x^2 + y^2 + xy^2 - 1$, $D = \{(x, y) \mid |x| \leq 1, |y| \leq 2\}$
- (iii) $f(x, y) = yx^2$, $D = \{(x, y) \mid x \geq 0, y \geq 0, x^2 + y^2 \leq 4\}$
- (iv) $f(x, y, z) = xy^2 + z$, $D = \{(x, y, z) \mid |x| \leq 1, |y| \leq 1, |z| \leq 1\}$
- (v) $f(x, y, z) = xy^2 + z$, $D = \{(x, y, z) \mid 1 \leq x^2 + y^2 \leq 4, -4 - x \leq z \leq -4 + x\}$

- (7) Find the point on the plane $x + y - z = 1$ that is closest to the point $(1, 2, 3)$. *Hint:* Let the point in question have the coordinates (x, y, z) . Then the squared distance between it and $(1, 2, 3)$ is $f(x, y, z) = (x - 1)^2 + (y - 2)^2 + (z - 3)^2$, where $z = x + y - 1$ because (x, y, z) is in the plane.
- (8) Find the point on the cone $z^2 = x^2 + y^2$ that is closest to the point $(1, 2, 3)$.
- (9) Find an equation of the plane that passes through the point $(3, 2, 1)$ and cuts off the smallest volume in the first octant.
- (10) Find the extreme values of $f(x, y) = ax^2 + 2bxy + cy^2$ on the circle $x^2 + y^2 = 1$.
- (11) Find the extreme values of $f(x, y, z) = x^2/a^2 + y^2/b^2 + z^2/c^2$ on the sphere $x^2 + y^2 + z^2 = 1$.
- (12) Find two positive numbers whose product is fixed, while the sum of their reciprocals is minimal.
- (13) Find m positive numbers whose product is fixed, while the sum of their reciprocals is minimal.
- (14) A solid object consists of a solid cylinder and a solid circular cone such that the base of the cone coincides with the base of the cylinder. If the total surface area of the object is fixed, what are the dimensions of the cone and cylinder at which the object has maximal volume?
- (15) Find a linear approximation $y = mx + b$ to the parabola $y = x^2$ such that the deviation $\Delta = \max |x^2 - mx - b|$ is minimal in the interval $1 \leq x \leq 3$.
- (16) Let N points of masses m_j , $j = 1, 2, \dots, N$, be positioned in a plane at $P_j = (x_j, y_j)$. Recall from Calculus II that the moment of inertia of this system about a point $P = (x, y)$ is

$$I(x, y) = \sum_{j=1}^N m_j |PP_j|^2.$$

Find P about which the moment of inertia is minimal.

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 13.49. *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 along 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 minimum value at $t = \pi/4$ and $t = \pi/4 + \pi$ and its absolute maximum 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 $(-\sqrt{2}, -\sqrt{2})$. The solution is illustrated in the right panel of Figure 13.13. \square

However, in many similar problems, 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 $g_a(\mathbf{r}) = 0$, $a = 1, 2, \dots, M$? Naturally, the number of independent constraints should not exceed the number of variables.

DEFINITION 13.29. (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 $\mathbf{r} = (x, y)$. Let \mathbf{r}_0 be a point at which $f(\mathbf{r})$ has a local extremum on the set S defined by the constraint $g(\mathbf{r}) = 0$. Let us assume that the function g has continuous partial derivatives in a neighborhood of \mathbf{r}_0 and $\nabla g(\mathbf{r}_0) \neq \mathbf{0}$. Then the equation $g(\mathbf{r}) = 0$ defines a *smooth* curve through the point \mathbf{r}_0 (recall the argument given before Theorem 13.16). Let $\mathbf{r}(t)$ be parametric equations of this curve in a neighborhood of \mathbf{r}_0 , that is, for some $t = t_0$, $\mathbf{r}(t_0) = \mathbf{r}_0$. The function $F(t) = f(\mathbf{r}(t))$ defines values of f along the curve and has a local extremum at t_0 . Let f be differentiable at \mathbf{r}_0 and $\nabla f(\mathbf{r}_0) \neq \mathbf{0}$. Since the curve is smooth, the vector function $\mathbf{r}(t)$ is differentiable, and it is concluded that F has no rate of change at $t = t_0$, $F'(t_0) = 0$. The chain rule yields

$$\begin{aligned} F'(t_0) &= f'_x(\mathbf{r}_0)x'(t_0) + f'_y(\mathbf{r}_0)y'(t_0) \\ &= \nabla f(\mathbf{r}_0) \cdot \mathbf{r}'(t_0) = 0 \implies \nabla f(\mathbf{r}_0) \perp \mathbf{r}'(t_0), \end{aligned}$$

provided $\mathbf{r}'(t_0) \neq \mathbf{0}$ (*the curve is smooth*). The gradient $\nabla f(\mathbf{r}_0)$ is orthogonal to a tangent vector to the curve at *the point where f has a local extremum on the curve*. By Theorem 13.16, 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 , *provided* $\nabla g(\mathbf{r}_0) \neq \mathbf{0}$. Therefore, the gradients $\nabla f(\mathbf{r}_0)$ and $\nabla g(\mathbf{r}_0)$ must be parallel at \mathbf{r}_0 (see Figure 13.14). The characteristic geometrical property of the point \mathbf{r}_0 is that *the level curve of f and the curve $g(x, y) = 0$ intersect at \mathbf{r}_0 and are tangential to one another*. For this very reason, f has no rate of change along the curve $g(x, y) = 0$ 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 13.22. (Critical Points Subject to a Constraint).

Suppose that f has a local extreme value at a point \mathbf{r}_0 on the set defined by a constraint $g(\mathbf{r}) = 0$. Suppose that g has continuous partial derivatives in a neighborhood of \mathbf{r}_0 and $\nabla g(\mathbf{r}_0) \neq \mathbf{0}$. If f is differentiable at

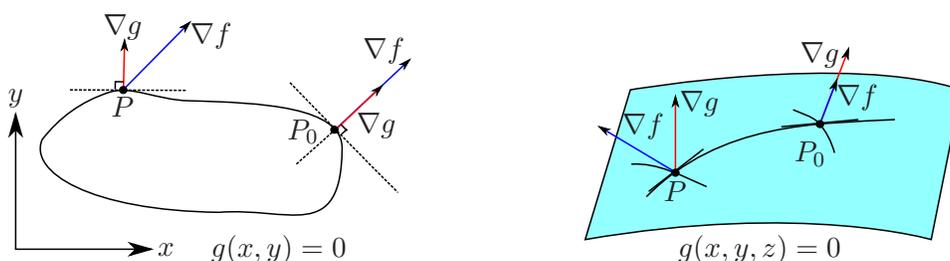


FIGURE 13.14. **Left:** Relative orientations of the gradients ∇f and ∇g along the curve $g(x, y) = 0$. At the point P_0 , the function f has a local extreme value along the curve $g = 0$. At this point, the gradients are parallel, and the level curve of f through P_0 and the curve $g = 0$ have a common tangent line. **Right:** Relative orientations of the gradients ∇f and ∇g along any curve in the constraint surface $g(x, y, z) = 0$. At the point P_0 , the function f has a local extreme value on the surface $g = 0$. At this point, the gradients are parallel, and the level surface of f through P_0 and the surface $g = 0$ have a common tangent plane.

\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 . On the other hand, by the properties of the gradient, the vector $\nabla g(\mathbf{r}_0)$ is orthogonal to $\mathbf{r}'(t_0)$ for every such curve. 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 critical points of f subject to a constraint $g = 0$. It is called the *method of Lagrange multipliers*. To find the critical points of f , the following system of equations must be solved:

$$(13.17) \quad \nabla f(\mathbf{r}) = \lambda \nabla g(\mathbf{r}), \quad g(\mathbf{r}) = 0.$$

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 13.50. *Use the method of Lagrange multipliers to solve the problem in Example 13.33.*

SOLUTION: Put $g(x, y) = x^2 + y^2 - 4$. The functions $f(x, y) = -xy$ and g have continuous partial derivatives everywhere as they are polynomials. Then

$$\begin{cases} f'_x = \lambda g'_x \\ f'_y = \lambda g'_y \\ g = 0 \end{cases} \implies \begin{cases} -y = 2\lambda x \\ -x = 2\lambda y \\ x^2 + y^2 = 4 \end{cases} .$$

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 $f(\pm\sqrt{2}, \mp\sqrt{2}) = 2$ is the maximal value and $f(\pm\sqrt{2}, \pm\sqrt{2}) = -2$ is the minimal one. \square

EXAMPLE 13.51. *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

$$\begin{cases} f'_x = \lambda g'_x \\ f'_y = \lambda g'_y \\ f'_z = \lambda g'_z \\ g = 0 \end{cases} \implies \begin{cases} y + 2z = \lambda yz \\ x + 2z = \lambda xz \\ 2x + 2y = \lambda xy \\ xyz = V \end{cases} \implies \begin{cases} xy + 2xz = \lambda V \\ xy + 2zy = \lambda V \\ 2xz + 2yz = \lambda V \\ xyz = V \end{cases} ,$$

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 that $2z(y - x) = 0$. Since $z \neq 0$ ($V \neq 0$), one has $y = x$. Combining the first and third equations, one infers that $y(x - 2z) = 0$ and hence $x = 2z$. The substitution of $y = x = 2z$ 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. \square

The method of Lagrange multipliers can be used to determine extreme values of a function on a set D . Recall that the extreme values may occur on the boundary of D . In Example 13.47, explicit parametric equations of the boundary of D have been used (Step 2). Instead, an algebraic equation of the boundary, $g(x, y) = x^2 + y^2 - 4 = 0$, can be used in combination with the method of Lagrange multipliers. Indeed, if $f(x, y) = x^2 + y^2 + xy$, then its critical points along the boundary circle satisfy the system of equations:

$$\begin{cases} f'_x = \lambda g'_x \\ f'_y = \lambda g'_y \\ g = 0 \end{cases} \implies \begin{cases} 2x + y = 2\lambda x \\ 2y + x = 2\lambda y \\ x^2 + y^2 = 4 \end{cases} .$$

By subtracting the second equation from the first one, it follows that $x - y = 2\lambda(x - y)$. Hence, either $x = y$ or $\lambda = 1/2$. In the former case, the constraint yields $2x^2 = 4$ or $x = \pm\sqrt{2}$. The corresponding critical points are $(\pm\sqrt{2}, \pm\sqrt{2})$. If $\lambda = 1/2$, then from the first two equations in the system, one infers that $x = -y$. The constraint becomes $2x^2 = 4$ or $x = \pm\sqrt{2}$, and the critical points are $(\pm\sqrt{2}, \mp\sqrt{2})$.

Remark. The condition $\nabla g(\mathbf{r}_0) \neq \mathbf{0}$ is crucial for the method of Lagrange multipliers to work. If $\nabla f(\mathbf{r}_0) \neq \mathbf{0}$ (i.e. \mathbf{r}_0 is not a critical point of f without the constraint), then (13.17) have no solution when $\nabla g(\mathbf{r}_0) = \mathbf{0}$, and the method of Lagrange multipliers fails. Recall that the derivation of (13.17) requires that a curve defined by the equation $g(\mathbf{r}) = 0$ is smooth near \mathbf{r}_0 , which may no longer be the case if $\nabla g(\mathbf{r}_0) = \mathbf{0}$ (see the implicit function theorem). So, if f attains its local extreme value at a point which is a cusp of the curve $g(x, y) = 0$, then it cannot be determined by (13.17). For example, the equation $g(x, y) = x^3 - y^2 = 0$ defines a curve that has a cusp at $(0, 0)$ and $\nabla g(0, 0) = \mathbf{0}$, i.e., the curve has no normal vector at the origin. Since $x = y^{2/3} \geq 0$ on the curve, the function $f(x, y) = x$ attains its absolute minimum value 0 along this curve at the origin. However, $\nabla f(0, 0) = (1, 0)$ and the method of Lagrange multipliers fails to detect this point because there is no λ at which Eqs. (13.17) are satisfied. On the other hand, the function $f(x, y) = x^2$ also attains its absolute minimum value at the origin. Equations (13.17) have the

solution $(x, y) = (0, 0)$ and $\lambda = 0$. The difference between the two cases is that in the latter case $\nabla f(0, 0) = \mathbf{0}$, i.e., $(0, 0)$ is also a critical point of f without the constraint. The method of Lagrange multipliers also becomes inapplicable if g is not differentiable at \mathbf{r}_0 (see the exercises). The behavior of a function f at the points where ∇g does not exist or vanishes has to be investigated by different means.

96.2. The Case of Two or More Constraints. Let a function of three variables f have a local extreme value at point \mathbf{r}_0 on the set defined by two constraints $g_1(\mathbf{r}) = 0$ and $g_2(\mathbf{r}) = 0$. Provided g_1 and g_2 have continuous partial derivatives in a neighborhood of \mathbf{r}_0 and their gradients do not vanish at \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{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} . Suppose that *the vectors $\nabla g_1(\mathbf{r}_0)$ and $\nabla g_2(\mathbf{r}_0)$ are not parallel or, equivalently, $\nabla g_1(\mathbf{r}_0)$ is not proportional to $\nabla g_2(\mathbf{r}_0)$* . Then any vector in the plane normal to \mathbf{v} is a linear combination of them (see Study Problem 11.6). Therefore, there exist numbers λ_1 and λ_2 such that

$$\nabla f(\mathbf{r}) = \lambda_1 \nabla g_1(\mathbf{r}) + \lambda_2 \nabla g_2(\mathbf{r}), \quad g_1(\mathbf{r}) = g_2(\mathbf{r}) = 0$$

when $\mathbf{r} = \mathbf{r}_0$. 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 13.23. (Critical Points Subject to Constraints).

Suppose that functions g_a , $a = 1, 2, \dots, M$, of m variables, $m > M$, have continuous partial derivatives in a neighborhood of a point \mathbf{r}_0 and a function f has a local extreme value at \mathbf{r}_0 in the set defined by the constraints $g_a(\mathbf{r}) = 0$. Suppose that $\nabla g_a(\mathbf{r}_0)$ are nonzero vectors any of which cannot be expressed as a linear combination of the others and f is differentiable at \mathbf{r}_0 . Then there exist numbers λ_a such that

$$\nabla f(\mathbf{r}_0) = \lambda_1 \nabla g_1(\mathbf{r}_0) + \lambda_2 \nabla g_2(\mathbf{r}_0) + \cdots + \lambda_M \nabla g_M(\mathbf{r}_0).$$

EXAMPLE 13.52. *Find extreme values of the functions $f(x, y, z) = xyz$ on the curve that is an intersection of the sphere $x^2 + y^2 + z^2 = 1$ and the plane $x + y + z = 0$.*

SOLUTION: Put $g_1(x, y, z) = x^2 + y^2 + z^2 - 1$ and $g_2(x, y, z) = x + y + z$. One has $\nabla g_1 = (2x, 2y, 2z)$, which can only vanish at $(0, 0, 0)$, and hence $\nabla g_1 \neq \mathbf{0}$ on the sphere. Also, $\nabla g_2 = (1, 1, 1) \neq \mathbf{0}$. Therefore, critical points of f on the surface of constraints are determined by the equations:

$$\begin{array}{rcl} f'_x = \lambda_1 g'_x + \lambda_2 g'_x & & yz = 2\lambda_1 x + \lambda_2 \\ f'_y = \lambda_1 g'_y + \lambda_2 g'_y & & xz = 2\lambda_1 y + \lambda_2 \\ f'_z = \lambda_1 g'_z + \lambda_2 g'_z & \Rightarrow & xy = 2\lambda_1 z + \lambda_2 \\ g_1 = 0 & & 1 = x^2 + y^2 + z^2 \\ g_2 = 0 & & 0 = x + y + z \end{array}$$

Subtract the second equation from the first one to obtain $(y - x)z = 2\lambda_1(x - y)$. It follows, then, that either $x = y$ or $z = -2\lambda_1$. Suppose first that $y = x$. Then $z = -2x$ by the fifth equation. The substitution of $x = y$ and $z = -2x$ into the fourth equation yields $6x^2 = 1$ or $x = \pm 1/\sqrt{6}$. Therefore, the points $\mathbf{r}_1 = (1/\sqrt{6}, 1/\sqrt{6}, -2/\sqrt{6})$ and $\mathbf{r}_2 = (-1/\sqrt{6}, -1/\sqrt{6}, 2/\sqrt{6})$ are critical points provided there exist the corresponding values λ_1 and λ_2 such that all equations are satisfied. For example, take \mathbf{r}_1 . Then the second and third equations become

$$\begin{cases} -\frac{2}{6} = \frac{2}{\sqrt{6}}\lambda_1 + \lambda_2 \\ \frac{1}{6} = -\frac{4}{\sqrt{6}}\lambda_1 + \lambda_2 \end{cases} \Leftrightarrow \begin{cases} -\frac{3}{6} = \frac{6}{\sqrt{6}}\lambda_1 \\ \frac{1}{6} = -\frac{4}{\sqrt{6}}\lambda_1 + \lambda_2 \end{cases}$$

So $\lambda_1 = -1/(2\sqrt{6})$ and $\lambda_2 = -1/6$. The existence of λ_1 and λ_2 for the point \mathbf{r}_2 is verified similarly. Next, suppose that $z = -2\lambda_1$. Subtract the third equation from the second one to obtain $(z - y)x = 2\lambda_1(y - z)$. It follows that either $y = z$ or $x = -2\lambda_1$. Let $y = z$. The fifth equation yields $x = -2y$, and the fourth equation is reduced to $6y^2 = 1$. Therefore, there are two more critical points: $\mathbf{r}_3 = (-2/\sqrt{6}, 1/\sqrt{6}, 1/\sqrt{6})$ and $\mathbf{r}_4 = (2/\sqrt{6}, -1/\sqrt{6}, -1/\sqrt{6})$. The reader is to verify the existence of λ_1 and λ_2 in these cases (note that $\lambda_1 = -z/2$). Finally, let $x = -2\lambda_1$ and $z = -2\lambda_1$. These conditions imply that $x = z$ and, by the fifth equation, $y = -2x$. The fourth equation yields $6x^2 = 1$ so that there is another pair of critical points: $\mathbf{r}_5 = (1/\sqrt{6}, -2/\sqrt{6}, 1/\sqrt{6})$ and $\mathbf{r}_6 = (-1/\sqrt{6}, 2/\sqrt{6}, -1/\sqrt{6})$ (the reader is to verify the existence of λ_1 and λ_2). The intersection of the sphere and the plane is a circle. So it is sufficient to compare values of f at the critical points to determine the extreme values of f . It follows that f attains the maximum value $2/\sqrt{6}$ at $\mathbf{r}_2, \mathbf{r}_4$, and \mathbf{r}_6 and the minimum value $-2/\sqrt{6}$ at $\mathbf{r}_1, \mathbf{r}_3$, and \mathbf{r}_5 . \square

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 (13.17). 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 (13.17). Similarly, if there are several constraints, critical points of the function with additional variables λ_a , $a = 1, 2, \dots, M$,

$$(13.18) \quad F(\mathbf{r}, \lambda_1, \lambda_2, \dots, \lambda_n) = f(\mathbf{r}) - \sum_{a=1}^M \lambda_a g_a(\mathbf{r})$$

coincide with the critical points of f subject to the constraints $g_a = 0$ as stated in Theorem 13.23. 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*. The above observation provides a simple way to formulate the equations for critical points subject to constraints.

96.3. Finding Local Maxima and Minima. In the simplest case of a function f of two variables subject to a constraint, the nature of critical points (local maximum or minimum) can be determined 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 13.51, 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 g has continuous partial derivatives in a neighborhood of \mathbf{r}_0 and $\nabla g(\mathbf{r}_0) \neq \mathbf{0}$. 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, 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

$$(13.19) \quad F'' = (d/dx)(f'_x + f'_y h') = f''_{xx} + 2f''_{xy} h' + f''_{yy} (h')^2 + f'_y h''.$$

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

$$(13.20) \quad G'' = g''_{xx} + 2g''_{xy} h' + g''_{yy} (h')^2 + g'_y h'' = 0,$$

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 (13.19) 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 13.53. Show that the point $\mathbf{r}_0 = (0, 0)$ is a critical point of the function $f(x, y) = x^2 y + y + x$ 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

$$g''_{xx} = y^2 e^{xy}, \quad g''_{yy} = x^2 e^{xy}, \quad g''_{xy} = e^{xy} + xy e^{xy}.$$

The derivative $h''(0)$ is found from (13.20), 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 (13.19) 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. \square

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 13.21).

96.4. Study Problems.

Problem 13.20. *An axially symmetric solid consists of a circular cylinder and a right-angled circular cone attached to one of the cylinder's bases. What are the dimensions of the solid at which it has a maximal volume if the surface area of the solid has a fixed value S ?*

SOLUTION: Let r and h be the radius and height of the cylinder. Since the cone is right-angled, its height is r . The surface area is the sum of three terms: the area of the base (disk) πr^2 , the area of the side of the cylinder $2\pi r h$, and the surface area S_c of the cone. A cone with an angle α at the vertex is obtained by rotation of a straight line $y = mx$, where $m = \tan(\alpha/2)$, about the x axis. In the present case, $\alpha = \pi/2$ and $m = 1$. If a is the height of the cone, then the surface area of the cone is (see Calculus II)

$$S_c = \int_0^a 2\pi y \, dx = \int_0^r 2\pi x \, dx = \pi r^2$$

because $a = r$ in the present case. Similarly, the volume of the cone is

$$V_c = \int_0^a \pi y^2 dx = \int_0^r \pi x^2 dx = \frac{\pi r^3}{3}.$$

Therefore, the problem is reduced to finding the maximal value of the function (volume) $V(r, h) = \pi r^2 h + \pi r^3/3$ subject to the constraint $2\pi r h + 2\pi r^2 = S$. Put $g(r, h) = 2\pi r h + 2\pi r^2 - S$. Then critical points of V satisfy the equations:

$$\begin{cases} V'_r = \lambda g'_r \\ V'_h = \lambda g'_h \\ 0 = g(r, h) \end{cases} \Rightarrow \begin{cases} 2\pi r h + \pi r^2 h = \lambda(2\pi h + 4\pi r) \\ \pi r^2 = 2\pi \lambda r \\ \frac{S}{2\pi} = r h + r^2 \end{cases}.$$

Since $r \neq 0$ (the third equation is not satisfied if $r = 0$), the second equation implies that $\lambda = r/2$. The substitution of the latter into the first equation yields $r h = r^2$ or $h = r$. Then it follows from the third equation that the sought-after dimensions are $h = r = \sqrt{S/(4\pi)}$. \square

Problem 13.21. Let functions f and g of three variables $\mathbf{r} = (x, y, z)$ have continuous partial derivatives up to order 2. 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 h''_{xx} , h''_{xy} , and h''_{yy} are found from the equations

$$\begin{aligned} G''_{xx} = 0 &\implies g''_{xx} + 2g''_{xz} h'_x + g''_{zz} (h'_x)^2 + g'_z h''_{xx} = 0, \\ G''_{yy} = 0 &\implies g''_{yy} + 2g''_{yz} h'_y + g''_{zz} (h'_y)^2 + g'_z h''_{yy} = 0, \\ G''_{xy} = 0 &\implies g''_{xy} + g''_{xz} h'_x + g''_{yz} h'_y + g''_{zz} h'_x h'_y + g'_z h''_{xy} = 0. \end{aligned}$$

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 first and second partial derivatives of h have been found earlier. 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 . \square

96.5. Exercises.

(1) Use Lagrange multipliers to find the maximum and minimum values of each of the following functions subject to the specified constraints:

- (i) $f(x, y) = xy$, $x + y = 1$
- (ii) $f(x, y) = x^2 + y^2$, $x/a + y/b = 1$
- (iii) $f(x, y) = xy^2$, $2x^2 + y^2 = 6$
- (iv) $f(x, y) = y^2$, $x^2 + y^2 = 4$
- (v) $f(x, y) = x + y$, $x^2/16 + y^2/9 = 1$
- (vi) $f(x, y) = 2x^2 - 2y^2$, $x^4 + y^4 = 32$
- (vii) $f(x, y) = Ax^2 + 2Bxy + Cy^2$, $x^2 + y^2 = 1$
- (viii) $f(x, y, z) = xyz$, $3x^2 + 2y^2 + z^2 = 6$
- (ix) $f(x, y, z) = x - 2y + 2z$, $x^2 + y^2 + z^2 = 1$
- (x) $f(x, y, z) = x^2 + y^2 + z^2$, $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$
- (xi) $f(x, y, z) = -x + 3y - 3z$, $x + y - z = 0$, $y^2 + 2z^2 = 1$
- (xii) $f(x, y, z) = xy + yz$, $xy = 1$, $y^2 + 2z^2 = 1$
- (xiii) $f(x, y, z) = xy + yz$, $x^2 + y^2 = 2$, $y + z = 2$ ($x > 0$, $y > 0$, $z > 0$)
- (xiv) $f(x, y, z) = \sin(x) \sin(y) \sin(z)$, $x + y + z = \pi/2$ ($x > 0$, $y > 0$, $z > 0$)
- (xv) $f(x, y, z) = x^2/a^2 + y^2/b^2 + z^2/c^2$, $x^2 + y^2 + z^2 = 1$, $n_1x + n_2y + n_3z = 0$, where $\hat{\mathbf{n}} = (n_1, n_2, n_3)$ is a unit vector
- (xvi) $f(\mathbf{r}) = \hat{\mathbf{u}} \cdot \mathbf{r}$, $\|\mathbf{r}\| = R$, where $\mathbf{r} = (x_1, \dots, x_m)$, $\hat{\mathbf{u}}$ is a constant unit vector, and R is a constant
- (xvii) $f(\mathbf{r}) = \mathbf{r} \cdot \mathbf{r}$, $\mathbf{n} \cdot \mathbf{r} = 1$, where \mathbf{n} has strictly positive components and $\mathbf{r} = (x_1, x_2, \dots, x_m)$
- (xviii) $f(\mathbf{r}) = x_1^n + x_2^n + \dots + x_m^n$, $x_1 + x_2 + \dots + x_m = a$, where $n > 0$ and $a > 0$

(2) Prove the inequality

$$\frac{x^n + y^n}{2} \geq \left(\frac{x + y}{2} \right)^n$$

if $n \geq 1$, $x \geq 0$, and $y \geq 0$. *Hint:* Minimize the function $f = (x^n + y^n)/2$ under the condition $x + y = s$.

(3) Find the minimal value of the function $f(x, y) = y$ on the curve $x^2 + y^4 - y^3 = 0$. Explain why the method of Lagrange multipliers fails.

Hint: Sketch the curve near the origin.

(4) Use the method of Lagrange multipliers to maximize the function $f(x, y) = 3x + 2y$ on the curve $\sqrt{x} + \sqrt{y} = 5$. Compare the obtained value with $f(0, 25)$. Explain why the method of Lagrange multipliers fails.

(5) Find three positive numbers whose sum is a fixed number $c > 0$ and whose product is maximal.

(6) Use the method of Lagrange multipliers to solve the following exercises from Section 95.5:

- (i) Exercise 10
- (ii) Exercise 11
- (iii) Exercise 12
- (iv) Exercise 13
- (v) Exercise 14

(7) The cross section of a cylindrical tab is a half-disk. If the tab has total area S , what are the dimensions at which the tab has maximal volume?

(8) Find a rectangle with a fixed perimeter $2p$ that forms a solid of the maximal volume under rotation about one of its sides.

(9) Find a triangle with a fixed perimeter $2p$ that forms a solid of the maximal volume under rotation about one of its sides.

(10) Find a rectangular box with the maximal volume that is contained in a half-ball of radius R .

(11) Find a rectangular box with the maximal volume that is contained in an ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$.

(12) Consider a circular cone obtained by rotation of a straight line segment of length l about the axis through an endpoint of the segment. If the angle between the segment and the axis is θ , find a rectangular box within the cone that has a maximal volume.

(13) The solid consists of a rectangular box and two identical pyramids whose bases are opposite faces of the box. The edges of the pyramid adjacent at the vertex opposite to its base have equal lengths. If the solid has a fixed volume V , at what angle between the edges of the pyramid and its base is the surface area of the solid minimal?

(14) Use Lagrange multipliers to find the distance between the parabola $y = x^2$ and the line $x - y = 2$.

(15) Find the maximum value of the function $f(\mathbf{r}) = \sqrt[m]{x_1 x_2 \cdots x_m}$ given that $x_1 + x_2 + \cdots + x_m = c$, where c is a positive constant. Deduce from the result that if $x_i > 0$, $i = 1, 2, \dots, m$, then

$$\sqrt[m]{x_1 x_2 \cdots x_m} \leq \frac{1}{m}(x_1 + x_2 + \cdots + x_m);$$

that is, the *geometrical mean* of m numbers is no larger than the *arithmetic mean*. When is the equality reached?

(16) Give an alternative proof of the Cauchy-Schwarz inequality in a Euclidean space (Theorem 13.1) using the method of Lagrange multipliers to maximize the function of $2m$ variables $f(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ subject to the constraints $\mathbf{x} \cdot \mathbf{x} = 1$ and $\mathbf{y} \cdot \mathbf{y} = 1$, where $\mathbf{x} = (x_1, \dots, x_m)$ and $\mathbf{y} = (y_1, \dots, y_m)$. *Hint:* After maximizing the function, put $\mathbf{x} = \mathbf{a}/\|\mathbf{a}\|$ and $\mathbf{y} = \mathbf{b}/\|\mathbf{b}\|$ for any two nonzero vectors \mathbf{a} and \mathbf{b} .

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 $p = 1, 2, \dots, N$ 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 (see the left panel of Figure 14.1). 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

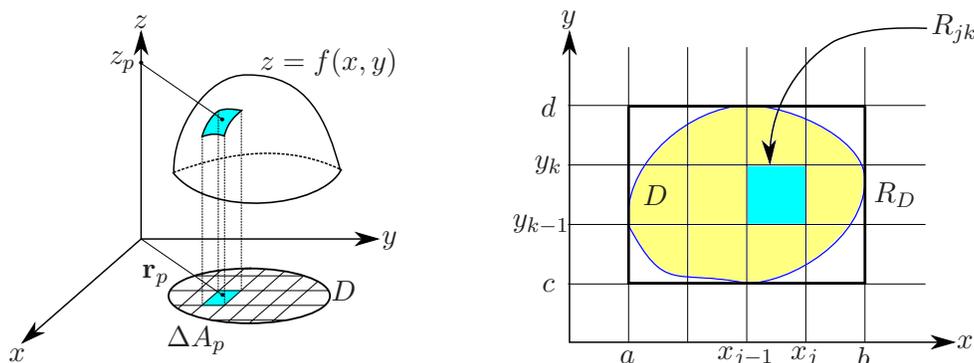


FIGURE 14.1. **Left:** The volume of a solid region bounded from above by the graph $z = f(x, y)$ and from below by a portion D of the xy plane is approximated by the sum of volumes $\Delta V_p = z_p \Delta A_p$ of columns with the base area ΔA_p and the height $z_p = f(\mathbf{r}_p)$, where \mathbf{r}_p is a sample point within the base and p enumerates the columns. **Right:** A rectangular partition of a region D is obtained by embedding D into a rectangle R_D . Then the rectangle R_D is partitioned into smaller rectangles R_{kj} .

therefore expect that the exact value of the volume is obtained in the limit

$$(14.1) \quad V = \lim_{\substack{N \rightarrow \infty \\ (R_N \rightarrow 0)}} \sum_{p=1}^N f(\mathbf{r}_p) \Delta A_p.$$

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 , $k = 0, 1, \dots, N$, $x_0 = a < x_1 < \dots < x_{N-1} < x_N = b$ is the partition of $[a, b]$, then ΔA_p is the analog of $\Delta x_j = x_j - x_{j-1}$, $j = 1, 2, \dots, N$, the number R_N is the analog of $\Delta_N = \max_j \Delta x_j$, and the values $f(\mathbf{r}_p)$ are analogous to $f(x_j^*)$, where $x_j^* \in [x_{j-1}, x_j]$. The area under the graph is then

$$A = \lim_{\substack{N \rightarrow \infty \\ (\Delta_N \rightarrow 0)}} \sum_{j=1}^N f(x_j^*) \Delta x_j = \int_a^b f(x) dx.$$

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, $m \leq f(\mathbf{r}) \leq M$ 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 $R_D = \{(x, y) \mid x \in [a, b], y \in [c, d]\}$ (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 $x_j, j = 0, 1, \dots, N_1$, of the interval $[a, b]$, where $x_j = a + j \Delta x$, $\Delta x = (b - a)/N_1$, and a partition $y_k, k = 0, 1, \dots, N_2$, of the interval $[c, d]$, where $y_k = c + k \Delta y$ and $\Delta y = (d - c)/N_2$. These partitions induce a partition of the rectangle R_D by rectangles $R_{jk} = \{(x, y) \mid x \in [x_{j-1}, x_j], y \in [y_{k-1}, y_k]\}$, where $j = 1, 2, \dots, N_1$ and $k = 1, 2, \dots, N_2$. The area of each partition rectangle R_{jk} is $\Delta A = \Delta x \Delta y$. This partition is called a *rectangular partition* of R_D . It is depicted in the right panel of Figure 14.1. For every partition rectangle R_{jk} , there are numbers $M_{jk} = \sup f(\mathbf{r})$ and $m_{jk} = \inf f(\mathbf{r})$, the supremum and infimum of f on R_{jk} .

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_{jk} 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

$$U(f, N_1, N_2) = \sum_{j=1}^{N_1} \sum_{k=1}^{N_2} M_{jk} \Delta A, \quad L(f, N_1, N_2) = \sum_{j=1}^{N_1} \sum_{k=1}^{N_2} m_{jk} \Delta A$$

are called the upper and lower sums.

The upper and lower sums are examples of double sequences.

DEFINITION 14.3. (Double Sequence).

A double sequence is a rule that assigns a number a_{nm} to an ordered pair of integers (n, m) , $n, m = 1, 2, \dots$

In other words, a double sequence is a function f of two variables (x, y) whose domain consists of points with integer-valued coordinates, $a_{nm} = f(n, m)$. Similarly to ordinary numerical sequences, one can define a limit of a double sequence.

DEFINITION 14.4. (Limit of a Double Sequence).

If, for any positive number ε , there exists an integer N such that $|a_{nm} - a| < \varepsilon$ for all $n, m > N$, then the sequence is said to converge to a and the number a is called the limit of the sequence and denoted $\lim_{n, m \rightarrow \infty} a_{nm} = a$.

The limit of a double sequence is analogous to the limit of a function of two variables. A limit of a double sequence can be found by studying the corresponding limit of a function of two variables whose range contains the double sequence. Suppose $a_{nm} = f(1/n, 1/m)$ and $f(x, y) \rightarrow 0$ as $(x, y) \rightarrow (0, 0)$. The latter means that, for any $\varepsilon > 0$, there is a number $\delta > 0$ such that $|f(x, y)| < \varepsilon$ for all $\|\mathbf{r}\| < \delta$, where $\mathbf{r} = (x, y)$. In particular, for $\mathbf{r} = (1/n, 1/m)$, the condition $\|\mathbf{r}\|^2 = 1/n^2 + 1/m^2 < \delta^2$ is satisfied for all $n, m > N > 2/\delta$. Hence, for all such n, m , $|a_{nm}| < \varepsilon$, which means that $a_{nm} \rightarrow 0$ as $n, m \rightarrow \infty$.

Continuing the analogy with the volume problem, the upper and lower sums represent the smallest upper estimate and the greatest lower estimate of the volume. They should become closer and closer to the volume as the partition becomes finer and finer. This leads to the following definition of the double integral.

DEFINITION 14.5. (Double Integral).

If the limits of the upper and lower sums exist as $N_{1,2} \rightarrow \infty$ (or

$(\Delta x, \Delta y) \rightarrow (0, 0)$ and coincide, then f is said to be Riemann integrable on D , and the limit of the upper and lower sums

$$\iint_D f(x, y) dA = \lim_{N_{1,2} \rightarrow \infty} U(f, N_1, N_2) = \lim_{N_{1,2} \rightarrow \infty} L(f, N_1, N_2)$$

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 $(\Delta x, \Delta y) \rightarrow (0, 0)$ or as the limit of double sequences. The existence of the limit and its value must be established according to Definition 14.4.

Let us discuss Definition 14.5 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 14.6. (Riemann Sum).

Let f be a bounded 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 rectangular 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:

$$\lim_{N_{1,2} \rightarrow \infty} R(f, N_1, N_2) = \iint_D f dA.$$

PROOF. For any partition rectangle R_{jk} and any sample point \mathbf{r}_{jk}^* in it, $m_{jk} \leq f(\mathbf{r}_{jk}^*) \leq M_{jk}$. It follows from this inequality that $L(f, N_1, N_2) \leq R(f, N_1, N_2) \leq U(f, N_1, N_2)$. 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. \square

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. Not every bounded function is integrable. There are functions whose behavior is so irregular that one cannot give any meaning to the volume under their graph by converging upper and lower sums.

An Example of a Nonintegrable Function. Let f be 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, or 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, $2 \neq 0$, 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.

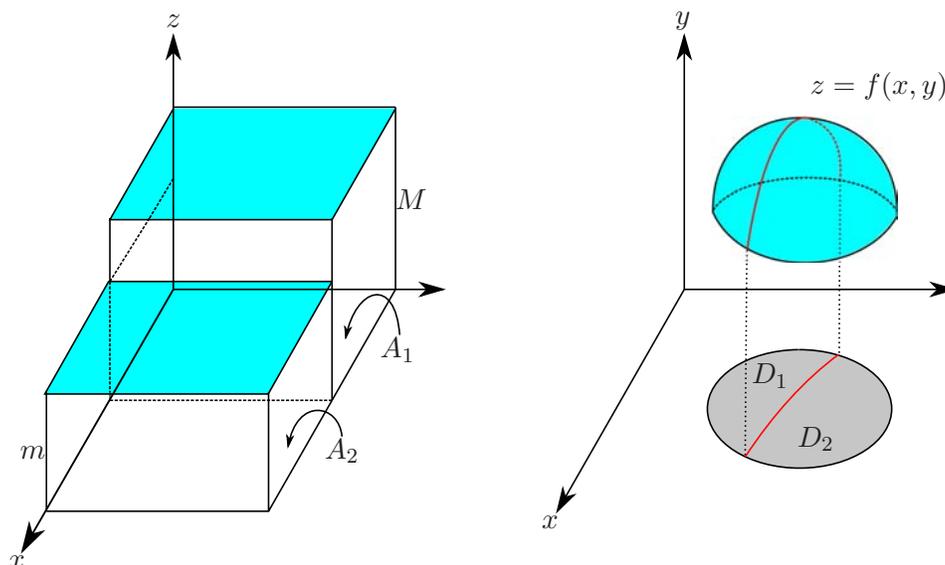


FIGURE 14.2. **Left:** The graph of a piecewise-constant function. The function has a jump discontinuity along a straight line. The volume under the graph is $V = MA_1 + mA_2$. Despite the jump discontinuity, the function is integrable and the value of the double integral coincides with the volume V . **Right:** Additivity of the double integral. If a region D is split by a curve into two regions D_1 and D_2 , then the double integral of f over D is the sum of integrals over D_1 and D_2 . The additivity of the double integral is analogous to the additivity of the volume: The volume under the graph $z = f(x, y)$ and above D is the sum of volumes above D_1 and D_2 .

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 . Its graph is shown in the left panel of Figure 14.2. The volume below the graph $z = f(x, y)$ and above D is easy to find; it is the sum of the 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} differ 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 $\Delta x \rightarrow 0$, 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 14.1. *Let D be a closed, bounded region whose boundaries are piecewise-smooth curves. If a function f is bounded on D and discontinuous only on a finite number of smooth curves, then it is integrable on D .*

97.4. Exercises.

(1) For each of the following functions and the specified rectangular domain D , find the double integral using its definition:

- (i) $f(x, y) = k = \text{const}$, $D = \{(x, y) | a \leq x \leq b, c \leq y \leq d\}$
- (ii) $f(x, y) = k_1 = \text{const}$ if $y > 0$ and $f(x, y) = k_2 = \text{const}$ if $y \leq 0$, $D = \{(x, y) | 0 \leq x \leq 1, -1 \leq y \leq 1\}$
- (iii) $f(x, y) = xy$, $D = \{(x, y) | 0 \leq x \leq 1, 0 \leq y \leq 1\}$
Hint: $1 + 2 + \cdots + N = N(N + 1)/2$.

(2) Let D be the rectangle $1 \leq x \leq 2$, $1 \leq y \leq 3$. Consider a rectangular partition of D by lines $x = 1 + j/N$ and $y = 1 + 2k/N$, $j, k = 1, 2, \dots, N$. For the function $f(x, y) = x^2 + y^2$, find

- (i) The lower and upper sums, U and L
- (ii) The limit of the difference $U - L$ as $N \rightarrow \infty$
- (iii) The limit of the sums as $N \rightarrow \infty$

(3) For each of the following functions, use a Riemann sum with specified N_1 and N_2 and sample points at lower right corners to estimate the double integral over a given region D :

- (i) $f(x, y) = x + y^2$, $(N_1, N_2) = (2, 2)$, $D = \{(x, y) | 0 \leq x \leq 2, 0 \leq y \leq 4\}$
- (ii) $f(x, y) = \sin(x + y)$, $(N_1, N_2) = (3, 3)$, $D = \{(x, y) | 0 \leq x \leq \pi, 0 \leq y \leq \pi\}$

(4) Approximate the integral of $f(x, y) = (24 + x^2 + y^2)^{-1/2}$ over the disk $x^2 + y^2 \leq 25$ by a Riemann sum. Use a partition by squares whose vertices have integer-valued coordinates and sample points at vertices of the squares that are farthest from the origin.

(5) Evaluate each of the following double integrals by first identifying it as the volume of a solid:

- (i) $\iint_D k \, dA$ if D is the disk $x^2 + y^2 \leq 1$ and k is a constant
- (ii) $\iint_D \sqrt{1 - x^2 - y^2} \, dA$ if D is the disk $x^2 + y^2 \leq 1$
- (iii) $\iint_D (1 - x - y) \, dA$ if D is the triangle with vertices $(0, 0)$, $(0, 1)$, and $(1, 0)$
- (iv) $\iint_D (k - x) \, dA$ if D is the rectangle $0 \leq x \leq k$ and $0 \leq y \leq a$
- (v) $\iint_D (2 - \sqrt{x^2 + y^2}) \, dA$ if D is the part of the disk $x^2 + y^2 \leq 1$ in the first quadrant (*Hint*: The volume of a circular solid cone with the base being the disk of radius R and the height h is $\pi R^2 h/3$.)

(6) Let I be the integral of $\sin(x+y)$ over the disk $x^2 + y^2 \leq 1$. Suppose that the integration region is partitioned by rectangles of area ΔA . If R is a Riemann sum, find ΔA such that $|I - R| < 0.001$ for any choice of sample points.

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

$$\begin{aligned} \iint_D (f + g) \, dA &= \iint_D f \, dA + \iint_D g \, dA, \\ \iint_D cf \, dA &= c \iint_D f \, dA. \end{aligned}$$

Area. The double integral

$$(14.2) \quad A(D) = \iint_D dA$$

is called the *area* of D (if it exists). If D is bounded by piecewise-smooth curves, then it exists because the unit function $f = 1$ is continuous on D . By the geometrical interpretation of the double integral, the number $A(D)$ is the volume of the solid cylinder with the cross section D and the *unit* height ($f = 1$). Intuitively, the region D can always be covered by the union of adjacent rectangles of area $\Delta A = \Delta x \Delta y$. In the limit $(\Delta x, \Delta y) \rightarrow (0, 0)$, 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 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 (see the right panel of Figure 14.2).

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 the double integral is the difference of the volumes above and below the xy plane. Therefore, it may vanish or take negative values, depending on which volume is larger. This property is analogous to the familiar relation between the ordinary integral and the area under the graph. It is illustrated in Figure 14.3 (left panel).

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 property for the double integrals of $f(x, y) - m \geq 0$ and $M - f(x, y) \geq 0$ over D and (14.2), 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

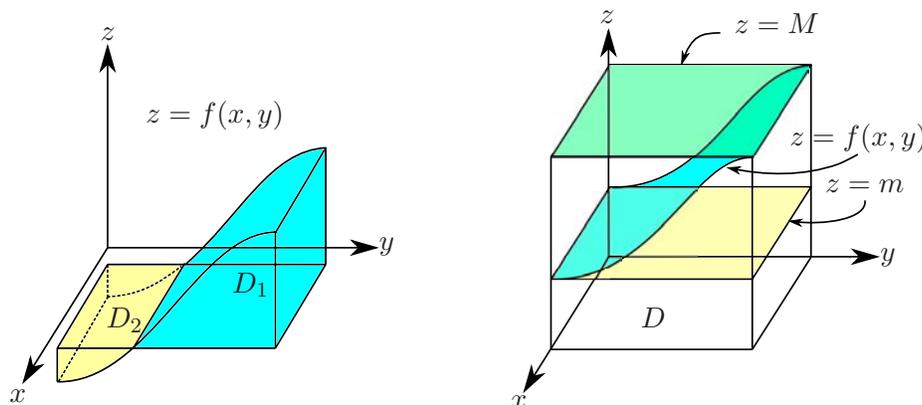


FIGURE 14.3. **Left:** A function f is nonnegative on the region D_1 and nonpositive on D_2 . The double integral of $f(x, y)$ over the union of regions D_1 and D_2 is the difference of the indicated volumes. The volume below the xy plane and above the graph of f contributes to the double integral with the negative sign. **Right:** An illustration to the upper and lower bounds of the double integral of a function f over a region D . If $A(D)$ is the area of D and $m \leq f(x, y) \leq M$ in D , then the volume under the graph of f is no less than the volume $mA(D)$ and no larger than $MA(D)$.

lies in the cylinder with 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 as shown in the right panel of Figure 14.3.

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 (Extreme Value Theorem 12.21). Therefore, for any $m \leq h_0 \leq M$, there is $\mathbf{r}_0 \in D$ such that $f(\mathbf{r}_0) = h_0$. \square

A geometrical interpretation of the integral mean value theorem is rather simple. Imagine that the solid below the graph of f is made of

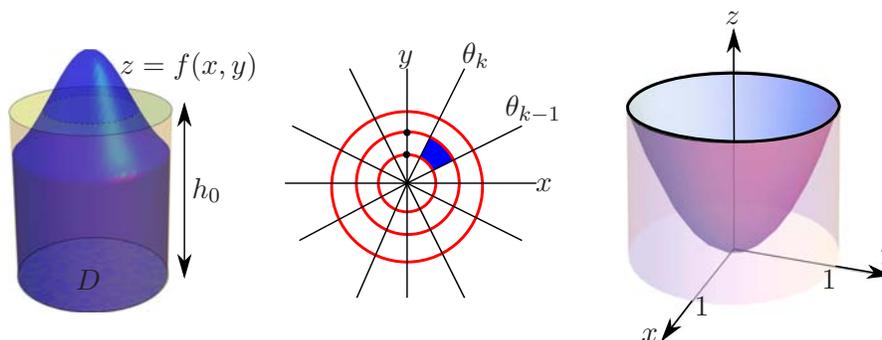


FIGURE 14.4. **Left:** A clay solid with a nonflat top (the graph of a continuous function f) may be deformed to the solid of the same volume and with the same horizontal cross section D , but with a flat top h_0 . The function f takes the value h_0 at some point of D . This illustrates the integral mean value theorem. **Middle:** A partition of a disk by concentric circles of radii $r = r_p$ and rays $\theta = \theta_k$ as described in Example 14.1. A partition element is the region $r_{p-1} \leq r \leq r_p$ and $\theta_{k-1} \leq \theta_k$. **Right:** The volume below the graph $z = x^2 + y^2$ and above the disk D , $x^2 + y^2 \leq 1$. The corresponding double integral is evaluated in Example 14.1 by taking the limit of Riemann sums for the partition of D shown in the middle panel.

clay (see the left panel of Figure 14.4). 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.

DEFINITION 14.7. (Average Value of a Function).

Let f be integrable on D and let $A(D)$ be the area of D . The average value of f on D is the integral:

$$\frac{1}{A(D)} \iint_D f dA.$$

If f is continuous on D , then the integral mean value theorem asserts that f attains its average value at some point in D . The continuity hypothesis is crucial here. For example, the function depicted in the left panel of Figure 14.2 is discontinuous. Its average value is $(MA_1 + mA_2)/(A_1 + A_2)$, which generally does not coincide with either M or m .

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

$$\left| \iint_D f \, dA \right| \leq \iint_D |f| \, dA.$$

A proof of the integrability of $|f|$ is rather technical. Once the integrability of $|f|$ is established, the inequality is a simple consequence of $|a + b| \leq |a| + |b|$ applied to a Riemann sum of f . Making the analogy between the double integral and the volume, suppose that $f \geq 0$ on D_1 and $f \leq 0$ 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, $|V_1 - V_2| \leq V_1 + V_2$ for positive $V_{1,2}$. The converse is not true. The integrability of the absolute value $|f|$ does not generally imply the integrability of f . The reader is advised to consider the function $f(x, y) = 1$ if x and y are rational and $f(x, y) = -1$ otherwise where (x, y) span the rectangle $[0, 1] \times [0, 1]$. Note that $|f(x, y)| = 1$ is integrable.

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 analysis is limited to the case when f is continuous.

DEFINITION 14.8. (Uniform Continuity).

Let f be a function on a region D in a Euclidean space. If, for any number $\varepsilon > 0$, there exists a number $\delta > 0$ such that $|f(\mathbf{r}) - f(\mathbf{r}')| < \varepsilon$ whenever $\|\mathbf{r} - \mathbf{r}'\| < \delta$ for any \mathbf{r} and \mathbf{r}' in D , then f is called uniformly continuous on D .

Definition 14.8 implies continuity of f at \mathbf{r}' : $f(\mathbf{r}) \rightarrow f(\mathbf{r}')$ as $\mathbf{r} \rightarrow \mathbf{r}'$ according to Definition 13.9. So the uniform continuity implies continuity. The converse is not true. The uniform continuity imposes a stronger condition on the behavior of the function. The difference is that the number δ can be chosen *independently* of a point \mathbf{r}_0 , whereas

for a continuous function, δ generally depends on both ε and \mathbf{r}_0 . In other words, the uniform continuity means that, for any preassigned positive number ε , there is a positive number δ such that the values of the function do not differ more than ε within a ball of radius δ *no matter where in D the ball is centered*. The relation between continuous and uniformly continuous functions is established in the following theorem.

THEOREM 14.4. *If f is continuous on a bounded, closed region D in a Euclidean space, then f is uniformly continuous on D .*

The proof is omitted. The hypothesis of the closedness of D is essential. Take $f(x, y) = 1/x$, which is continuous in the rectangle $D = (0, 1] \times [0, 1]$. Note D is not closed. Then in a disk whose center is sufficiently close to the line $x = 0$, the values of f can have variations as large as desired within this disk because $1/x$ diverges as x approaches 0. For example, take an interval $[x_1, x_2]$ of length $\delta = x_2 - x_1$. Then $1/x_1 - 1/x_2$ can be made arbitrarily large by moving x_1 closer to 0 for any choice of $\delta > 0$. So the variations of f cannot be bounded by a fixed number ε uniformly in any disk of some nonzero radius in D , and f is not uniformly continuous in D . Similarly, take $f(x, y) = x^2$, which is continuous in the unbounded rectangle $D = [0, \infty) \times [0, 1]$. Then in a disk whose center is sufficiently far from the line $x = 0$, the values of f can have variations as large as desired within this disk. For an interval $[x_1, x_2]$ of length $\delta > 0$, the variation $x_2^2 - x_1^2 = \delta(x_2 + x_1)$ can be made as large as desired by taking x_2 large enough no matter how small δ is. Hence, f is not uniformly continuous in D .

Let f be continuous in a closed, bounded region D . Let D be partitioned by piecewise-smooth curves into partition elements D_p , $p = 1, 2, \dots, N$, 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. Under the aforementioned conditions, the following theorem holds.

THEOREM 14.5. (Independence of the Partition).

For any choice of sample points \mathbf{r}_p^ and any choice of partition elements D_p ,*

$$(14.3) \quad \iint_D f \, dA = \lim_{\substack{N \rightarrow \infty \\ (R_N \rightarrow 0)}} \sum_{p=1}^N f(\mathbf{r}_p^*) \Delta A_p.$$

PROOF. As 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 $\mathbf{r}_p^* \neq \mathbf{r}_p$, then the Riemann sum does not coincide with the double integral. However, its limit as $N \rightarrow \infty$ equals the double integral. Indeed, put $c_p = |f(\mathbf{r}_p^*) - f(\mathbf{r}_p)|$ and $c_N = \max c_p$, $p = 1, 2, \dots, N$. By Theorem 14.4, f is uniformly continuous on D . For any $\varepsilon > 0$, there is $\delta > 0$ such that variations of f in any disk of radius δ in D do not exceed ε . Since $R_N \rightarrow 0$ as $N \rightarrow \infty$, $R_N < \delta$ for all N larger than some N_0 . Hence, $c_N < \varepsilon$ because any partition element D_p is contained in a disk of radius $R_p \leq R_N < \delta$, which implies that $c_N \rightarrow 0$ as $N \rightarrow \infty$. Therefore, the deviation of the Riemann sum from the double integral converges to 0:

$$\begin{aligned} \left| \iint_D f \, dA - R(f, N) \right| &= \left| \sum_{p=1}^N (f(\mathbf{r}_p) - f(\mathbf{r}_p^*)) \Delta A_p \right| \\ &\leq \sum_{p=1}^N |f(\mathbf{r}_p) - f(\mathbf{r}_p^*)| \Delta A_p \\ &= \sum_{p=1}^N c_p \Delta A_p \leq c_N \sum_{p=1}^N \Delta A_p = c_N A(D) \rightarrow 0 \end{aligned}$$

as $N \rightarrow \infty$. □

A practical significance of this theorem is that the double integral can be approximated by Riemann sums for *any convenient partition* of the integration region. 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 $p = 0, 1, 2, \dots, N$. 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 $\theta = \theta_k = k \Delta\theta$, $\Delta\theta = 2\pi/n$, $k = 1, 2, \dots, n$. Each partition element lies in the sector of angle $\Delta\theta$ and is bounded by two circles whose radii differ by Δr (see the middle panel of Figure 14.4). 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 $\Delta A_p = r_{p+1}^2 \Delta\theta/2 - r_p^2 \Delta\theta/2 = (r_{p+1}^2 - r_p^2) \Delta\theta/2 = (r_{p+1} + r_p) \Delta r \Delta\theta/2$. 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 $N \rightarrow \infty$, this sum converges to the integral of g over the interval $[0, 1]$, that is,

$$\iint_D (x^2 + y^2) dA = 2\pi \lim_{N \rightarrow \infty} \sum_{p=1}^N \bar{r}_p^3 \Delta r = 2\pi \int_0^1 r^3 dr = \pi/2.$$

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. \square

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 .

98.1. Exercises.

(1) Evaluate each of the following double integrals by using the properties of the double integral and its interpretation as the volume of a solid:

- (i) $\iint_D k \, dA$, where k is a constant and D is the square $-2 \leq x \leq 2$, $-2 \leq y \leq 2$ with a circular hole of radius 1 (i.e., $x^2 + y^2 \geq 1$ in D)
- (ii) $\iint_D f \, dA$, where D is a disk $x^2 + y^2 \leq 4$ and f is a piecewise-constant function: $f(x, y) = 2$ if $1 \leq x^2 + y^2 \leq 4$ and $f(x, y) = -3$ if $0 \leq x^2 + y^2 < 1$
- (iii) $\iint_D (4 + 3\sqrt{x^2 + y^2}) \, dA$, where D is the disk $x^2 + y^2 \leq 1$
- (iv) $\iint_D (\sqrt{4 - x^2 - y^2} - 2) \, dA$ if D is the part of the disk $x^2 + y^2 \leq 4$ in the first quadrant
- (v) $\iint_D (4 - x - y) \, dA$, where D is the rectangle with vertices $(0, 0)$, $(1, 0)$, and $(0, 1)$ (*Hint*: Use the identity $4 - x - y = 3 + (1 - x - y)$ and the linearity of the double integral.)
- (2) Use the positivity of the double integral to show that
- (i) $\iint_D \sin(xy)/(xy) \, dA \leq A(D)$, where D is a bounded region in which $x > 0$ and $y > 0$
- (ii) $\iint_D (ax^2 + by^2) \, dA \leq (a+b)\pi/2$, where D is the disk $x^2 + y^2 \leq 1$ (*Hint*: Put $r^2 = x^2 + y^2$. Then use $x^2 \leq r^2$ and $y^2 \leq r^2$ and apply the result of Example 14.1.)
- (3) Find the lower and upper bounds for each of the following integrals:
- (i) $\iint_D xy^3 \, dA$, where D is the square $1 \leq x \leq 2$, $1 \leq y \leq 2$
- (ii) $\iint_D \sqrt{1 + xe^{-y}} \, dA$, where D is the square $0 \leq x \leq 1$, $0 \leq y \leq 1$
- (iii) $\iint_D \sin(x + y) \, dA$, where D is the triangle with vertices $(0, 0)$, $(0, \pi)$, and $(\pi/4, 0)$ (*Hint*: Graph D . Then graph the set of point at which $\sin(x + y)$ attains its maximal value.)
- (iv) $\iint_D (100 + \cos^2 x + \cos^2 y)^{-1} \, dA$, where D is defined by $|x| + |y| \leq 10$
- (4) Let f be continuous on a bounded region D with a nonzero area. If the double integral of f over D vanishes, prove that there is a point in D at which f vanishes.
- (5) Use the method of Example 14.1 to find $\iint_D e^{x^2 + y^2} \, dA$, where D is the part of the disk $x^2 + y^2 \leq 1$.
- (6) Use a Riemann sum to approximate the double integral of $f(x, y) = \sqrt{x + y}$ over the triangle bounded by the lines $x = 0$, $y = 0$, and $x + y = 1$. Partition the integration region into four equal triangles by the lines $x = \text{const}$, $y = \text{const}$, and $x + y = \text{const}$. Choose sample points to be centroids of the triangle.
- (7) Determine the sign of each of the following integrals:
- (i) $\iint_D \ln(x^2 + y^2) \, dA$, where D is defined by $|x| + |y| \leq 1$
- (ii) $\iint_D \sqrt[3]{1 - x^2 - y^2} \, dA$, where D is defined by $x^2 + y^2 \leq 4$

- (iii) $\iint_D \sin^{-1}(x+y) dA$, where D is defined by $0 \leq x \leq 1$ and $0 \leq y \leq 1-x$

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. In combination with the properties of the double integral, it is sufficient for many applications.

Recall Section 87.3 where it has been shown that if a multivariable limit exists, then the *repeated limits* exist and coincide. A similar statement is true for double sequences.

THEOREM 14.6. *Suppose that a double sequence a_{nm} converges to a as $n, m \rightarrow \infty$. Then*

$$\lim_{n \rightarrow \infty} \left(\lim_{m \rightarrow \infty} a_{nm} \right) = \lim_{m \rightarrow \infty} \left(\lim_{n \rightarrow \infty} a_{nm} \right) = a.$$

A proof of this simple theorem is left to the reader as an exercise. The limit $\lim_{m \rightarrow \infty} a_{nm} = b_n$ is taken for a fixed value of n . Similarly, the limit $\lim_{n \rightarrow \infty} a_{nm} = c_m$ is taken for a fixed m . The theorem states that the limits of two generally *different* sequences b_n and c_m coincide and are equal to the limit of the double sequence. This property of double sequences will be applied to Riemann sums to reduce a double integral to ordinary *iterated* integrals.

99.1. Rectangular Domains. Let a function f be continuous on D . Suppose first that D is a rectangle $x \in [a, b]$ and $y \in [c, d]$. In what follows, a brief notation for rectangles will be used: $D = [a, b] \times [c, d]$. Let R_{jk} be a rectangular partition of D as defined earlier. For any choice of sample points (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 $R(f, N_1, N_2)$ converges to the double integral of f over D by Theorem 14.1. Since the limit of the double sequence $R(f, N_1, N_2)$ exists, it should not depend on the order in which the limits $N_1 \rightarrow \infty$ (or $\Delta x \rightarrow 0$) and $N_2 \rightarrow \infty$ (or $\Delta y \rightarrow 0$) are computed (Theorem 14.6). Suppose the limit $\Delta y \rightarrow 0$ is to be evaluated first:

$$\begin{aligned} \iint_D f dA &= \lim_{N_{1,2} \rightarrow \infty} R(f, N_1, N_2) \\ &= \lim_{N_1 \rightarrow \infty} \sum_{j=1}^{N_1} \left(\lim_{N_2 \rightarrow \infty} \sum_{k=1}^{N_2} f(x_j^*, y_k^*) \Delta y \right) \Delta x. \end{aligned}$$

The expression in parentheses 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,

$$(14.4) \quad \lim_{N_2 \rightarrow \infty} \sum_{k=1}^{N_2} f(x_j^*, y_k^*) \Delta y = \int_c^d f(x_j^*, y) dy.$$

Define a function $A(x)$ by

$$(14.5) \quad A(x) = \int_c^d f(x, y) dy.$$

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 (see the right panel of Figure 14.5).

The second sum in the Riemann sum for the double integral in the Riemann sum of $A(x)$ on the interval $[a, b]$:

$$\begin{aligned} \iint_D f dA &= \lim_{N_1 \rightarrow \infty} \sum_{j=1}^{N_1} A(x_j^*) \Delta x = \int_a^b A(x) dx \\ &= \int_a^b \left(\int_c^d f(x, y) dy \right) dx, \end{aligned}$$

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 parentheses 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 $\Delta x \rightarrow 0$ 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.

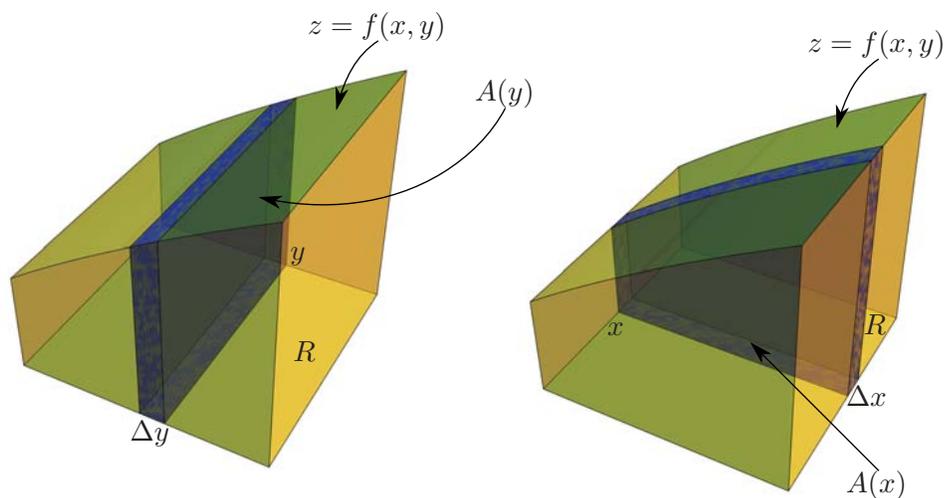


FIGURE 14.5. An illustration to Fubini's theorem. The volume of a solid below the graph $z = f(x, y)$ and above a rectangle R is the sum of the volumes of the slices. **Left:** The slicing is done parallel to the x axis so that the volume of each slice is $\Delta y A(y)$, where $A(y)$ is the area of the cross section by a plane with a fixed value of y . **Right:** The slicing is done parallel to the y axis so that the volume of each slice is $\Delta x A(x)$, where $A(x)$ is the area of the cross section by a plane with a fixed value of x as given in (14.5).

THEOREM 14.7. (Fubini's Theorem).

If f is continuous on the rectangle $D = [a, b] \times [c, d]$, then

$$\iint_D f(x, y) dA = \int_c^d \int_a^b f(x, y) dx dy = \int_a^b \int_c^d f(x, y) dy dx.$$

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 $(x, y) \in [0, 1] \times [0, 1]$.

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 14.2. (Factorization of Iterated Integrals).

Let D be a rectangle $[a, b] \times [c, d]$. Suppose $f(x, y) = g(x)h(y)$, where the functions g and h are integrable on $[a, b]$ and $[c, d]$, respectively. Then

$$\iint_D f(x, y) dA = \int_a^b g(x) dx \int_c^d h(y) dy.$$

So the double integral becomes the product of two ordinary integrals in this case. 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 $[0, \pi] \times [-\pi/2, \pi/2]$.

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,

$$\begin{aligned}\iint_{D_2} f \, dA &= \iint_D f \, dA + \iint_{D_1} f \, dA \quad \Rightarrow \\ \iint_D f \, dA &= \iint_{D_2} f \, dA - \iint_{D_1} f \, dA.\end{aligned}$$

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, $1024 - 4 = 1020$. \square

99.2. Study Problem.

Problem 14.1. Suppose a function f has continuous second derivatives on the rectangle $D = [0, 1] \times [0, 1]$. Find $\iint_D f''_{xy} \, dA$ if $f(0, 0) = 1$, $f(0, 1) = 2$, $f(1, 0) = 3$, and $f(1, 1) = 5$.

SOLUTION: By Fubini's theorem,

$$\begin{aligned}\iint_D f''_{xy} \, dA &= \int_0^1 \int_0^1 \frac{\partial}{\partial x} f'_y(x, y) \, dx \, dy = \int_0^1 f'_y(x, y) \Big|_0^1 \, dy \\ &= \int_0^1 [f'_y(1, y) - f'_y(0, y)] \, dy = \int_0^1 \frac{d}{dy} [f(1, y) - f(0, y)] \, dy \\ &= [f(1, y) - f(0, y)] \Big|_0^1 \\ &= [f(1, 1) - f(0, 1)] - [f(1, 0) - f(0, 0)] = 1.\end{aligned}$$

By Clairaut's theorem, $f''_{xy} = f''_{yx}$, and the value of the integral is independent of the order of integration. \square

99.3. Exercises.

(1) Evaluate the following double integrals over specified rectangular regions:

- (i) $\iint_D (x + y) \, dA$, $D = [0, 1] \times [0, 2]$
- (ii) $\iint_D xy^2 \, dA$, $D = [0, 1] \times [-1, 1]$
- (iii) $\iint_D \sqrt{x + 2y} \, dA$, $D[1, 2] \times [0, 1]$
- (iv) $\iint_D (1 + 3x^2y) \, dA$, $D = [0, 1] \times [0, 2]$
- (v) $\iint_D xe^{yx} \, dA$, $D = [0, 1] \times [0, 1]$
- (vi) $\iint_D \cos(x + 2y) \, dA$, $D = [0, \pi] \times [0, \pi/4]$

- (vii) $\iint_D \frac{1+2x}{1+y^2} dA$, $D = [0, 1] \times [0, 1]$
 (viii) $\iint_D \frac{y}{x^2+y^2} dA$, $D = [0, 1] \times [1, 2]$
 (ix) $\iint_D (x-y)^n dA$, $D = [0, 1] \times [0, 1]$, where n is a positive integer
 (x) $\iint_D e^x \sqrt{y+e^x} dA$, $D = [0, 1] \times [0, 2]$
 (xi) $\iint_D \sin^2(x) \sin^2(y) dA$, $D = [0, \pi] \times [0, \pi]$
 (xii) $\iint_D \ln(x+y) dA$, $D = [1, 2] \times [1, 2]$
 (xiii) $\iint_D \frac{1}{2x+y} dA$, $D = [0, 1] \times [1, 2]$
 (xiv) $\iint_D x2^{x-y} dA$, $D = [0, 1] \times [0, 1]$
 (xv) $\iint_D \frac{1-(xy)^3}{1-xy} dA$, $D = [0, \frac{1}{2}] \times [0, \frac{1}{2}]$
- (2) Find the volume of each of the following solids E :
- E lies under the paraboloid $z = 1 + 3x^2 + 6y^2$ and above the rectangle $[-1, 1] \times [0, 2]$.
 - E lies in the first octant and is bounded by the cylinder $z = 4 - y^2$ and the plane $x = 3$.
 - E lies in the first octant and is bounded by the planes $x + y - z = 0$, $y = 2$, and $x = 1$.
- (3) Evaluate $\iint_D xy dA$, where D is the part of the square $[-1, 1] \times [-1, 1]$ that does not lie in the first quadrant.
- (4) Let f be continuous on $[a, b] \times [c, d]$ and let $g(u, v) = \iint_{D_{uv}} f(x, y) dA$, where $D_{uv} = [a, u] \times [b, v]$ for $a < u < b$ and $c < v < d$. Show that $g''_{uv} = g''_{vu} = f(u, v)$.
- (5) Let $f(x)$ be continuous in $[a, b]$. Prove that

$$\left(\int_a^b f(x) dx \right)^2 \leq (b-a) \int_a^b (f(x))^2 dx,$$

where the equality is reached only if $f(x) = \text{const}$. *Hint:* Consider the iterated integral

$$\int_a^b \int_a^b [f(x) - f(y)]^2 dy dx.$$

- (6) Find the average value of the squared distance from the origin to a point of the disk $(x-a)^2 + (y-b)^2 \leq R^2$.

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 14.9. (Simple and Convex Regions).

A region D is said to be simple in the direction \mathbf{u} if any line parallel

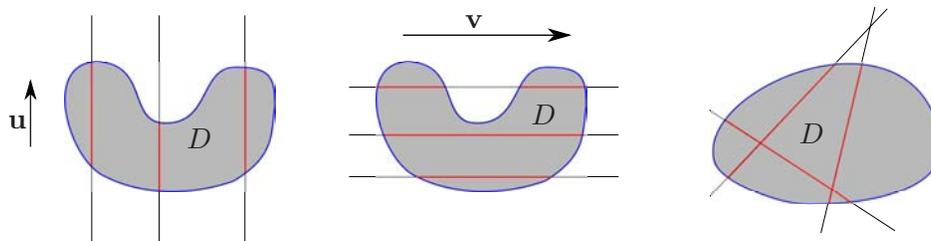


FIGURE 14.6. **Left:** A region D is simple in the direction \mathbf{u} . **Middle:** A region D is not simple in the direction \mathbf{v} . **Right:** A region D is simple or convex. Any straight line intersects it along at most one segment, or a straight line segment connecting any two points of D lies in D .

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.

This definition is illustrated in Figure 14.6. 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 $a \leq x \leq b$. 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 $x_0 \in [a, b]$, $y_{\text{bot}} \leq y \leq y_{\text{top}}$, where the numbers y_{bot} and y_{top} depend on x_0 . So all 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 $a \leq x \leq b$ and is bounded from below by the graph $y = y_{\text{bot}}(x)$ and from above by the graph $y = y_{\text{top}}(x)$:

$$(14.6) \quad D = \{(x, y) \mid y_{\text{bot}}(x) \leq y \leq y_{\text{top}}(x), \quad x \in [a, b]\}.$$

The numbers a and b are, respectively, the smallest and the largest values of the x coordinate of points of D .

EXAMPLE 14.5. Give an algebraic description of the half-disk $x^2 + y^2 \leq 1$, $y \geq 0$, as a vertically simple region.

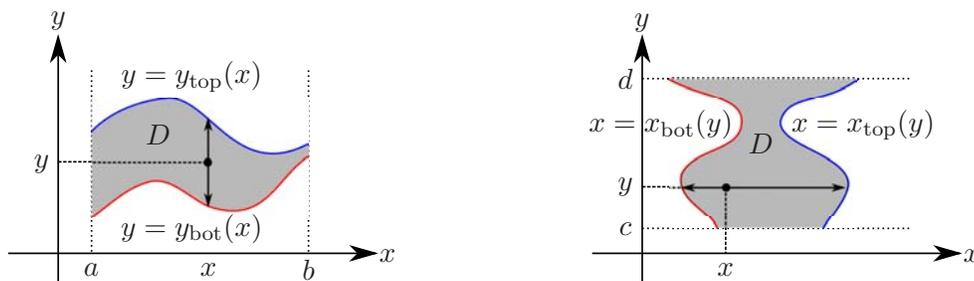


FIGURE 14.7. Left: An algebraic description of a vertically simple region as given in (14.6): for every $x \in [a, b]$, the y coordinate ranges over the interval $y_{bot}(x) \leq y \leq y_{top}(x)$.

Right: An algebraic description of a horizontally simple region D as given in (14.7): for every $y \in [c, d]$, the x coordinate ranges over the interval $x_{bot}(y) \leq x \leq x_{top}(y)$.

SOLUTION: The x coordinate of any point in the disk lies in the interval $[a, b] = [-1, 1]$ (see Figure 14.8, left panel). Take a vertical line corresponding to a fixed value of x in this interval. This line intersects the half-disk along the segment whose one endpoint lies on the x axis; that is, $y = 0 = y_{bot}(x)$. The other endpoint lies on the circle. Solving the equation of the circle for y , one finds $y = \pm\sqrt{1-x^2}$. Since $y \geq 0$ in the half-disk, the positive solution has to be taken, $y = \sqrt{1-x^2} = y_{top}(x)$. So the region is bounded by two graphs $y = 0$ and $y = \sqrt{1-x^2}$. For every $-1 \leq x \leq 1$, $0 \leq y \leq \sqrt{1-x^2}$. \square

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 $c \leq y \leq d$. 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 $y_0 \in [c, d]$, $x_{bot} \leq x \leq x_{top}$, where the numbers x_{bot} and x_{top} depend on y_0 . So all 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 $c \leq y \leq b$ and is bounded from below by the graph $x = x_{bot}(y)$ and from above by the graph $x = x_{top}(y)$:

$$(14.7) \quad D = \{(x, y) \mid x_{bot}(y) \leq x \leq x_{top}(y), \quad y \in [c, d]\}.$$

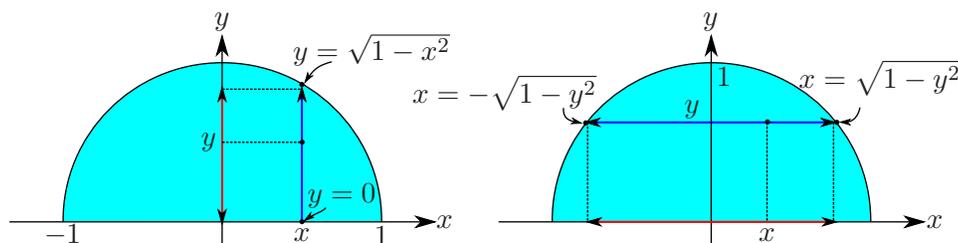


FIGURE 14.8. The half-disk D , $x^2 + y^2 \leq 1$, $y \geq 0$, is a simple region. **Left:** An algebraic description of D as a vertically simple region as given in (14.6). The maximal range of x in D is $[-1, 1]$. For every such x , the y coordinate in D has the range $0 \leq y \leq \sqrt{1 - x^2}$. **Right:** An algebraic description of D as a horizontally simple region as given in (14.7). The maximal range of y in D is $[0, 1]$. For every such y , the x coordinate in D has the range $-\sqrt{1 - y^2} \leq x \leq \sqrt{1 - y^2}$.

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.

EXAMPLE 14.6. Give an algebraic description of the half-disk $x^2 + y^2 \leq 1$, $y \geq 0$, as a horizontally simple region.

SOLUTION: The y coordinate of any point in the disk lies in the interval $[c, d] = [0, 1]$. Take a horizontal line corresponding to a fixed value of y from this interval. The line intersects the half-disk along a segment whose endpoints lie on the circle. Solving the equation of the circle for x , the x coordinates of the endpoints are obtained: $x = \pm\sqrt{1 - y^2}$. So, for every $0 \leq y \leq 1$, $-\sqrt{1 - y^2} \leq x \leq \sqrt{1 - y^2}$. When viewed in the horizontal direction, the top boundary of the region is the graph $x = \sqrt{1 - y^2} = x_{\text{top}}(y)$ and the bottom boundary is the graph $x = -\sqrt{1 - y^2} = x_{\text{bot}}(y)$ (see Figure 14.8, right panel). \square

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 $c \leq y_{\text{bot}}(x) \leq y_{\text{top}}(x) \leq d$ 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 $c \leq y < y_{\text{bot}}(x)$ and $y_{\text{top}}(x) < y \leq d$, 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 $\Delta x \rightarrow 0$ and $\Delta y \rightarrow 0$ are taken (Theorem 14.6). For a vertically simple D , the limit $\Delta y \rightarrow 0$ is taken first. Similarly to (14.4), one infers that

$$\lim_{N_2 \rightarrow \infty} \sum_{k=1}^{N_2} f(x_j^*, y_k^*) \Delta y = \int_c^d f(x_j^*, y) dy = \int_{y_{\text{bot}}(x_j^*)}^{y_{\text{top}}(x_j^*)} f(x_j^*, y) dy$$

because the function f vanishes outside the interval $y_{\text{bot}}(x) \leq y \leq y_{\text{top}}(x)$ for any $x \in [a, b]$.

Suppose that $f(x, y) \geq 0$ and consider the solid bounded from above by the graph $z = f(x, y)$ and from below by the region D . The area of the cross section of the solid by the coordinate plane corresponding to a fixed value of x is given by (14.5):

$$A(x) = \int_c^d f(x, y) dy = \int_{y_{\text{bot}}(x)}^{y_{\text{top}}(x)} f(x, y) dy.$$

So just like in the case of rectangular domains, the above limit equals $A(x_j^*)$. That the area is given by an integral over a *single* interval 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 $[a, b]$. Recall that the volume of a slice of width dx and cross section area $A(x)$ is $dV = A(x) dx$ so that the total volume of the solid is given by the integral $V = \int_a^b A(x) dx$ (as the sum of volumes of all slices in the solid).

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

$$(14.8) \quad \iint_D f(x, y) dA = \int_a^b \int_{y_{\text{bot}}(x)}^{y_{\text{top}}(x)} f(x, y) dy dx.$$

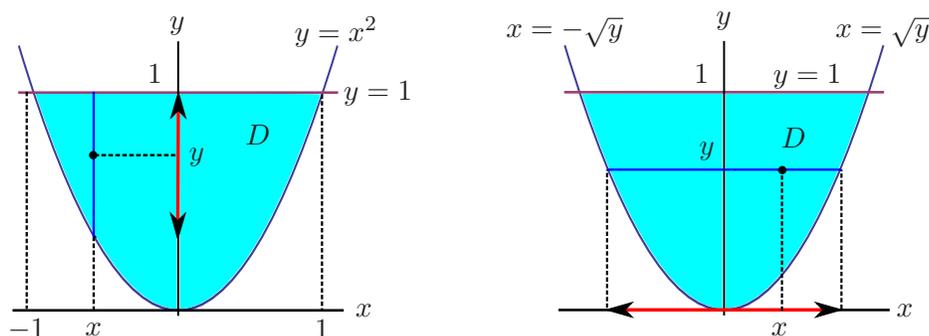


FIGURE 14.9. Illustration to Example 14.8. **Left:** The integration region as a vertically simple region: $-1 \leq x \leq 1$ and, for every such x , $x^2 \leq y \leq 1$. **Right:** The integration region as a horizontally simple region: $0 \leq y \leq 1$ and, for every such y , $-\sqrt{y} \leq x \leq \sqrt{y}$.

Iterated Integral for Horizontally Simple Regions. Naturally, for horizontally simple regions, the integration with respect to x should be carried out first. Therefore, the limit $\Delta x \rightarrow 0$ should be taken first in the Riemann sum. The technicalities are similar to the case of vertically simple regions. 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

$$(14.9) \quad \iint_D f(x, y) \, dA = \int_c^d \int_{x_{\text{bot}}(y)}^{x_{\text{top}}(y)} f(x, y) \, dx \, dy.$$

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 by suitable smooth curves into simple regions D_p , $p = 1, 2, \dots, n$. 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. Sometimes, it is also convenient to cut the integration region into two or more pieces even if the region is simple (see Example 14.8).

EXAMPLE 14.7. 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 (see the left panel of Figure 14.9). For any $x \in [-1, 1]$, the y coordinate of points of D attains the smallest value on the parabola (i.e., $y_{\text{bot}}(x) = x^2$), and the largest value on the line (i.e., $y_{\text{top}}(x) = 1$). 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 on the parabola $y = x^2$ or $x = \pm\sqrt{y}$, that is, $x_{\text{bot}}(y) = -\sqrt{y}$ and $x_{\text{top}}(y) = \sqrt{y}$ (see the right panel of Figure 14.9). 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 14.8. 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 $[a, b] = [-1, 8]$. For any fixed $x \in [-1, 0]$, the range of the y coordinate is determined by the parabola $x = y^2 - 1$.

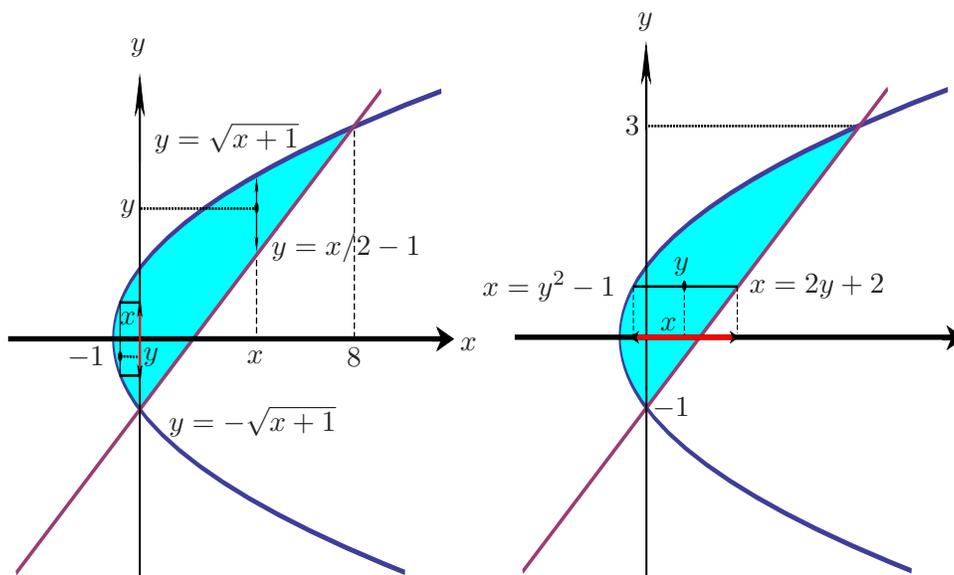


FIGURE 14.10. Illustration to Example 14.8. **Left:** The integration region D as a vertically simple region. An algebraic description requires to splitting the maximal range of x into two intervals. For every $-1 \leq x \leq 0$, the y coordinate ranges over the interval $-\sqrt{x+1} \leq y \leq \sqrt{x+1}$, whereas for every $0 \leq x \leq 8$, $x/2 - 1 \leq y \leq \sqrt{x+1}$. Accordingly, when converting the double integral to the iterated integral, the region D has to be split into two parts in which $x \leq 0$ and $x \geq 0$. **Right:** The integration region D as a horizontally simple region. For every $-1 \leq y \leq 3$, the x coordinate ranges over the interval $y^2 - 1 \leq x \leq 2y + 2$. So the double integral can be converted to a single iterated integral.

Solutions of this equation are $y = \pm\sqrt{x+1}$, and the range of the y coordinate is $-\sqrt{x+1} \leq y \leq \sqrt{x+1}$. For any fixed $x \in [0, 8]$, the largest value of y still occurs on the parabola, $y = \sqrt{x+1}$, while the smallest value occurs on the line, $x = 2y + 2$ or $y = (x-2)/2$, so that $-\sqrt{x+1} \leq y \leq (x-2)/2$. The boundaries of D are

$$y = y_{\text{top}}(x) = \sqrt{x+1}, \quad y = y_{\text{bot}}(x) = \begin{cases} -\sqrt{x+1} & \text{if } -1 \leq x \leq 0 \\ x/2 - 1 & \text{if } 0 \leq x \leq 8 \end{cases}.$$

That the bottom boundary consists of two graphs 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

$$\begin{aligned}\iint_D 2x \, dA &= \iint_{D_1} 2x \, dA + \iint_{D_2} 2x \, dA \\ &= 2 \int_{-1}^0 x \int_{-\sqrt{x+1}}^{\sqrt{x+1}} dy \, dx + 2 \int_0^8 x \int_{-\sqrt{x+1}}^{x/2+1} dy \, dx.\end{aligned}$$

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, $x_{\text{bot}}(y) = y^2 - 1 \leq x \leq 2y + 2 = x_{\text{top}}(y)$. 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. \square

Sometimes the iterated integration cannot even be carried out in one order, but it can still be done in the other order.

EXAMPLE 14.9. Evaluate the double integral of $f(x, y) = \sin(y^2)$ over the region D , which is the triangle bounded by the lines $x = 0$, $y = x$, and $y = \sqrt{\pi}$.

SOLUTION: Suppose that 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 $y_{\text{bot}}(x) = x \leq y \leq \sqrt{\pi} = y_{\text{top}}(x)$ 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 $x_{\text{bot}}(y) = 0 \leq x \leq y = x_{\text{top}}(y)$ in D . Therefore, the iterated integral reads

$$\begin{aligned}\iint_D \sin(y^2) \, dA &= \int_0^{\sqrt{\pi}} \sin(y^2) \int_0^y dx \, dy \\ &= \int_0^{\sqrt{\pi}} \sin(y^2) y \, dy = -\frac{1}{2} \cos(y^2) \Big|_0^{\sqrt{\pi}} = 1.\end{aligned}$$

\square

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 is quite useful. For example, an indefinite integral of $\sin(x^{2011})$ cannot be expressed in elementary functions. Nevertheless, to find its definite integral over any *symmetric* interval $[-a, a]$, an explicit form of the indefinite integral is not necessary. Indeed, the function $\sin(x^{2011})$ is antisymmetric, and hence its integral over any symmetric interval vanishes. 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 $(x, y) \rightarrow (x_s, y_s)$ 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, $(x, y) \rightarrow (x_s, y_s) = (-x, y)$, $(x, y) \rightarrow (x_s, y_s) = (x, -y)$, or $(x, y) \rightarrow (x_s, y_s) = (-x, -y)$. A transformation of the plane $(x, y) \rightarrow (x_s, y_s)$ 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 14.8. (Symmetry Property).

Let a region D be symmetric under an area-preserving transformation $(x, y) \rightarrow (x_s, y_s)$ 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 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} , $p = 1, 2, \dots, N$. 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 as shown in the left panel of Figure 14.11. 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}, y_{ps})

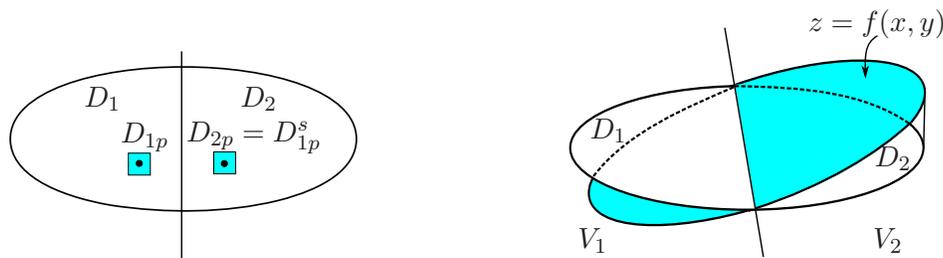


FIGURE 14.11. **Left:** The region D is symmetric relative to the reflection about the line. Under this reflection, $D_1 \rightarrow D_2$ and $D_2 \rightarrow D_1$. Any partition of D_1 by elements D_{1p} induces the partition of D_2 by taking the images of D_{1p} under the reflection. **Right:** The graph of a function f that is skew-symmetric under the reflection. If f is positive in D_2 , then it is negative in D_1 . The volume V_2 of the solid below the graph and above D_2 is exactly the same as the volume $V_1 = V_2$ of the solid above the graph and below D_1 . But the latter solid lies below the xy plane, and hence the double integral over D is $V_2 - V_1 = 0$.

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:

$$\iint_D f \, dA = \lim_{N \rightarrow \infty} \sum_{p=1}^N \left(f(x_p, y_p) \Delta A_p + f(x_{ps}, y_{ps}) \Delta A_p \right) = 0,$$

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 (see the right panel of Figure 14.11).

EXAMPLE 14.10. Evaluate the double integral of $\sin[(x - y)^3]$ over the portion D of the disk $x^2 + y^2 \leq 1$ that lies in the first quadrant ($x, y \geq 0$).

SOLUTION: The region D is symmetric under the reflection about the line $y = x$ (see the left panel of Figure 14.12), that is, $(x, y) \rightarrow (x_s, y_s) = (y, x)$, whereas the function is antisymmetric, $f(x_s, y_s) = f(y, x) = \sin[(y - x)^3] = \sin[-(x - y)^3] = -\sin[(x - y)^3] = -f(x, y)$. By the symmetry property, the double integral vanishes. \square

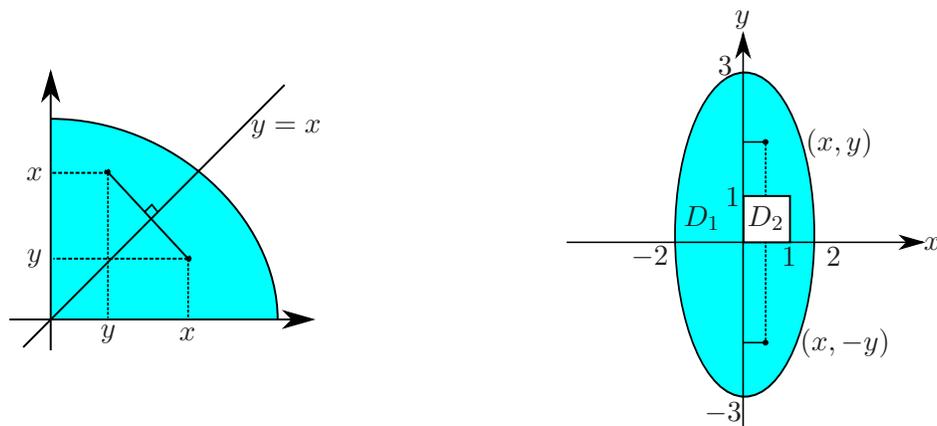


FIGURE 14.12. **Left:** Illustration to Example 14.8. The region is symmetric under the reflection about the line $y = x$. **Right:** The integration region D in Example 14.9. It can be viewed as the difference of the elliptic region D_1 and the square D_2 . The elliptic region is symmetric under the reflection about the x axis, whereas the function $f(x, y) = x^2y^3$ is skew-symmetric, $f(x, -y) = -f(x, y)$. So the integral over D_1 must vanish, and the double integral over D is the negative of the integral over D_2 .

EXAMPLE 14.11. 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 as shown in the right panel of Figure 14.12. Therefore, the additivity of the double integral can be used (compare Example 14.4) to transform the double integral over a nonsimple 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 $(x, y) \rightarrow (x_s, y_s) = (x, -y)$, whereas the integrand is antisymmetric, $f(x, -y) = x^2(-y)^3 = -x^2y^3 = -f(x, y)$. \square

100.5. Study Problems.

Problem 14.2. *Prove the Dirichlet formula*

$$\int_0^a \int_0^x f(x, y) dy dx = \int_0^a \int_y^a f(x, y) dx dy, \quad a > 0.$$

SOLUTION: The left side of the equation is an iterated integral for the double integral $\iint_D f dA$. Let us find the shape of D . According to the limits of integration, D admits the following algebraic description (as a vertically simple region). For every $0 \leq x \leq a$, the y coordinate changes in the interval $0 \leq y \leq x$. So the region D is the triangle bounded by the lines $y = 0$, $y = x$, and $x = a$. To reverse the order of integration, let us find an algebraic description of D as a horizontally simple region. The maximal range of y in D is the interval $[0, a]$. For every fixed $0 \leq y \leq a$, the x coordinate spans the interval $y \leq x \leq a$ in D . So the two sides of the Dirichlet formula represent the same double integral as iterated integrals in different orders and hence are equal. \square

Problem 14.3. *Reverse the order of integration*

$$\int_1^2 \int_{2-x}^{\sqrt{2x-x^2}} f(x, y) dy dx.$$

SOLUTION: The given iterated integral represents a double integral $\iint_D f dA$, where the integration region admits the following description (as a vertically simple region). For every fixed $1 \leq x \leq 2$, the y coordinates span the interval $2 - x \leq y \leq \sqrt{2x - x^2}$. So D is bounded by the graphs $y = 2 - x$ (a line) and $y = \sqrt{2x - x^2}$ or $y^2 = 2x - x^2$ or, after completing the squares, $(x - 1)^2 + y^2 = 1$ (a circle of radius 1 centered at $(1, 0)$). The circle and the line intersect at the points $(1, 1)$ and $(2, 0)$. Thus, the region D is the part of the disk $(x - 1)^2 + y^2 \leq 1$ that lies above the line $y = 2 - x$. The reader is advised to sketch it. To reverse the order of integration, let us find an algebraic description of D as a horizontally simple region. The maximal range of y is the interval $[0, 1]$, which is determined by the points of intersection of the circle and the line. Viewing the region D along the x axis, one can see that, for every fixed $0 \leq y \leq 1$, the smallest value of x in D is attained on the line $y = 2 - x$ or $x = 2 - y = x_{\text{bot}}(y)$, while its greatest value in D is attained on the circle $(x - 1)^2 + y^2 = 1$ or $x - 1 = \pm \sqrt{1 - y^2}$ or $x = 1 + \sqrt{1 - y^2} = x_{\text{top}}(y)$ because the solution with the plus sign corresponds to the part of the circle that lies above the line. Hence,

the integral in the reversed order reads

$$\int_0^1 \int_{2-y}^{1+\sqrt{1-y^2}} f(x, y) dx dy.$$

□

100.6. Exercises.

(1) For each of the two orders of integration, specify the limits in the iterated integrals for $\iint_D f(x, y) dA$, splitting the integration region when necessary, if

- (i) D is the triangle with vertices $(0, 0)$, $(2, 1)$, and $(-2, 1)$
- (ii) D is a the trapezoid with vertices $(0, 0)$, $(1, 0)$, $(1, 2)$, and $(0, 1)$
- (iii) D is the disk $x^2 + y^2 \leq 1$
- (iv) D is the disk $x^2 + y^2 \leq y$
- (v) D is the ring $1 \leq x^2 + y^2 \leq 4$

(2) Evaluate the following double integrals over the specified region:

- (i) $\iint_D xy dA$, where D is bounded by the curves $y = x^2$ and $y = x$
- (ii) $\iint_D (2 + y) dA$, where D is the region bounded by the graphs of $x = 3$ and $x = 4 - y^2$
- (iii) $\iint_D (x + y) dA$, where D is bounded by the curves $x = y^4$ and $x = y$
- (iv) $\iint_D (2 + y) dA$, where D is the region bounded by the three lines of $x = 3$, $y + x = 0$, and $y - x = 0$; find the value of the integral by geometric means
- (v) $\iint_D x^2 y dA$, where D is the region bounded by the graphs of $y = 2 + x^2$ and $y = 4 - x^2$
- (vi) $\iint_D \sqrt{1 - y^2} dA$, where D is the triangle with vertices $(0, 0)$, $(0, 1)$, and $(1, 1)$
- (vii) $\iint_D xy dA$, where D is bounded by the lines $y = 1$, $x = -3y$, and $x = 2y$
- (viii) $\iint_D y\sqrt{x^2 - y^2} dA$, where D is the triangle with vertices $(0, 0)$, $(0, 1)$, and $(1, 1)$
- (ix) $\iint_D (2a - x)^{-1/2} dA$, where D is bounded by the coordinate axes and by the shortest arc of the circle of radius a and centered at (a, a)
- (x) $\iint_D |xy| dA$, where D is the disk of radius a centered at the origin
- (xi) $\iint_D (x^2 + y^2) dA$, where D is the parallelogram with the sides $y = x$, $y = x + a$, $y = a$, and $y = 3a$ ($a > 0$)

- (xii) $\iint_D y^2 dA$, where D is bounded by the x axis and by one arc of the cycloid $x = a(t - \sin t)$, $y = a(1 - \cos t)$, $0 \leq t \leq 2\pi$
- (3) Sketch the solid region whose volume is given by the following integrals:

- (i) $\int_0^1 \int_0^{1-x} (x^2 + y^2) dy dx$
- (ii) $\iint_D (x + y) dA$, where D is defined by the inequalities $0 \leq x + y \leq 1$, $x \geq 0$, and $y \geq 0$
- (iii) $\iint_D \sqrt{x^2 + y^2} dA$, where D is defined by the inequality $x^2 + y^2 \leq x$
- (iv) $\iint_D (x^2 + y^2) dA$, where D is defined by the inequality $|x| + |y| \leq 1$
- (v) $\iint_D \sqrt{1 - (x/2)^2 - (y/3)^2} dA$, where D is defined by the inequality $(x/2)^2 + (y/3)^2 \leq 1$

- (4) Use the double integral to find the volume of the specified solid region E :

- (i) E is bounded by the plane $x + y + z = 1$ and the coordinate planes.
- (ii) E lies under the paraboloid $z = 2x^2 + y^2$ and above the region in the xy plane bounded by the curves $x = y^2$ and $x = 1$.
- (iii) E is bounded by the cylinder $x^2 + y^2 = 1$ and the planes $y = z$, $x = 0$, and $z = 0$ in the first octant.
- (iv) E is bounded by the cylinders $x^2 + y^2 = a^2$ and $y^2 + z^2 = a^2$.
- (v) E is enclosed by the parabolic cylinders $y = 1 - x^2$, $y = x^2 - 1$ and the planes $x + y + z = 2$, $2x + 2y - z = 10$.

- (5) Sketch the region of integration and reverse the order of integration in each of the following iterated integrals. Evaluate the integral if the integrand is specified:

- (i) $\int_1^4 \int_{\sqrt{y}}^2 f(x, y) dx dy$
- (ii) $\int_0^{\sqrt{\pi}} \int_y^{\sqrt{\pi}} \cos(x^2) dx dy$ *Hint:* After reversing the integration order, make the substitution $u = x^2$ to do the integral.
- (iii) $\int_0^1 \int_{x^3}^{\sqrt{x}} f(x, y) dy dx$
- (iv) $\int_0^1 \int_{y^2}^y f(x, y) dx dy$
- (v) $\int_0^1 \int_1^{e^x} f(x, y) dy dx$
- (vi) $\int_1^4 \int_{\sqrt{y}}^2 f(x, y) dx dy$
- (vii) $\int_0^3 \int_0^{y/3} f(x, y) dx dy + \int_3^6 \int_0^{6-y} f(x, y) dx dy$
- (viii) $\int_0^4 \int_{\sqrt{x}}^2 (1 + y^3)^{-1} dy dx$
- (ix) $\int_{-6}^2 \int_{(x^2/4)-1}^{2-x} f(x, y) dy dx$

- (x) $\int_{-1}^1 \int_{-\sqrt{1-x^2}}^{1-x^2} f(x, y) dy dx$
- (xi) $\int_0^{2a} \int_{\sqrt{2ax-x^2}}^{\sqrt{2ax}} f(x, y) dy dx$ ($a > 0$)
- (xii) $\int_0^{2\pi} \int_0^{\sin x} f(x, y) dy dx$
- (6) Use the symmetry and the properties of the double integral to find:
- (i) $\iint_D e^{x^2} \sin(y^3) dA$, where D is the triangle with vertices $(0, 1)$, $(0, -1)$, and $(1, 0)$
- (ii) $\iint_D (y^9 + px^9) dA$, where $p = \pm 1$ and $D = \{(x, y) | 1 \leq |x| + |y| \leq 2\}$
- (iii) $\iint_D x dA$, where D is bounded by the ellipse $x^2/a^2 + y^2/b^2 = 1$ and has the triangular hole with vertices $(0, b)$, $(0, -b)$, and $(a, 0)$
- (iv) $\iint_D (\cos(x^2) + \sin(y^2)) dA$, where D is the disk $x^2 + y^2 \leq a^2$
- (7) Find the area of the following regions:
- (i) D is bounded by the curves $xy = a^2$ and $x + y = 5a/2$, $a > 0$.
- (ii) D is bounded by the curves $y^2 = 2px + p^2$ and $y^2 = -2qx + q^2$, where p and q are positive numbers.
- (iii) D is bounded by $(x - y)^2 + x^2 = a^2$.

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 $(x, y) \neq (0, 0)$ of the plane and points of the strip $(r, \theta) \in (0, \infty) \times [0, 2\pi)$. The pairs $(r, \theta) = (0, \theta)$ correspond to the origin $(x, y) = (0, 0)$. 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.

The relations $x = r \cos \theta$, $y = r \sin \theta$ define a *transformation* of any region D' in the polar plane to a region D in the xy plane; that is, to every ordered pair (r, θ) corresponding to a point of D' , an ordered pair (x, y) corresponding to a point of D is assigned. Accordingly, the *inverse transformation* $r = \sqrt{x^2 + y^2}$, $\theta = \tan^{-1}(y/x)$ maps a region

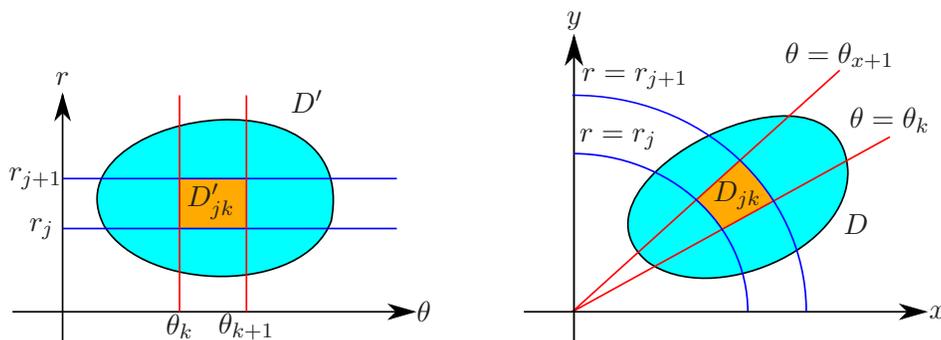


FIGURE 14.13. **Left:** A partition of D' by the coordinate lines $r = r_j$ and $\theta = \theta_k$, where $r_{j+1} - r_j = \Delta r$ and $\theta_{k+1} - \theta_k = \Delta\theta$. A partition element is a rectangle D'_{jk} . Its area is $\Delta A'_{jk} = \Delta r \Delta\theta$. **Right:** A partition of D by the images of the coordinate curves $r = r_j$ (concentric circles) and $\theta = \theta_k$ (rays extended from the origin). A partition element D_{jk} is the image of the rectangle D'_{jk} . Its area is $\Delta A_{jk} = \frac{1}{2}(r_{j+1}^2 - r_j^2)\Delta\theta = \frac{1}{2}(r_{j+1} + r_j)\Delta A'_{jk}$.

D in the xy plane to a region D' in the polar plane. 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 $x^2 + y^2 \leq 1$ in the first quadrant. Then the shape of D' can be found from the images of boundaries of D in the polar plane:

boundaries of $D \leftrightarrow$ boundaries of D'

$$x^2 + y^2 = 1 \leftrightarrow r = 1$$

$$y = 0, x \geq 0 \leftrightarrow \theta = 0$$

$$x = 0, y \geq 0 \leftrightarrow \theta = \pi/2$$

Since $r \geq 0$, the region D' is the rectangle $(r, \theta) \in [0, 1] \times [0, \pi/2] = D'$. The boundary of D' always contains $r = 0$ if the origin belongs to D . If D is invariant under rotations about the origin (a disk or a ring), then θ takes its full range $[0, 2\pi]$ in D' .

Let D' be a region in the polar plane and let D be its image in the xy plane. 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 R'_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 $\theta = \theta_{k+1} = \theta_k + \Delta\theta$ as shown in Figure 14.13 (left panel). 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 (concentric circles and rays, respectively). A rectangular partition of D' induces a partition of D by coordinate curves of the polar coordinates. 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 $(\Delta r, \Delta\theta) \rightarrow (0, 0)$. 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)

$$\lim_{\substack{N \rightarrow \infty \\ (R_N \rightarrow 0)}} \sum_{p=1}^N f(\mathbf{r}_p^*) \Delta A_p = \lim_{\substack{N_{1,2} \rightarrow \infty \\ (\Delta r, \Delta\theta) \rightarrow (0,0)}} \sum_{j=1}^{N_1} \sum_{k=1}^{N_2} f(r_j^* \cos \theta_k^*, r_j^* \sin \theta_k^*) r_j^* \Delta A',$$

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 14.10. (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

$$\iint_D f(x, y) dA = \iint_{D'} f(r \cos \theta, r \sin \theta) J(r) dA', \quad J(r) = r.$$

In particular, the area of a region D is given by the double integral

$$A(D) = \iint_D dA = \iint_{D'} r dA'$$

in the polar coordinates. 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 .* Suppose that D' is a vertically simple region as shown in Figure 14.14 (right panel):

$$D' = \{(r, \theta) | r_{\text{bot}}(\theta) \leq r \leq r_{\text{top}}(\theta), \theta_1 \leq \theta \leq \theta_2\}.$$

Then D is bounded by the *polar graphs* $r = r_{\text{bot}}(\theta)$, $r = r_{\text{top}}(\theta)$ and by the lines $y = \tan \theta_1 x$ and $y = \tan \theta_2 x$ (see the left panel of Figure 14.14). Recall that curves defined by the equation $r = g(\theta)$ are called *polar graphs*. 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 at a distance $r = g(\theta)$ from the origin. All such points obtained for all values of θ form the polar graph. The double integral over D can be written as the iterated integral over D' :

$$\iint_D f(x, y) dA = \int_{\theta_1}^{\theta_2} \int_{r_{\text{bot}}(\theta)}^{r_{\text{top}}(\theta)} f(r \cos \theta, r \sin \theta) r dr d\theta.$$

EXAMPLE 14.12. *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: When converting a double integral to polar coordinates, three essential steps should be followed.

- First, one has to find the region D' in the polar plane whose image is D under the transformation $x = r \cos \theta$, $y = r \sin \theta$. 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 integrand has to be written in polar coordinates. In this example, $f(r \cos \theta, r \sin \theta) = r^4 \cos \theta \sin^2 \theta$.
- Third, the double integral of the integrand written in polar coordinates and *multiplied by the Jacobian r* has to be evaluated over D' .

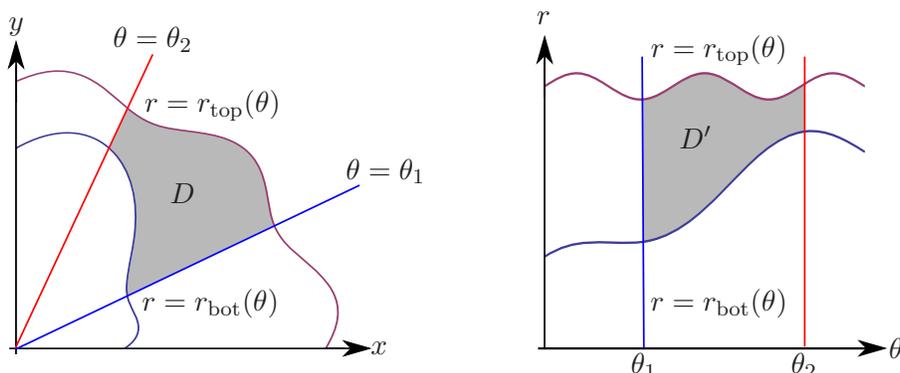


FIGURE 14.14. **Left:** In polar coordinates, the boundary of a region D , which is the image of a vertically simple region D' in the polar plane, can be viewed as polar graphs and lines through the origin. **Right:** A vertically simple region D' in the polar plane.

In this example, D' is a rectangle and hence, 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 using 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 $y/x = \tan \theta$.

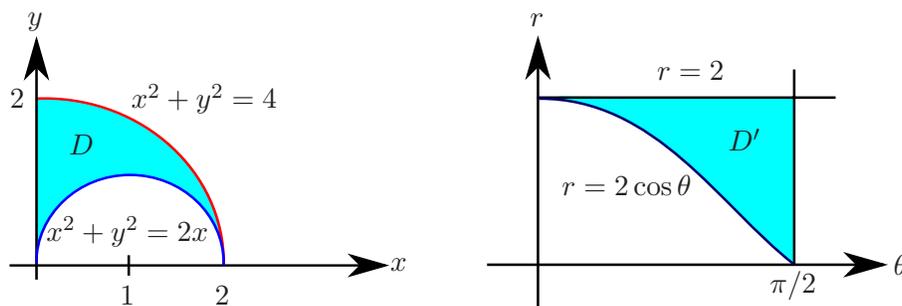


FIGURE 14.15. Illustration to Example 14.13.

Indeed, if D is bounded only by circles centered at the origin and lines through the origin, then 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 14.13. 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 region D' whose image is D must be found. Using the principle that the boundaries of D are related to the boundaries of D' by the change of variables, the equations of the boundaries of D' are obtained by converting the equations for the boundaries of D into polar coordinates. The boundary of the region D consists of three curves:

$$\begin{aligned} x^2 + y^2 = 4 &\rightarrow r^2 = 4 \rightarrow r = 2, \\ x^2 + y^2 = 2x &\rightarrow r^2 = 2r \cos \theta \rightarrow r = 2 \cos \theta, \\ x = 0, y \geq 0 &\rightarrow \theta = \pi/2. \end{aligned}$$

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 $\theta = \pi/2$. 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 $\theta \in [0, \pi/2] = [\theta_1, \theta_2]$ (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

$$\begin{aligned}
 \iint_D xy \, dA &= \iint_{D'} r^3 \sin \theta \cos \theta \, dA' = \int_{\theta_1}^{\theta_2} \sin \theta \cos \theta \int_{r_{\text{bot}}(\theta)}^{r_{\text{top}}(\theta)} r^3 \, dr \, d\theta \\
 &= \int_0^{\pi/2} \sin \theta \cos \theta \int_{2 \cos \theta}^2 r^3 \, dr \, d\theta \\
 &= 4 \int_0^{\pi/2} (1 - \cos \theta)^4 \cos \theta \sin \theta \, d\theta \\
 &= 4 \int_0^1 (1 - u)^4 u \, du = 4 \int_0^1 v^4 (1 - v) \, dv = \frac{4}{15},
 \end{aligned}$$

where two changes of variables have been used to simplify the calculations, $u = \cos \theta$ and $v = 1 - u$. \square

EXAMPLE 14.14. 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 polar graphs. Given a value of θ , $r = \theta$ (or $r = 2\theta$) is the distance from the point on the graph to the origin. As this distance increases monotonically with increasing θ , the polar graphs are spirals winding about the origin. The region D lies between two spirals; it is not simple in any direction (see the left panel of Figure 14.16). By converting the polar graph $r = \theta$ 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 by cutting the region D into simple pieces, 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 $y = 0$, $x \geq 0$. They are the images of the lines $r = \theta$, $r = 2\theta$, and $\theta = 2\pi$ in the polar plane as shown in the right panel of Figure 14.16. These lines form the boundaries of D' . An algebraic description of D' as a vertically simple region is convenient to use, $(r, \theta) \in D'$ if $r_{\text{bot}}(\theta) = \theta \leq r \leq 2\theta = r_{\text{top}}(\theta)$ and $\theta \in [0, 2\pi] = [\theta_1, \theta_2]$. Hence,

$$A(D) = \iint_D dA = \iint_{D'} r \, dA' = \int_0^{2\pi} \int_{\theta}^{2\theta} r \, dr \, d\theta = \frac{3}{2} \int_0^{2\pi} \theta^2 \, d\theta = 4\pi^3.$$

\square

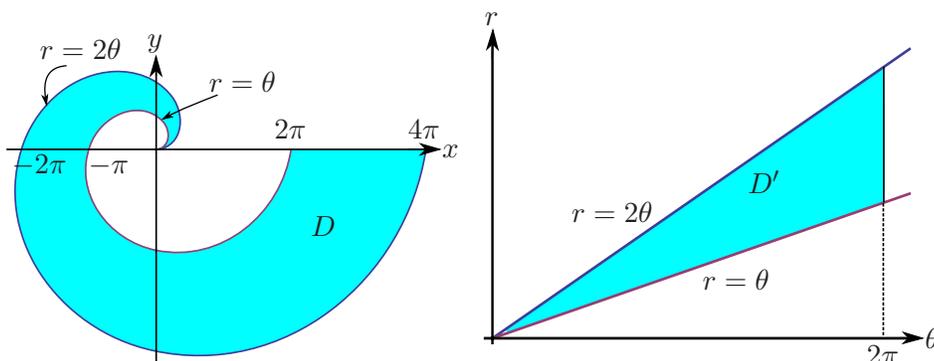


FIGURE 14.16. An illustration to Example 14.14. **Left:** The integration region D lies between two spirals. It is not simple in any direction. **Right:** The region D' in the polar plane whose image is D . The region D' is simple and is bounded by straight lines.

EXAMPLE 14.15. Find the volume of the part of the solid bounded by the cone $z = \sqrt{x^2 + y^2}$ and the paraboloid $z = 2 - x^2 - y^2$ that lies in the first octant.

SOLUTION: The solid is shown in the left panel of Figure 14.17. The intersection of the cone (bottom boundary) and paraboloid (top boundary) is a circle of unit radius. Indeed, put $r = \sqrt{x^2 + y^2}$. Then the points of intersection satisfy the condition $\sqrt{x^2 + y^2} = 2 - x^2 - y^2$ or $r = 2 - r^2$ or $r = 1$. So the projection D of the solid onto the xy plane along the z axis is the part of the disk $r \leq 1$ in the first quadrant. For any $(x, y) \in D$, the height is $h = z_{\text{top}}(x, y) - z_{\text{bot}}(x, y) = 2 - r^2 - r$ (i.e., independent of the polar angle θ). The region D is the image of the rectangle $D' = [0, 1] \times [0, \pi/2]$ in the polar plane. The volume is

$$\begin{aligned} \iint_D h(x, y) \, dA &= \iint_{D'} (2 - r^2 - r)r \, dA' \\ &= \int_0^{\pi/2} d\theta \int_0^1 (2r - r^3 - r^2) \, dr = \frac{5\pi}{24}. \end{aligned}$$

□

101.1. Study Problem.

Problem 14.4. Find the area of the four-leaved rose bounded by the polar graph $r = \cos(2\theta)$.

SOLUTION: The polar graph comes through the origin $r = 0$ four times when $\theta = \pi/4$, $\theta = \pi/4 + \pi/2$, $\theta = \pi/4 + \pi$, and $\theta = \pi/4 + 3\pi/2$. These

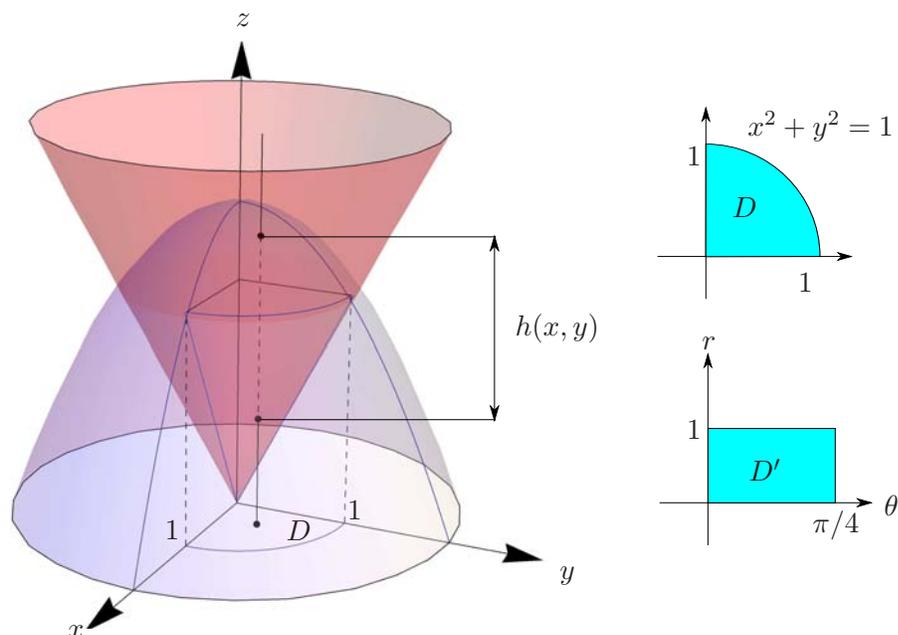


FIGURE 14.17. An illustration to Example 14.15. **Left:** The solid whose volume is sought. Its vertical projection onto the xy plane is D , which is the part of the disk $r \leq 1$ in the first quadrant. At a point (x, y) in D , the height $h(x, y)$ of the solid is the difference between the values of the z coordinate on the top and bottom boundaries (the paraboloid and the cone, respectively). **Right:** The region D' in the polar plane whose image is D .

angles may be changed by adding an integer multiple of π , owing to the periodicity of $\cos(2\theta)$. Therefore, each leaf of the rose corresponds to the range of θ between two neighboring zeros of $\cos(2\theta)$. Since all leaves have the same area, it is sufficient to find the area of one leaf, say, for $-\pi/4 \leq \theta \leq \pi/4$. With this choice, the leaf is the image of the vertically simple region

$$D' = \{(r, \theta) | 0 \leq r \leq \cos(2\theta), -\pi/4 \leq \theta \leq \pi/4\}$$

in the polar plane. Therefore, its area is given by the double integral

$$\begin{aligned} A(D) &= \iint_D dA = \iint_{D'} r \, dA' = \int_{-\pi/4}^{\pi/4} \int_0^{\cos(2\theta)} r \, dr \, d\theta \\ &= \frac{1}{2} \int_{-\pi/4}^{\pi/4} \cos^2(2\theta) \, d\theta = \frac{1}{4} \int_{-\pi/4}^{\pi/4} (1 + \cos(4\theta)) \, d\theta \\ &= \frac{1}{4} \left(\theta + \frac{1}{4} \sin(4\theta) \right) \Big|_{-\pi/4}^{\pi/4} = \frac{\pi}{8}. \end{aligned}$$

Thus, the total area is $4A(D) = \pi/2$. \square

101.2. Exercises.

(1) Sketch the region whose area is given by the iterated integral in polar coordinates and evaluate the integral:

- (i) $\int_0^\pi \int_1^2 r \, dr \, d\theta$
- (ii) $\int_{-\pi/2}^{\pi/2} \int_0^{2a \cos \theta} r \, dr \, d\theta$

(2) Convert the double integral $\iint_D f(x, y) \, dA$ to an iterated integral in polar coordinates if

- (i) D is the disk $x^2 + y^2 \leq R^2$
- (ii) D is the disk $x^2 + y^2 \leq ax$, $a > 0$
- (iii) D is the ring $a^2 \leq x^2 + y^2 \leq b^2$
- (iv) D is the parabolic segment $-a \leq x \leq a$, $x^2/a \leq y \leq a$

(3) Evaluate the double integral by changing to polar coordinates:

- (i) $\iint_D xy \, dA$, where D is the part of the ring $a^2 \leq x^2 + y^2 \leq b^2$ in the first quadrant
- (ii) $\iint_D \sin(x^2 + y^2) \, dA$, where D is the disk $x^2 + y^2 \leq a^2$
- (iii) $\iint_D \arctan(y/x) \, dA$, where D is the part of the ring $0 < a^2 \leq x^2 + y^2 \leq b^2$ between the lines $y = \sqrt{3}x$ and $y = x/\sqrt{3}$ in the first quadrant
- (iv) $\iint_D \ln(x^2 + y^2) \, dA$, where D is the portion of the ring $0 < a^2 \leq x^2 + y^2 \leq b^2$ between two half-lines $x = \pm y$, $y > 0$
- (v) $\iint_D \sin(\sqrt{x^2 + y^2}) \, dA$, where D is the ring $\pi^2 \leq x^2 + y^2 \leq 4\pi^2$

(4) If r and θ are polar coordinates, reverse the order of integration:

- (i) $\int_{-\pi/2}^{\pi/2} \int_0^{\cos \theta} f(r, \theta) \, dr \, d\theta$
- (ii) $\int_0^{\pi/2} \int_0^{a\sqrt{\sin(2\theta)}} f(r, \theta) \, dr \, d\theta$, $a > 0$
- (iii) $\int_a^0 \int_0^\theta f(r, \theta) \, dr \, d\theta$, $0 < a < 2\pi$

(5) Sketch the region of integration and evaluate the integral by converting it to polar coordinates:

- (i) $\int_{-1}^1 \int_0^{\sqrt{1-y^2}} e^{x^2+y^2} \, dx \, dy$
- (ii) $\int_{-1}^0 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} (x+y) \, dy \, dx$
- (iii) $\int_0^2 \int_0^{\sqrt{2y-y^2}} \sqrt{x^2+y^2} \, dx \, dy$
- (iv) $\int_{1/\sqrt{2}}^1 \int_{\sqrt{1-x^2}}^x xy \, dy \, dx + \int_1^{\sqrt{2}} \int_0^x xy \, dy \, dx + \int_{\sqrt{2}}^2 \int_0^{\sqrt{4-x^2}} xy \, dy \, dx$

(6) Convert the iterated integral in rectangular coordinates to an iterated integral in polar coordinates:

$$(i) \int_0^2 \int_x^{x\sqrt{3}} f(\sqrt{x^2 + y^2}) dy dx$$

$$(ii) \int_0^1 \int_0^{x^2} f(x, y) dy dx$$

(7) Convert the double integral to an iterated integral in polar coordinates:

$$(i) \iint_D f(\sqrt{x^2 + y^2}) dA, \text{ where } D \text{ is the disk } x^2 + y^2 \leq 1$$

$$(ii) \iint_D f(\sqrt{x^2 + y^2}) dA, \text{ where } D = \{(x, y) \mid |y| \leq |x|, |x| \leq 1\}$$

$$(iii) \iint_D f(y/x) dA, \text{ where } D \text{ is the disk } x^2 + y^2 \leq x$$

$$(iv) \iint_D f(\sqrt{x^2 + y^2}) dA \text{ where } D \text{ is bounded by the curve } (x^2 + y^2)^2 = a^2(x^2 - y^2)$$

(8) Find the area of the specified region D :

$$(i) D \text{ is enclosed by the polar graph } r = 1 + \cos \theta.$$

$$(ii) D \text{ is the bounded plane region between two spirals } r = \theta/4 \text{ and } r = \theta/2, \text{ where } \theta \in [0, 2\pi], \text{ and the positive } x \text{ axis.}$$

$$(iii) D \text{ is the part of the region enclosed by the cardioid } r = 1 + \sin \theta \text{ that lies outside the disk } x^2 + y^2 \leq 9/4$$

$$(iv) D \text{ is bounded by the curve } (x^2 + y^2)^2 = 2a^2(x^2 - y^2) \text{ and } x^2 + y^2 \geq a^2$$

$$(v) D \text{ is bounded by the curve } (x^3 + y^3)^2 = x^2 + y^2 \text{ and lies in the first quadrant}$$

$$(vi) D \text{ is bounded by the curve } (x^2 + y^2)^2 = a(x^3 - 3xy^2), a > 0$$

$$(vii) D \text{ is bounded by the curve } (x^2 + y^2)^2 = 8a^2xy \text{ and } (x - a)^2 + (y - a)^2 \leq a^2, a > 0.$$

(9) Find the volume of the specified solid E :

$$(i) E \text{ is bounded by the cones } z = 3\sqrt{x^2 + y^2} \text{ and } z = 4 - \sqrt{x^2 + y^2}.$$

$$(ii) E \text{ is bounded by the cone } z = \sqrt{x^2 + y^2}, \text{ the plane } z = 0, \text{ and the cylinders } x^2 + y^2 = 1, x^2 + y^2 = 4.$$

$$(iii) E \text{ is bounded by the paraboloid } z = 1 - x^2 - y^2 \text{ and the plane } z = -3.$$

$$(iv) E \text{ is bounded by the hyperboloid } x^2 + y^2 - z^2 = -1 \text{ and the plane } z = 2.$$

$$(v) E \text{ lies under the paraboloid } z = x^2 + y^2, \text{ above the } xy \text{ plane, and inside the cylinder } x^2 + y^2 = 2x.$$

(10) Find

$$\lim_{a \rightarrow 0} \frac{1}{\pi a^2} \iint_D f(x, y) dA, \quad D: x^2 + y^2 \leq a^2$$

if f is a continuous function.

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 variables 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-after 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 14.11. (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 $u(x, y) = u_i$, $i = 1, 2, \dots, N_1$, and $v(x, y) = v_j$, $j = 1, 2, \dots, N_2$, 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

$$\iint_D f \, dA = \lim_{N_1, N_2 \rightarrow \infty} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} f(\mathbf{r}_{ij}^*) \Delta A_{ij},$$

where ΔA_{ij} is the area of the partition element D_{ij} . The limit $N_1, N_2 \rightarrow \infty$ is understood as the limit of a double sequence (Definition 14.4) or as the two-variable limit $(\Delta u, \Delta v) \rightarrow (0, 0)$. 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

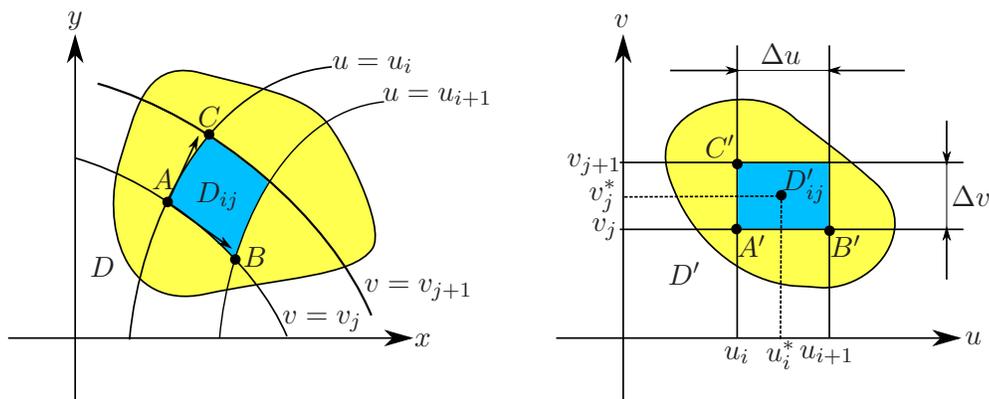


FIGURE 14.18. **Left:** A partition of a region D by the coordinate curves of the new variables $u(x, y) = u_i$ and $v(x, y) = v_j$, which are the images of the straight lines $u = u_i$ and $v = v_j$ in the uv plane. A partition element D_{ij} is bounded by the coordinate curves for which $u_{i+1} - u_i = \Delta u$ and $v_{j+1} - v_j = \Delta v$. **Right:** The region D' , whose image is the integration region D under the coordinate transformation, is partitioned by the coordinate lines $u = u_i$ and $v = v_j$. A partition element is the rectangle D'_{ij} whose area is $\Delta u \Delta v$. The change of variables establishes a one-to-one correspondence between points of D and D' . In particular, A , B , and C in D correspond to A' , B' , and C' in D' , respectively.

can be accomplished by finding a relation between ΔA_{ij} and $\Delta A'_{ij}$, that is, the rule of the area element transformation under a change of variables. Consider a rectangle D'_{ij} in the uv plane bounded by the lines $u = u_i$, $u = u_i + \Delta u$, $v = v_j$, and $v = v_j + \Delta v$. Let A' be the vertex (u_i, v_j) , B' be $(u_i + \Delta u, v_j)$, and C' be $(u_i, v_j + \Delta v)$. The image D_{ij} of D'_{ij} in the xy plane is a region bounded by the coordinate curves of the variables u and v as shown in Figure 14.18. The images A and B of the points A' and B' lie on the coordinate curve $v = v_j$, while A and C (the image of C') are on the coordinate curve $u = u_i$. Since the transformation $(u, v) \rightarrow (x, y)$ is a change of variables (see Definition 14.11), the functions $x(u, v)$ and $y(u, v)$ have continuous partial derivatives and hence are differentiable. In a small neighborhood, a differentiable function can be well approximated by its linearization (recall Definition 13.17). So, when calculating the area ΔA of D_{ij} , it is sufficient to consider variations of x and y within D_{ij} linear in variations of u and v within D'_{ij} . Consequently, the numbers Δu and Δv can be viewed as the differentials of u and v . In the limit

$(\Delta u, \Delta v) \rightarrow (0, 0)$, their higher powers can be neglected, and the area transformation law should have the form

$$\Delta A = J \Delta u \Delta v = J \Delta A',$$

where the coefficient J is to be found. Recall that $J = r$ for polar coordinates.

Suppose that the gradients $\nabla x(u, v)$ and $\nabla y(u, v)$ do not vanish. Then the level curves of $x(u, v)$ and $y(u, v)$, which are also the coordinate curves, are smooth (recall the discussion of Theorem 13.16). A sufficiently small part of a smooth curve between two points can be well approximated by a secant line through these points (recall Section 80.3). This argument suggests that the area of D_{ij} can be approximated by the area of a parallelogram with adjacent sides $\overrightarrow{AB} = \mathbf{b}$ and $\overrightarrow{AC} = \mathbf{c}$. The coordinates of A are $(x(u_i, v_j), y(u_i, v_j))$, while the coordinates of B are $(x(u_i + \Delta u, v_j), y(u_i + \Delta u, v_j))$ because they are images of A' and B' , respectively, under the inverse transformation $x = x(u, v)$ and $y = y(u, v)$. Therefore,

$$\begin{aligned} \mathbf{b} &= \left(x(u_i + \Delta u, v_j) - x(u_i, v_j), y(u_i + \Delta u, v_j) - y(u_i, v_j), 0 \right) \\ &= \left(x'_u(u_i, v_j) \Delta u, y'_u(u_i, v_j) \Delta u, 0 \right) \\ &= \Delta u \left(x'_u(u_i, v_j), y'_u(u_i, v_j), 0 \right), \end{aligned}$$

where, owing to the smallness of Δu and Δv , the variations of x and y have been linearized: $x(u_i + \Delta u, v_j) - x(u_i, v_j) = x'_u(u_i, v_j) \Delta u$ and $y(u_i + \Delta u, v_j) - y(u_i, v_j) = y'_u(u_i, v_j) \Delta u$; that is, higher powers of Δu have been neglected. 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

$$\mathbf{c} = \Delta v \left(x'_v(u_i, v_j), y'_v(u_i, v_j), 0 \right).$$

The area of the parallelogram reads

$$(14.10) \quad \Delta A_{ij} = \|\mathbf{b} \times \mathbf{c}\| = \left| \det \begin{pmatrix} x'_u & x'_v \\ y'_u & y'_v \end{pmatrix} \right| \Delta u \Delta v = J(u_i, v_j) \Delta u \Delta v,$$

where the partial derivatives are taken at the point (u_i, v_i) . 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 14.12. (Jacobian of a Transformation).

The Jacobian of a transformation defined by differentiable functions $x = x(u, v)$ 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 in 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 $\Delta A \neq 0$. If the partial derivatives of x and y with respect to u and v are continuous on D' , J is continuous on D' , too. Therefore, for any sample point (u_i^*, v_j^*) in D'_{ij} , the difference $(\Delta A_{ij} - J(u_i^*, v_j^*)\Delta A')/\Delta A' = J(u_i, v_j) - J(u_i^*, v_j^*)$ vanishes in the limit $(\Delta u, \Delta v) \rightarrow (0, 0)$. So, if in (14.10) the value of the Jacobian is taken at any sample point, then the corresponding change in the value of ΔA_{ij} depends on higher powers of Δu and Δv . Since ΔA_{ij} has been calculated in the linear approximation, such variations of ΔA_{ij} may be neglected and one can always put

$$\Delta A_{ij} = J(u_i^*, v_j^*) \Delta u \Delta v$$

in the Riemann sum for any choice of sample points. The limit of the Riemann sum

$$\iint_D f dA = \lim_{N_1, N_2 \rightarrow \infty} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} f(x(u_i^*, v_j^*), y(u_i^*, v_j^*)) \Delta A_{ij}$$

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 14.9. (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

$$\iint_D f(x, y) dA = \iint_{D'} f(x(u, v), y(u, v)) J(u, v) dA',$$

$$J(u, v) = \left| \frac{\partial(x, y)}{\partial(u, v)} \right|.$$

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 as well as a rigorous justification of the linear approximation in the area transformation law, which were excluded in the above analysis.

The change of variables in a double integral entails the following steps:

1. Finding the region D' whose image under the transformation $x = x(u, v)$, $y = y(u, v)$ is the integration region D . A useful rule to remember here is

$$\text{boundaries of } D' \longrightarrow \text{boundaries of } D$$

under the transformation. In particular, 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:

$$dA = \left| \frac{\partial(x, y)}{\partial(u, v)} \right| du dv = J dA', \quad J = \left| \frac{\partial(x, y)}{\partial(u, v)} \right|.$$

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).

EXAMPLE 14.16. Use the change of variables $x = u(1 - v)$, $y = uv$ to evaluate the integral $\iint_D (x+y)^5 y^5 dA$, where D is the triangle bounded by the lines $y = 0$, $x = 0$, and $x + y = 1$.

SOLUTION:

1. Note first that the line $u = 0$ is mapped to a single point, the origin, in the xy plane. So the line $u = 0$ must be a boundary of D' . The equation $x = 0$ in the new variables becomes $u(1 - v) = 0$, which means that either $u = 0$ or $v = 1$. Therefore, the line $v = 1$ is a boundary of D' as it is mapped to the boundary line $x = 0$. The equation $y = 0$ in the new variables reads $uv = 0$. Therefore, the line $v = 0$ is also a boundary of D' . The equation $x + y = 1$ in the new variables has the form $u = 1$. Thus, the region D' is bounded by four lines $u = 0$, $u = 1$, $v = 0$, and $v = 1$, which is the square $[0, 1] \times [0, 1]$.
2. Since $x + y = u$, the integrand in the new variables is $u^5(uv)^5 = u^{10}v^5$.
3. The Jacobian of the transformation is

$$\frac{\partial(x, y)}{\partial(u, v)} = \det \begin{pmatrix} x'_u & x'_v \\ y'_u & y'_v \end{pmatrix} = \det \begin{pmatrix} 1 - v & -u \\ v & u \end{pmatrix} = u(1 - v) + uv = u.$$

Therefore the area element transformation is $dA = |u|dA'$. The absolute value may be omitted because $u \geq 0$ in D' . Note that the Jacobian vanishes only on the boundary of D' and, hence, the hypotheses of Theorem 14.9 are fulfilled.

4. The double integral in the new variables is evaluated by Fubini's theorem:

$$\iint_D (x + y)^5 y^5 dA = \iint_{D'} u^{11} v^5 dA' = \int_0^1 u^{11} du \int_0^1 v^5 dv = \frac{1}{12} \cdot \frac{1}{6} = \frac{1}{72}.$$

□

This example and the example of polar coordinate show that the transformation is not one-to-one on the sets where the Jacobian vanishes (the line $u = 0$ is mapped to a single point) and the inverse transformation fails to exist. It turns out that this observation is of a general nature.

THEOREM 14.10. (Inverse Function Theorem).

Let the transformation $(u, v) \rightarrow (x, y)$ be defined on an open set U' containing a point (u_0, v_0) . Suppose that the functions $x(u, v)$ and $y(u, v)$ have continuous partial derivatives in U' and the Jacobian of the transformation does not vanish at the point (u_0, v_0) . Then there exists an inverse transformation $u = u(x, y)$, $v = v(x, y)$ in an open set U containing the image point $(x_0, y_0) = (x(u_0, v_0), y(u_0, v_0))$ and

the functions $u(x, y)$ and $v(x, y)$ have continuous partial derivatives in U .

By this theorem, the Jacobian of the *inverse transformation* can be calculated as $\partial(u, v)/\partial(x, y)$ so that the area transformation law is $du dv = |\partial(u, v)/\partial(x, y)| dx dy$ and the following statement holds.

COROLLARY 14.3. *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*

$$(14.11) \quad \frac{\partial(x, y)}{\partial(u, v)} = \frac{1}{\frac{\partial(u, v)}{\partial(x, y)}} = \frac{1}{\det \begin{pmatrix} u'_x & u'_y \\ v'_x & v'_y \end{pmatrix}}.$$

The analogy with a change of variables in the one-dimensional case can be made. If $x = f(u)$, where f has continuous derivative $f'(u)$ that does not vanish, then, by the inverse function theorem for functions of one variable (Theorem 12.6), there is an inverse function $u = g(x)$ whose derivative is continuous and $g'(x) = 1/f'(u)$, where $u = g(x)$. Then the transformation of the differential dx can be written in two equivalent forms, just like the transformation of the area element $dA = dx dy$:

$$dx = f'(u) du = \frac{du}{g'(x)} \quad \longleftrightarrow \quad dx dy = \left| \frac{\partial(x, y)}{\partial(u, v)} \right| du dv = \frac{du dv}{\left| \frac{\partial(u, v)}{\partial(x, y)} \right|}.$$

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 14.17. *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 equations of the lines can be written in the form $y/x = 1$ and $y/x = 3$ because $y, x > 0$ in D (see Figure 14.19). 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 $v = v(x, y) = xy$, 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

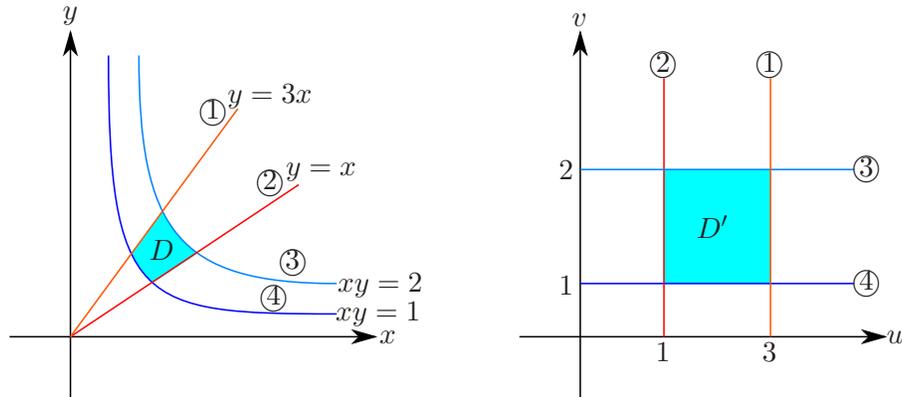


FIGURE 14.19. An illustration to Example 14.17. The transformation of the integration region D . Equations of the boundaries of D , $y = 3x$, $y = x$, $xy = 2$, and $xy = 1$, are written in the new variables $u = y/x$ and $v = xy$ to obtain the equations of the boundaries of D' , $u = 3$, $u = 1$, $v = 2$, and $v = 1$, respectively. The correspondence between the boundaries of D and D' is indicated by encircled numbers enumerating the boundary curves.

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):

$$J = \left| \det \begin{pmatrix} u'_x & u'_y \\ v'_x & v'_y \end{pmatrix} \right|^{-1} = \left| \det \begin{pmatrix} -y/x^2 & 1/x \\ y & x \end{pmatrix} \right|^{-1} = \left| -\frac{2y}{x} \right|^{-1} = \frac{x}{2y}.$$

The absolute value bars may be omitted as x and y are strictly positive in D . The integrand becomes $fJ = x^2y^2/2 = v^2/2$. So finding the functions $x = x(u, v)$ and $y = y(u, v)$ happens to be unnecessary in this example! Hence,

$$\iint_D xy^3 dA = \frac{1}{2} \iint_{D'} v^2 dA' = \frac{1}{2} \int_1^3 du \int_1^2 v^2 dv = \frac{7}{3}.$$

The reader is advised to evaluate the double integral in the original rectangular coordinates to compare the amount of work needed with this solution. \square

The following example illustrates how a change of variables can be used to simplify the integrand of a double integral.

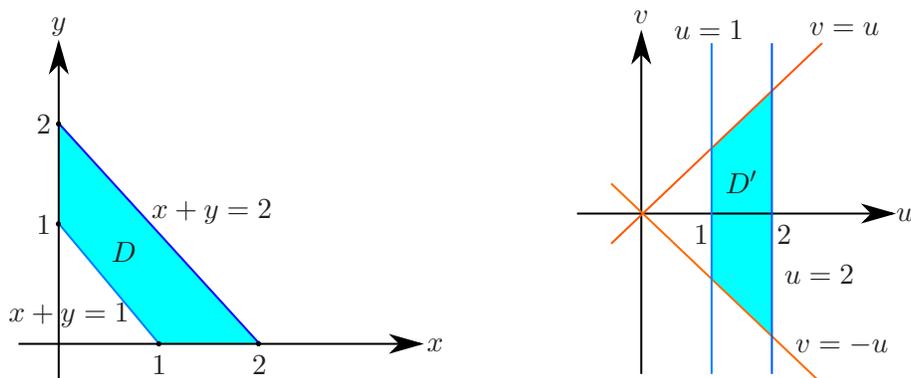


FIGURE 14.20. **Left:** The integration region D in Example 14.18 is bounded by the lines $x + y = 1$, $x + y = 2$, $x = 0$, and $y = 0$. **Right:** The image D' of D under the change of variables $u = x + y$ and $v = y - x$. The boundaries of D' are obtained by substituting the new variables into the equations for boundaries of D so that $x + y = 1 \rightarrow u = 1$, $x + y = 2 \rightarrow u = 2$, $x = 0 \rightarrow v = u$, and $y = 0 \rightarrow v = -u$.

EXAMPLE 14.18. 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 (with respect to either 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 $D' = \{(u, v) \mid -u \leq v \leq u, u \in [1, 2]\}$. The Jacobian of the change of variables is $J = 1/2$. Hence,

$$\begin{aligned} \iint_D \cos\left(\frac{y-x}{y+x}\right) dA &= \frac{1}{2} \iint_{D'} \cos\left(\frac{v}{u}\right) dA' = \frac{1}{2} \int_1^2 \int_{-u}^u \cos\left(\frac{v}{u}\right) dv du \\ &= \frac{1}{2} \int_1^2 u \sin\left(\frac{v}{u}\right) \Big|_{-u}^u du \\ &= \sin(1) \int_1^2 u du = 3 \sin(1)/2. \end{aligned}$$

□

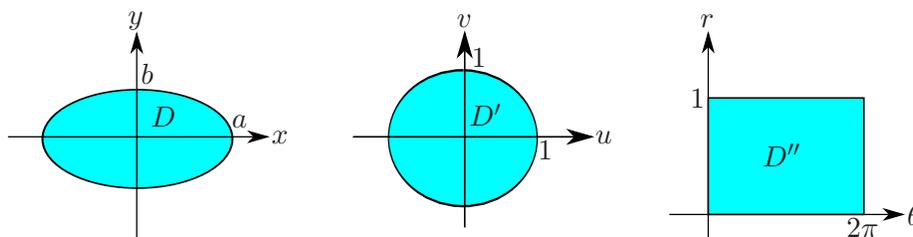


FIGURE 14.21. The transformation of the integration region D in Example 14.19. The region D , $x^2/a^2 + y^2/b^2 \leq 1$, is first transformed into the disk D' , $u^2 + v^2 \leq 1$, by $x = au$, $y = bv$, and then D' is transformed into the rectangle D'' by $u = r \cos \theta$, $v = r \sin \theta$.

EXAMPLE 14.19. (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 $J = ab$,

$$A(D) = \iint_D dA = \iint_{D'} J dA' = ab \iint_{D'} dA' = abA(D') = \pi ab.$$

Of course, the area $A(D')$ of the disk $u^2 + v^2 \leq 1$ can also be evaluated by converting the integral over D' to polar coordinates $u = r \cos \theta$, $v = r \sin \theta$. The disk D' is the image of the rectangle $D'' = [0, 1] \times [0, 2\pi]$, and the Jacobian is r . The transformations of the integration region are shown in Figure 14.21. \square

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. In Section 100.4, the symmetry properties of double integrals are shown to be quite helpful for their evaluation. Using the concept of a change of variables in double integrals, one can give an algebraic criterion of the symmetry transformation of a region D that preserves its area. 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, $dA = dA'$. Indeed, changing variables in the double integral for the area, $A(D) = \iint_D dA = \iint_{D'} dA' = A(D')$. For example, rotations, translations, and reflections are area-preserving transformations for obvious geometrical reasons. The following theorem holds.

THEOREM 14.11. *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 $dA = dA'$, 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 $I = 0$. □

102.4. Study Problem.

Problem 14.5. (Generalized Polar Coordinates).

Generalized polar coordinates are defined by the transformation

$$x = ar \cos^n \theta, \quad y = br \sin^n \theta,$$

where a , b , and n are parameters. Find the Jacobian of the transformation. Use the generalized polar coordinates with a suitable choice of parameters to find the area of the region in the first octant that is bounded by the curve $\sqrt[4]{x/a} + \sqrt[4]{y/b} = 1$.

SOLUTION: The Jacobian of the generalized polar coordinates is

$$\begin{aligned} \frac{\partial(x, y)}{\partial(r, \theta)} &= \det \begin{pmatrix} x'_r & x'_\theta \\ y'_r & y'_\theta \end{pmatrix} = \det \begin{pmatrix} a \cos^n \theta & -nar \sin \theta \cos^{n-1} \theta \\ b \sin^n \theta & nbr \cos \theta \sin^{n-1} \theta \end{pmatrix} \\ &= nabr (\cos^{n+1} \theta \sin^{n-1} \theta + \cos^{n-1} \theta \sin^{n+1} \theta) \\ &= nabr \cos^{n-1} \theta \sin^{n-1} \theta (\cos^2 \theta + \sin^2 \theta) \\ &= nabr \cos^{n-1} \theta \sin^{n-1} \theta. \end{aligned}$$

Choosing the parameter $n = 8$, the equation of the curve $\sqrt[4]{x/a} + \sqrt[4]{y/b} = 1$ becomes $\sqrt[4]{r} = 1$ or $r = 1$. Since the region in question lies in the first quadrant, it is also bounded by the lines $y = 0$ and $x = 0$, which are the images of the lines $\theta = \pi/2$ and $\theta = 0$ in the (r, θ) plane. Therefore, the rectangle $D' = [0, 1] \times [0, \pi/2]$ is mapped onto the region D in question. The Jacobian of the transformation is *positive* in D' so the absolute value of the Jacobian in the area element transformation

may be omitted. The area of D is

$$\begin{aligned} A(D) &= \iint_D dA = \iint_{D'} J dA' = 8ab \int_0^{\pi/2} \cos^7 \theta \sin^7 \theta d\theta \int_0^1 r dr \\ &= \frac{ab}{32} \int_0^{\pi/2} (\sin(2\theta))^7 d\theta = -\frac{ab}{64} \int_0^{\pi/2} (\sin(2\theta))^6 d\cos(2\theta) \\ &= \frac{ab}{64} \int_{-1}^1 (1-u^2)^3 du = \frac{ab}{70}, \end{aligned}$$

where first the double-angle formula $\cos \theta \sin \theta = \frac{1}{2} \sin(2\theta)$ has been used and then the integration has been carried out with the help of the substitution $u = \cos(2\theta)$. \square

102.5. Exercises.

(1) Find the Jacobian of the following transformations:

- (i) $x = 3u - 2v, y = u + 3v$
- (ii) $x = e^r \cos \theta, y = e^r \sin \theta$
- (iii) $x = uv, y = u^2 - v^2$
- (iv) $x = u \cosh v, y = u \sinh v$

(2) Consider *hyperbolic coordinates* in the first quadrant $x > 0, y > 0$ defined by the transformation $x = ve^u, y = ve^{-u}$. Calculate the Jacobian. Determine the range of (u, v) in which the transformation is one-to-one. Find the inverse transformation and sketch coordinate curves of hyperbolic coordinates.

(3) Find the conditions on the parameters of a linear transformation $x = a_1u + b_1v + c_1, y = a_2u + b_2v + c_2$ so that the transformation is area preserving. In particular, prove that the rotations discussed in Study Problem 11.2 are area preserving.

(4) Find the image D of the specified region D' under the given transformation:

- (i) $D' = [0, 1] \times [0, 1]$ and the transformation is $x = u, y = v(1 - u^2)$.
- (ii) D' is the triangle with vertices $(0, 0), (1, 0),$ and $(1, 1)$, and the transformation is $x = v^2, y = u$.
- (iii) D' is the region defined by the inequality $|u| + |v| \leq 1$, and the transformation is $x = u + v, y = u - v$.

(5) Find a linear transformation that maps the triangle D' with vertices $(0, 0), (0, 1),$ and $(1, 0)$ onto the triangle D with vertices $(0, 0), (a, b),$ and (b, a) , where a and b are positive, nonequal numbers. Use this transformation to evaluate the integral of $f(x, y) = bx - ay$ over the

triangle D .

(6) Evaluate the double integral using the specified change of variables:

- (i) $\iint_D (8x + 4y) dA$, where D is the parallelogram with vertices $(3, -1)$, $(-3, 1)$, $(-1, 3)$, and $(5, 1)$; the change of variables is $x = (v - 3u)/4$, $y = (u + v)/4$
- (ii) $\iint_D (x^2 - xy + y^2) dA$, where D is the region bounded by the ellipse $x^2 - xy + y^2 = 1$; the change of variables is $x = u - v/\sqrt{3}$, $y = u + v/\sqrt{3}$
- (iii) $\iint_D (x^2 - y^2)^{-1/2} dA$, where D is in the first quadrant and bounded by hyperbolas $x^2 - y^2 = 1$, $x^2 - y^2 = 4$ and by the lines $x = 2y$, $x = 4y$; the change of variables is $x = u \cosh v$, $y = u \sinh v$
- (iv) $\int \int_D e^{(x/y)} (x + y)^3 / y^2 dA$, where D is bounded by the lines $y = x$, $y = 2x$, $x + y = 1$, and $x + y = 2$; the change of variables $u = x/y$, $v = x + y$ *Hint:* Follow the procedure based on (14.11) as illustrated in Example 14.17.

(7) Find the image D' of the square $a < x < a + h$, $b < y < b + h$, where a , b , and h are positive numbers, under the transformation $u = y^2/x$, $v = \sqrt{xy}$. Find the ratio of the area $A(D')$ to the area $A(D)$. What is the limit of the ratio when $h \rightarrow 0$?

(8) Use the specified change of variables to convert the iterated integral to an iterated integral in the new variables:

- (i) $\int_a^b \int_{\alpha x}^{\beta x} f(x, y) dy dx$, where $0 < a < b$ and $0 < \alpha < \beta$ if $u = x$ and $v = y/x$
- (ii) $\int_0^2 \int_{1-x}^{2-x} f(x, y) dy dx$ if $u = x + y$ and $v = x - y$

(9) Convert the double integral $\iint_D f(x, y) dA$ to an iterated integral in the new variables, where D is bounded by the curve $\sqrt{x} + \sqrt{y} = \sqrt{a}$, $a > 0$, and the lines $x = 0$, $y = 0$ if $x = u \cos^4 v$ and $y = u \sin^4 v$.

(10) Evaluate the double integral by making a suitable change of variables:

- (i) $\iint_D yx^2 dA$, where D is in the first quadrant and bounded by the curves $xy = 1$, $xy = 2$, $yx^2 = 1$, and $yx^2 = 2$
- (ii) $\iint_D e^{x-y} dA$, where D is given by the inequality $|x| + |y| \leq 1$
- (iii) $\iint_D (1 + 3x^2) dA$, where D is bounded by the lines $x + y = 1$, $x + y = 2$ and by the curves $y - x^3 = 0$, $y - x^3 = 1$
- (iv) $\iint_D (y + 2x^2) dA$, where the domain D is bounded by two parabolas, $y = x^2$, $y = x^2 + 2$ and by two hyperbolas $xy = -1$ ($x < 0$), $xy = 1$ ($x > 0$)

- (v) $\iint_D (x+y)/x^2 dA$, where D is bounded by four lines $y = x$, $y = 2x$, $y + x = 1$, and $y + x = 2$
- (vi) $\iint_D \sqrt{y-x}/(x+y)$, where D is the square with vertices $(0, 2a)$, (a, a) , $(2a, 2a)$, and $(a, 3a)$ with $a > 0$
- (vii) $\iint_D \cos(x^2/a^2 + y^2/b^2) dA$, where D is bounded by the ellipse $x^2/a^2 + y^2/b^2 = 1$
- (viii) $\iint_D (x+y) dA$, where D is bounded by $x^2 + y^2 = x + y$
- (ix) $\iint_D (|x| + |y|) dA$, where D is defined by $|x| + |y| \leq 1$
- (x) $\iint_D (1 - \frac{x^2}{a^2} - \frac{y^2}{b^2})^{-1/2} dA$, where D is bounded by the ellipse $x^2/a^2 + y^2/b^2 = 1$
- (11)** Let f be continuous on $[0, 1]$. Show that $\iint_D f(x+y) dA = \int_0^1 uf(u) du$ if D is the triangle with vertices $(0, 0)$, $(0, 1)$, and $(1, 0)$.
- (12)** Use a suitable change of variables to reduce the double integral to a single integral:
- (i) $\iint_D f(x+y) dA$, where D is defined by $|x| + |y| \leq 1$
- (ii) $\iint_D f(ax + by + c) dA$, where D is the disk $x^2 + y^2 \leq 1$ and $a^2 + b^2 \neq 0$
- (iii) $\iint_D f(xy) dA$, where D lies in the first quadrant and is bounded by the curves $xy = 1$, $xy = 2$, $y = x$, and $y = 4x$
- (13)** Let n and m be positive integers. Prove that if $\iint_D x^n y^m dA = 0$, where D is bounded by an ellipse $x^2/a^2 + y^2/b^2 = 1$, then at least one of the numbers n and m is odd.
- (14)** Suppose that the level curves of a function $f(x, y)$ are simple closed curves and the region D is bounded by two level curves $f(x, y) = a$ and $f(x, y) = b$. Prove that

$$\iint_D f(x, y) dA = \int_a^b uF'(u) du,$$

where $F(u)$ is the area of the region between the curves $f(x, y) = a$ and $f(x, y) = u$. *Hint:* Split the region D by infinitesimally close level curves of the function f .

(15) Use the generalized polar coordinates with a suitable choice of parameters to find the area of a region D if

- (i) D is bounded by the curves $x^2/a^2 + y^3/b^3 = x^2 + y^2$ and lies in the first quadrant.
- (ii) D is bounded by the curves $x^3/a^3 + y^3/b^3 = x^2/c^2 - y^2/k^2$ and lies in the first quadrant.
- (iii) D is bounded by the curve $(x/a + y/b)^5 = x^2 y^2 / c^4$.

(16) Use the double integral and a suitable change of variables to find the area of D if

- (i) D is bounded by the curves $x + y = a$, $x + y = b$, $y = mx$, and $y = nx$ and lies in the first quadrant
- (ii) D is bounded by the curves $y^2 = 2ax$, $y^2 = 2bx$, $x^2 = 2cy$, and $x^2 = 2ky$, where $0 < a < b$ and $0 < c < k$
- (iii) D is bounded by the curves $(x/a)^{1/2} + (y/b)^{1/2} = 1$, $(x/a)^{1/2} + (y/b)^{1/2} = 2$, $x/a = y/b$, and $4x/a = y/b$, where $a > 0$ and $b > 0$
- (iv) D is bounded by the curves $(x/a)^{2/3} + (y/b)^{2/3} = 1$, $(x/a)^{2/3} + (y/b)^{2/3} = 4$, $x/a = y/b$, and $8x/a = y/b$ and lies in the first quadrant
- (v) D is bounded by the ellipses $x^2/\cosh^2 u + y^2/\sinh^2 u = 1$, where $u = u_1$ and $u = u_2 > u_1$, and by the hyperbolas $x^2/\cos^2 v - y^2/\sin^2 v = 1$, where $v = v_1$ and $v = v_2 > v_1$. *Hint:* Consider the transformation $x = \cosh u \cos v$, $y = \sinh u \sin v$.

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

$$\sigma(\mathbf{r}) = \lim_{\Delta V \rightarrow 0} \frac{\Delta m(\mathbf{r})}{\Delta V}.$$

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 $R \rightarrow 0$ (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 $\sigma(\mathbf{r}) = 5 \text{ g/cm}^3$ means that a piece of material of volume 1 cm^3 cut out around the point \mathbf{r} has a mass of 5 g.

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 , $p = 1, 2, \dots, N$, 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 $\mathbf{r}_p \in E_p$ (see the left panel of Figure 14.22). 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

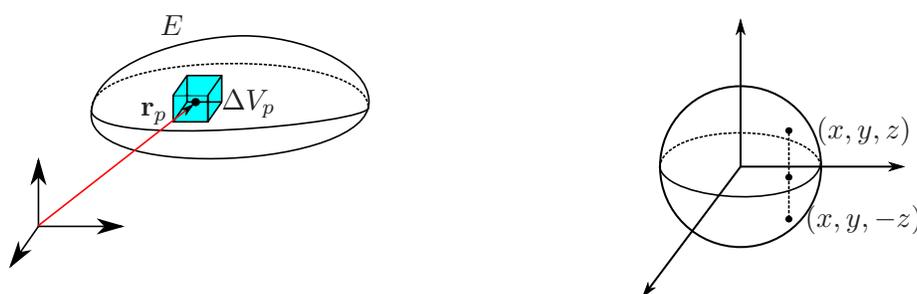


FIGURE 14.22. **Left:** A partition element of a solid region, where \mathbf{r}_p is the position vector of a sample point in it. If $\sigma(\mathbf{r})$ is the mass density, then the mass of the partition element is $\Delta m(\mathbf{r}_p) \approx \sigma(\mathbf{r}_p) \Delta V_p$, where ΔV_p is the volume of the partition element. The total mass is the sum of $\Delta m(\mathbf{r}_p)$ over the partition of the solid E as given in (14.12). **Right:** An illustration to Example 14.17. A ball is symmetric under the reflection about the xy plane: $(x, y, z) \rightarrow (x, y, -z)$. If the function f is skew-symmetric under this reflection, $f(x, y, -z) = -f(x, y, z)$, then the triple integral of f over the ball vanishes.

partition elements so that $R_p \leq R_N \rightarrow 0$ as $N \rightarrow \infty$, 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 $\Delta V_p \rightarrow 0$ for all p . So the total mass is

$$(14.12) \quad M = \lim_{\substack{N \rightarrow \infty \\ (R_N \rightarrow 0)}} \sum_{p=1}^N \sigma(\mathbf{r}_p) \Delta V_p,$$

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.

Smooth Surface. In Section 85.5, a surface was defined as a continuous deformation of an open set in a plane that has continuous inverse. A small piece of a surface can be viewed as the graph of a continuous function of two variables. Similarly to the notion of a smooth curve, a smooth surface can be defined. If the graph has a tangent plane at every point and the normal to the tangent plane changes continuously along the graph, then the surface is called *smooth*. Consider a level set

of a function $g(\mathbf{r})$ of three variables $\mathbf{r} = (x, y, z)$. Suppose that g has continuous partial derivatives and the gradient ∇g does not vanish. As explained in Section 93.2 (see the discussion of Theorem 13.16), a level set $g(\mathbf{r}) = k$ is a surface whose normal vector is the gradient ∇g . If g has continuous partial derivatives, then the components of the normal are continuous. Thus, *a surface is said to be smooth in a neighborhood of a point \mathbf{r}_0 if it coincides with a level set $g(\mathbf{r}) = g(\mathbf{r}_0)$ of a function g that has continuous partial derivatives and whose gradient does not vanish in at \mathbf{r}_0 . A surface is smooth if it is smooth in a neighborhood of its every point. A surface is piecewise smooth if it consists of several smooth pieces adjacent along smooth curves.*

Rectangular Partition. A region E in space is assumed to be closed and bounded; that is, it is contained in a ball of some (finite) radius. The boundaries of E are assumed to be piecewise-smooth surfaces. The region E is then embedded in a rectangular box $R_E = [a, b] \times [c, d] \times [s, q]$, 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 $x = x_i = a + i \Delta x$, $i = 0, 1, \dots, N_1$, where $\Delta x = (b - a)/N_1$; $y = y_j = c + j \Delta y$, $j = 0, 1, \dots, N_2$, where $\Delta y = (d - c)/N_2$; and $z = z_k = s + k \Delta z$, $k = 0, 1, \dots, N_3$, 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 $\mathbf{N} = (N_1, N_2, N_3)$. So the upper and lower sums are *triple* sequences (a rule that assigns a number a_{nmk} to an ordered triple of integers (n, m, k) is a triple sequence). The limit of a triple sequence is defined similarly to the limit of a double sequence (a_{nm} is replaced by a_{nmk} in Definition 14.4).

DEFINITION 14.13. (Triple Integral).

If the limits of the upper and lower sums exist as $N_{1,2,3} \rightarrow \infty$ (or $(\Delta x, \Delta y, \Delta z) \rightarrow (0, 0, 0)$) and coincide, then f is said to be Riemann

integrable on E , and the limit of the upper and lower sums

$$\iiint_E f(x, y, z) dV = \lim_{\mathbf{N} \rightarrow \infty} U(f, \mathbf{N}) = \lim_{\mathbf{N} \rightarrow \infty} L(f, \mathbf{N})$$

is called the triple integral of f over the region E .

The limit is understood as a three-variable limit $(\Delta x, \Delta y, \Delta z) \rightarrow (0, 0, 0)$ or as the limit of a triple sequence.

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 hold 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 14.12. (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

$$V(E) = \iiint_E dV.$$

If $m \leq f(\mathbf{r}) \leq M$ for all \mathbf{r} in E , then

$$mV(E) \leq \iiint_E f dV \leq MV(E).$$

The Integral Mean Value Theorem. The integral mean value theorem (Theorem 14.3) is extended to triple integrals. If f is continuous in E , then there is exists a point \mathbf{r}_0 in E such that

$$\iiint_E f(\mathbf{r}) dV = V(E)f(\mathbf{r}_0).$$

Its proof follows the same line of reasoning 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:

$$(14.13) \quad \iiint_E f(\mathbf{r}) \, dV = \lim_{\substack{N \rightarrow \infty \\ (R_N \rightarrow 0)}} \sum_{p=1}^N f(\mathbf{r}_p) \Delta V_p.$$

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 14.20. 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 $g(x, y, z) = x^2 \sin(y^4 z)$ 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 $(x, y, z) \rightarrow (x, y, -z)$, 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.

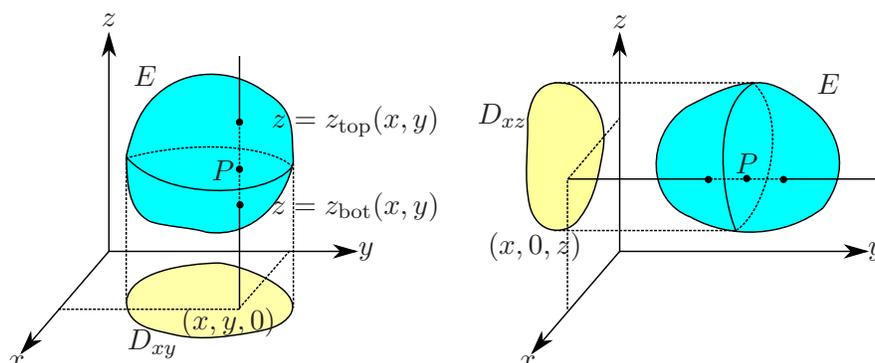


FIGURE 14.23. **Left:** An algebraic description of a solid region simple in the direction of the z axis. The solid E is vertically projected into the xy plane: every point (x, y, z) of E goes into the point $(x, y, 0)$. The projection points form the region D_{xy} . Since E is simple in the z direction, for every $(x, y, 0)$ in D_{xy} , the z coordinate of the point $P(x, y, z)$ in E ranges over the interval $z_{\text{bot}}(x, y) \leq z \leq z_{\text{top}}(x, y)$. In other words, E lies between the graphs $z = z_{\text{bot}}(x, y)$ and $z = z_{\text{top}}(x, y)$. **Right:** An illustration to the algebraic description (14.14) of a solid E as simple in the y direction. E is projected along the y axis to the xz plane, forming a region D_{xz} . For every $(x, 0, z)$ in D_{xz} , the y coordinate of the point $P(x, y, z)$ in E ranges over the interval $y_{\text{bot}}(x, z) \leq y \leq y_{\text{top}}(x, z)$. In other words, E lies between the graphs $y = y_{\text{bot}}(x, z)$ and $y = y_{\text{top}}(x, z)$.

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 14.14. (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 $\mathbf{v} = \hat{\mathbf{e}}_3$; that is, E is simple along the z axis. Then the region E admits the following description:

$$E = \{(x, y, z) \mid z_{\text{bot}}(x, y) \leq z \leq z_{\text{top}}(x, y), (x, y) \in D_{xy}\}.$$

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_{bot} and z_{top} are the minimal and maximal values of the z coordinate along the intersection segment, then, for any $(x, y, z) \in E$, $z_{\text{bot}} \leq z \leq z_{\text{top}}$ and any $(x, y) \in D_{xy}$. Naturally, the values z_{bot} and z_{top} may depend on $(x, y) \in D_{xy}$. Thus, the region E is bounded from the top by the graph $z = z_{\text{top}}(x, y)$ and from the bottom by the graph $z = z_{\text{bot}}(x, y)$. If E is simple along the y or x axis, then E admits similar descriptions:

$$(14.14) \quad E = \{(x, y, z) \mid y_{\text{bot}}(x, z) \leq y \leq y_{\text{top}}(x, z), (x, z) \in D_{xz}\},$$

$$(14.15) \quad E = \{(x, y, z) \mid x_{\text{bot}}(y, z) \leq x \leq x_{\text{top}}(y, z), (y, z) \in D_{yz}\},$$

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 (a generalization of Theorem 14.5 to the three-dimensional case is trivial as its proof is based on Theorem 14.4, which holds in any number of dimensions). Let D_p , $p = 1, 2, \dots, N$, be a partition of 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 $z = s + k \Delta z$, $k = 0, 1, \dots, N_3$, $\Delta z = (q - s)/N_3$. 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 $k = 1, 2, \dots, N_3$. The three-variable limit (14.13) exists and hence can be taken in any particular order (recall Theorem 14.6). Take first the limit $N_3 \rightarrow \infty$ or $\Delta z \rightarrow 0$. The double limit of the sum over the partition of D_{xy} is understood as before; that is, as $N \rightarrow 0$, the radii R_p of smallest disks

containing D_p go to 0 uniformly, $R_p \leq R_N \rightarrow 0$. Therefore,

$$\begin{aligned} \iiint_E f \, dV &= \lim_{\substack{N \rightarrow \infty \\ (R_N \rightarrow 0)}} \sum_{p=1}^N \left(\lim_{N_3 \rightarrow \infty} \sum_{k=1}^{N_3} f(x_p, y_p, z_k^*) \Delta z \right) \Delta A_p \\ &= \lim_{\substack{N \rightarrow \infty \\ (R_N \rightarrow 0)}} \sum_{p=1}^N \left(\int_{z_{\text{bot}}(x_p, y_p)}^{z_{\text{top}}(x_p, y_p)} f(x_p, y_p, z) \, dz \right) \Delta A_p \end{aligned}$$

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 parentheses. 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 14.13. (Iterated Triple Integral).

Let f be integrable on a solid region E bounded by a piecewise smooth surface. 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

$$\begin{aligned} \iiint_E f(x, y, z) \, dV &= \iint_{D_{xy}} \int_{z_{\text{bot}}(x, y)}^{z_{\text{top}}(x, y)} f(x, y, z) \, dz \, dA \\ &= \iint_{D_{xy}} F(x, y) \, dA. \end{aligned}$$

103.4. 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 14.14. (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). In particular, if $f(x, y, z) = g(x)h(y)w(z)$, then

$$\iiint_E f(x, y, z) \, dV = \int_a^b g(x) \, dx \int_c^d h(y) \, dy \int_s^q w(z) \, dz,$$

which is an extension of the factorization property stated in Corollary 14.2 to triple integrals.

EXAMPLE 14.21. 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,

$$\iiint_E xy^2z^3 \, dV = \int_0^2 x \, dx \int_1^2 y^2 \, dy \int_0^3 z^3 \, dz = 2 \cdot (7/3) \cdot 9 = 42.$$

□

EXAMPLE 14.22. 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 $r = 1$,

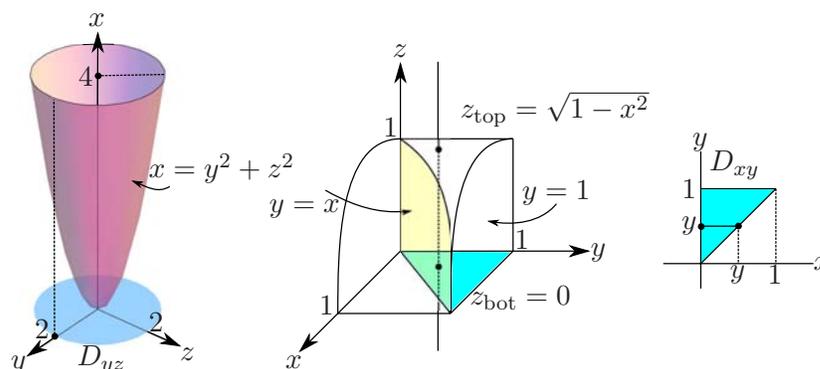


FIGURE 14.24. **Left:** The integration region in Example 14.23. The x axis is vertical. The region is bounded by the plane $x = 4$ (top) and the paraboloid $x = y^2 + z^2$ (bottom). Its projection into the yz plane is the disk of radius 2 as the plane and paraboloid intersect along the circle $4 = y^2 + z^2$. **Right:** An illustration to Study Problem 14.6.

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 been 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 substitution $u = r^2$. \square

EXAMPLE 14.23. 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 (see the left panel of Figure 14.24). One has

$$\begin{aligned} \iiint_E \sqrt{y^2 + z^2} dV &= \iint_{D_{yz}} \sqrt{y^2 + z^2} \int_{y^2+z^2}^4 dx dA \\ &= \iint_{D_{yz}} \sqrt{y^2 + z^2} [4 - (y^2 + z^2)] dA \\ &= \int_0^{2\pi} d\theta \int_0^2 r[4 - r^2]r dr = \frac{128\pi}{15}, \end{aligned}$$

where the double integral over D_{yz} has been converted to polar coordinates in the yz plane. \square

103.5. Study Problems.

Problem 14.6. Evaluate the triple integral of $f(x, y, z) = z$ over the region E bounded by the cylinder $x^2 + z^2 = 1$ 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., $z_{\text{top}}(x, y) = \sqrt{1 - x^2}$) (by taking the positive solution of $x^2 + z^2 = 1$). The integration region is shown in the right panel of Figure 14.24. The region D_{xy} is bounded by the lines of intersection of the planes $x = 0$, $y = x$, and of the planes $x = 0$, $y = x$, and $y = 1$. Thus, D_{xy} is the triangle bounded by the lines $x = 0$, $y = 1$, and $y = x$. One has

$$\begin{aligned} \iiint_E z dV &= \iint_{D_{xy}} \int_0^{\sqrt{1-x^2}} z dz dA \\ &= \frac{1}{2} \iint_{D_{xy}} (1 - x^2) dA = \frac{1}{2} \int_0^1 (1 - x^2) \int_x^1 dy dx = \frac{5}{24}, \end{aligned}$$

where the double integral has been evaluated by using the description of D_{xy} as a vertically simple region, $y_{\text{bot}} = x \leq y \leq 1 = y_{\text{top}}$ for all $x \in [0, 1] = [a, b]$. \square

Problem 14.7. 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,

$$\begin{aligned} \iiint_E xy^2z^3 dV &= \iiint_{E_1} xy^2z^3 dV - \iiint_{E_2} xy^2z^3 dV \\ &= 0 - \int_0^1 x dx \int_0^1 y^2 dy \int_0^1 z^3 dz = -\frac{1}{24}. \end{aligned}$$

The triple integral over E_1 vanishes by the symmetry argument (the ball is symmetric under the reflection $(-x, y, z) \rightarrow (-x, y, z)$ whereas $f(-x, y, z) = -f(x, y, z)$). The second integral is evaluated by Fubini's theorem. \square

103.6. Exercises.

(1) Evaluate the triple integral over the specified solid region by converting it to an appropriate iterated integral:

- (i) $\iiint_E (xy - 3z^2) dV$, where $E = [0, 1] \times [1, 2] \times [0, 2]$
- (ii) $\iiint_E 6xz dV$, where E is defined by the inequalities $0 \leq x \leq z$, $0 \leq y \leq x + z$, and $0 \leq z \leq 1$
- (iii) $\iiint_E ze^{y^2} dV$, where E is defined by the inequalities $0 \leq x \leq y$, $0 \leq y \leq z$, and $0 \leq z \leq 1$
- (iv) $\iiint_E 6xy dV$, where E lies under the plane $x + y - z = -1$ and above the region in the xy plane bounded by the curves $x = \sqrt{y}$, $x = 0$, and $y = 1$
- (v) $\iiint_E xy dV$, where E is bounded by the parabolic cylinders $y = x^2$, $x = y^2$, and by the planes $x + y - z = 0$, $x + y + z = 0$
- (vi) $\iiint_E dV$, where E is bounded by the coordinate planes and the plane through the points $(a, 0, 0)$, $(0, b, 0)$, and $(0, 0, c)$ with a, b, c being positive numbers
- (vii) $\iiint_E zx dV$ where E lies in the first octant between two planes $x = y$ and $x = 0$ and bounded by the cylinder $y^2 + z^2 = 1$
- (viii) $\iiint_E (x^2z + y^2z) dV$, where E is enclosed by the paraboloid $z = 1 - x^2 - y^2$ and the plane $z = 0$
- (ix) $\iiint_E z dV$, where E is enclosed by the elliptic paraboloid $z = 1 - x^2/a^2 - y^2/b^2$ and the plane $z = 0$. *Hint:* Use a suitable change of variables in the double integral.

- (x) $\iiint_E xy^2z^3 dV$, where E is bounded by the surfaces $z = xy$, $y = x$, $x = 1$, and $z = 0$
- (xi) $\iiint_E (1 + x + y + z)^{-3} dV$, where E is bounded by the plane $x + y + z = 1$ and by the coordinate planes
- (2) Use the triple integral to find the volume of the specified solid E .
- (i) E is bounded by the parabolic cylinder $x = y^2$ and the planes $z = 0$ and $z + x = 1$.
- (ii) E lies in the first octant and is bounded by the parabolic sheet $z = 4 - y^2$ and by two planes $y = x$ and $y = 2x$.
- (iii) E is bounded by the surfaces $z^2 = xy$, $x + y = a$, and $x + y = b$, where $0 < a < b$.
- (iv) E is bounded by the surfaces $z = x^2 + y^2$, $xy = a^2$, $xy = 2a^2$, $y = x/2$, $y = 2x$, and $z = 0$. *Hint:* Use a suitable change of variables in the double integral.
- (v) E is bounded by the surfaces $z = x^{3/2} + y^{3/2}$, $z = 0$, $x + y = 1$, $x = 0$, and $y = 0$.
- (vi) E is bounded by the surfaces $x^2/a^2 + y^2/b^2 + z/c = 1$, $(x/a)^{2/3} + (y/b)^{2/3} = 1$, and $z = 0$, where $c > 0$. *Hint:* Use the generalized polar coordinates to evaluate the integral. (see Study Problem 14.5)
- (vii) E is bounded by the surfaces $z = x + y$, $z = xy$, $x + y = 1$, $x = 0$, and $y = 0$.
- (viii) E is bounded by the surfaces $x^2 + z^2 = a^2$, $x + y = \pm a$, and $x - y = \pm a$.
- (ix) E is bounded by the surfaces $az = x^2 + y^2$, and $z = a - x - y$ and by the coordinate surfaces, where $a > 0$.
- (x) E is bounded by the surfaces $z = 6 - x^2 - y^2$ and $z = \sqrt{x^2 + y^2}$.
- (3) Use symmetry and other properties of the triple integral to evaluate:
- (i) $\iiint_E 24xy^2z^3 dV$, where E is bounded by the elliptic cylinder $(x/a)^2 + (y/b)^2 = 1$ and by the paraboloids $z = \pm[c - (x/a)^2 - (y/b)^2]$ and has the rectangular cavity $x \in [0, 1]$, $y \in [-1, 1]$, and $z \in [0, 1]$. Assume that a , b , and c are larger than 2.
- (ii) $\iiint_E (\sin^2(xz) - \cos^2(xy)) dV$, where E lies between the spheres: $1 \leq x^2 + y^2 + z^2 \leq 4$.
- (4) Express the integral $\iiint_E f dV$ as an iterated integral in six different ways, where E is the solid bounded by the specified surfaces:
- (i) $x^2 + y^2 = 4$, $z = -1$, $z = 2$
- (ii) $z + y = 1$, $x = 0$, $x = y^2$
- (5) Reverse the order of integration in all possible ways:

- (i) $\int_0^1 \int_0^{1-x} \int_0^{x+y} f(x, y, z) dz dy dx$
(ii) $\int_{-1}^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \int_{\sqrt{x^2+y^2}}^1 f(x, y, z) dz dy dx$
(iii) $\int_0^1 \int_0^1 \int_0^{x^2+y^2} f(x, y, z) dz dy dx$

(6) Reduce the following iterated integral to a single integral:

- (i) $\int_0^a \int_0^x \int_0^y f(z) dz dy dx$
(ii) $\int_0^1 \int_0^1 \int_0^{x+y} f(z) dz dy dx$ *Hint:* Reverse the order of integration in a suitable way.

(7) Use the interpretation of the triple integral f over E as the total amount of some quantity in E distributed with the density f to find E for which $\iiint_E (1 - x^2/a^2 - y^2/b^2 - z^2/c^2) dV$ is maximal.

(8) Prove the following representation of the triple integral by iterated integrals:

$$\iiint_E f(x, y, z) dV = \int_a^b \iint_{D_z} f(x, y, z) dA dz,$$

where D_z is the cross section of E by the plane $z = \text{const.}$

(9) Prove that if $f(x, y, z)$ is continuous in E and for any subregion W of E , $\iiint_W f dV = 0$, then $f(x, y, z) = 0$ in E .

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

$$(14.16) \quad x = r \cos \theta, \quad y = r \sin \theta, \quad z = z.$$

In any plane parallel to the xy plane, the points are labeled by polar coordinates, while the z coordinate is not transformed. Equation (14.16) defines 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 to an *image* region E . From the study of polar coordinates, the transformation (14.16) is one-to-one if $r \in (0, \infty)$,

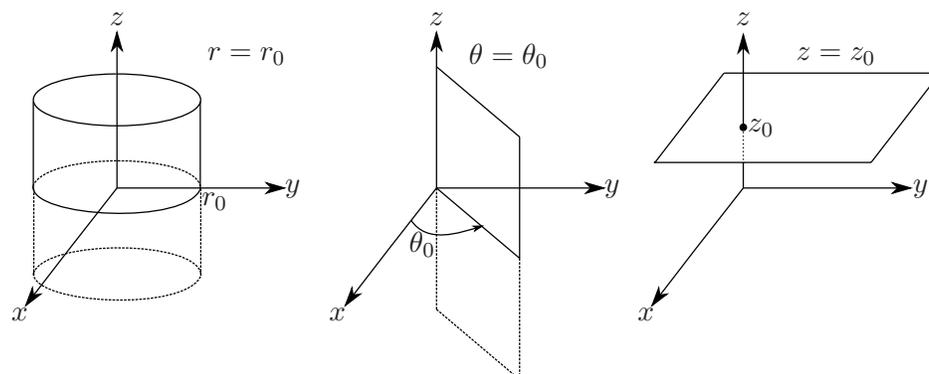


FIGURE 14.25. Coordinate surfaces of cylindrical coordinates: cylinders $r = r_0$, half-planes $\theta = \theta_0$ bounded by the z axis, and horizontal planes $z = z_0$. Any point in space can be viewed as the point of intersection of three coordinate surfaces.

$\theta \in [0, 2\pi)$, 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) belongs (see the discussion of polar coordinates). It maps any region E in the Euclidean space spanned by (x, y, z) to 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. It is particularly important 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 it is an image of a rectangular box E' , 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

$$\begin{aligned}r = r_0 &\leftrightarrow x^2 + y^2 = r_0^2 && \text{(cylinder),} \\ \theta = \theta_0 &\leftrightarrow y \cos \theta_0 = x \sin \theta_0 && \text{(half-plane),} \\ z = z_0 &\leftrightarrow z = z_0 && \text{(plane).}\end{aligned}$$

The coordinate surfaces of cylindrical coordinates are shown in Figure 14.25. A point in space corresponding to an ordered triple (r_0, θ_0, z_0) is an intersection point of a cylinder, a half-plane bounded by the cylinder axis, and a plane perpendicular to the cylinder axis.

EXAMPLE 14.24. Find the region E' whose image under the transformation (14.16) is 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,

$$E' = \left\{ (r, \theta, z) \mid r^2 \leq z \leq 4, (r, \theta) \in [0, 2] \times [0, \pi/4] \right\}.$$

□

104.2. 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 14.13, 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 coordinates, one infers

that

$$\begin{aligned}
 \iiint_E f(x, y, z) dV &= \iint_{D'_{xy}} \int_{z_{\text{bot}}(r, \theta)}^{z_{\text{top}}(r, \theta)} f(r \cos \theta, r \sin \theta, z) r dz dA' \\
 (14.17) \qquad \qquad &= \iiint_{E'} f(r \cos \theta, r \sin \theta, z) r dV',
 \end{aligned}$$

where the region E' is the image of E under the transformation from rectangular to cylindrical coordinates,

$$E' = \{(r, \theta, z) \mid z_{\text{bot}}(r, \theta) \leq z \leq z_{\text{top}}(r, \theta), (r, \theta) \in D'_{xy}\},$$

and $z = z_{\text{bot}}(r, \theta)$, $z = z_{\text{top}}(r, \theta)$ 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 as shown in the left panel of Figure 14.26. 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, $dA = r dA'$. 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 14.25. Evaluate the triple integral of $f(x, y, z) = x^2 z$ 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 below by the plane $z = 0$, by the circular paraboloid from above, 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

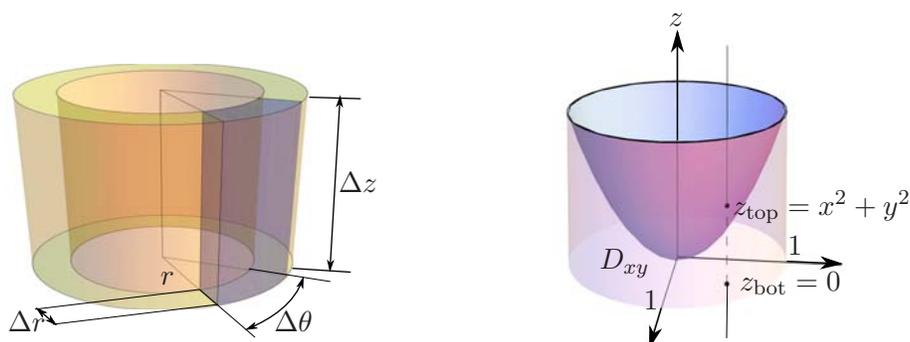


FIGURE 14.26. **Left:** A partition element of the partition of E by cylinders, half-planes, and horizontal planes (coordinate surfaces of cylindrical coordinates). The partition is the image of a rectangular partition of E' . Keeping only terms linear in the differentials $dr = \Delta r$, $d\theta = \Delta\theta$, $dz = \Delta z$, the volume of the partition element is $dV = dA dz = r dr d\theta dz = r dV'$, where $dA = r dr d\theta$ is the area element in the polar coordinates. So the Jacobian of cylindrical coordinates is $J = r$. **Right:** An illustration to Example 14.25.

bottom boundaries are $z = z_{\text{top}}(r, \theta) = r^2$ and $z = z_{\text{bot}}(r, \theta) = 0$. Hence,

$$\begin{aligned} \iiint_E x^2 z \, dV &= \int_0^{2\pi} \int_0^1 \int_0^{r^2} r^2 \cos^2 \theta \, z \, r \, dz \, dr \, d\theta \\ &= \frac{1}{2} \int_0^{2\pi} \cos^2 \theta \, d\theta \int_0^1 r^7 \, dr = \frac{\pi}{16}, \end{aligned}$$

where the double-angle formula, $\cos^2 \theta = (1 + \cos(2\theta))/2$, has been used to evaluate the integral. \square

104.3. 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

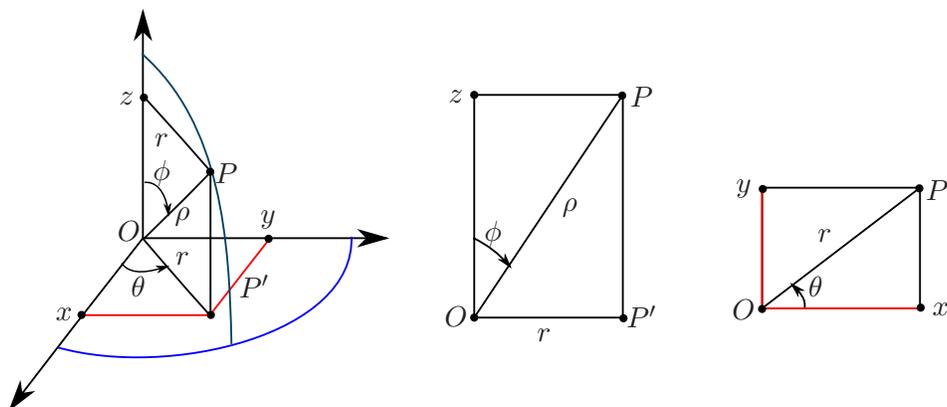


FIGURE 14.27. Spherical coordinates and their relation to the rectangular coordinates. A point P in space is defined by its distance to the origin ρ , the angle ϕ between the positive z axis and the ray OP , and the polar angle θ .

from the origin through $P = (x, y, z)$ and the rectangle with vertices $(0, 0, 0)$, $(0, 0, z)$, $P' = (x, y, 0)$, and (x, y, z) in this plane (see Figure 14.27). 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

$$(14.18) \quad x = \rho \sin \phi \cos \theta, \quad y = \rho \sin \phi \sin \theta, \quad z = \rho \cos \phi.$$

The inverse transformation follows from the geometrical interpretation of the spherical coordinates:

$$(14.19) \quad \rho = \sqrt{x^2 + y^2 + z^2}, \quad \cot \phi = \frac{z}{r} = \frac{z}{\sqrt{x^2 + y^2}}, \quad \tan \theta = \frac{y}{x}.$$

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.

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 rectangular box. 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 14.26. *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 the region E' that is mapped onto E by the transformation $(\rho, \phi, \theta) \rightarrow (x, y, z)$.*

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 the images of $\rho = 2$, $\cot \phi = \sqrt{3}$ or $\phi = \pi/3$, $\theta = 0$, and $\theta = \pi/2$, respectively. So E' is the rectangular box $[0, 2] \times [0, \pi/3] \times [0, \pi/2]$. 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 $0 \leq \theta \leq \pi/2$. \square

104.4. Triple Integrals in Spherical Coordinates. A triple integral in spherical coordinates is obtained by partitioning the integration region E by spheres, cones, and half-planes, constructing the Riemann sum (14.13), and taking its limit under a refinement of the partition. Let E' be

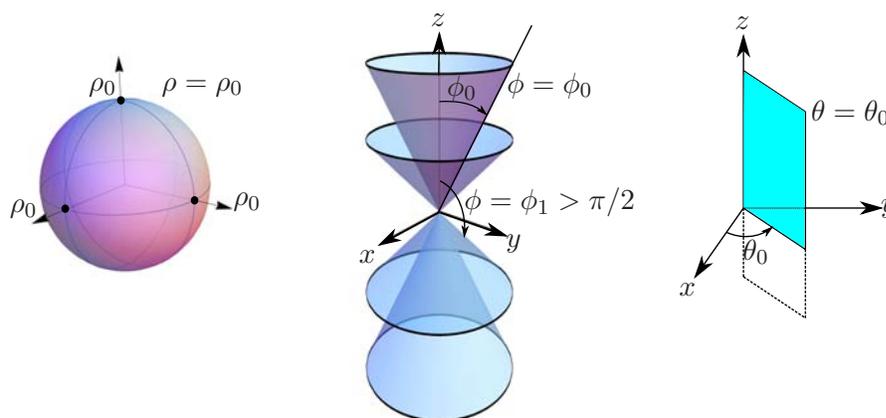


FIGURE 14.28. Coordinate surfaces of spherical coordinates: spheres $\rho = \rho_0$, circular cones $\phi = \phi_0$, and half-planes $\theta = \theta_0$ bounded by the z axis. In particular, $\phi = 0$ and $\phi = \pi$ describe the positive and negative z axes, respectively, and the cone with the angle $\phi = \pi/2$ becomes the xy plane.

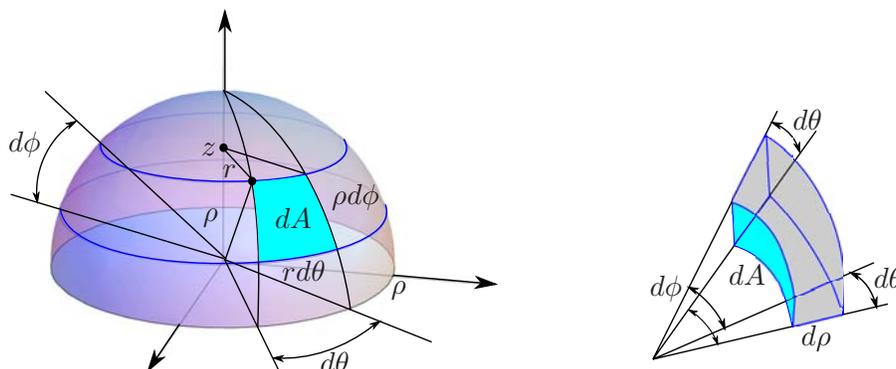


FIGURE 14.29. **Left:** The base of a partition element in spherical coordinates is a portion of a sphere of radius ρ cut out by two cones with the angles ϕ and $\phi + d\phi = \phi + \Delta\phi$ and by two half-planes with the angles θ and $\theta + d\theta = \theta + \Delta\theta$. Its area is $dA = (\rho d\phi) \cdot (r d\theta) = \rho \sin \phi d\phi d\theta$. **Right:** A partition element has the height $d\rho = \Delta\rho$ as it lies between two spheres whose radii differ by $d\rho$. So its volume is $dV = dA d\rho = \rho^2 \sin \phi d\rho d\phi d\theta = J dV'$, and the Jacobian of spherical coordinates is $J = \rho^2 \sin \phi$.

mapped onto a region E under the transformation (14.18). 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 element 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 $\Delta\theta$ as shown in Figure 14.29. 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 $(\Delta\rho, \Delta\phi, \Delta\theta) \rightarrow (0, 0, 0)$, 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 rectangular box $\rho \in [\rho_0, \rho_0 + \Delta\rho]$, $\phi \in [\phi_0, \phi_0 + \Delta\phi]$, $\theta \in [\theta_0, \theta_0 + \Delta\theta]$ 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 $\Delta b = r_0 \Delta\theta = \rho_0 \sin \phi_0 \Delta\theta$. 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

$$dV = J dV', \quad J = \rho^2 \sin \phi.$$

In a Euclidean space spanned by ordered triples (ρ, ϕ, θ) , the Jacobian vanishes on the “planes” $\rho = 0$, $\phi = 0$, and $\phi = \pi$. The transformation (14.18) is not one-to-one on them. All points $(0, \phi, \theta)$ are mapped to a single point $(0, 0, 0)$ by (14.18), and all points $(\rho, 0, \theta)$ and (ρ, π, θ) are mapped onto the z axis, that is, the line $(0, 0, z)$, $-\infty < z < \infty$.

By the continuity of the Jacobian, the difference between the values of J at any two sample points in a partition rectangle in E' vanishes in the limit $(\Delta\rho, \Delta\phi, \Delta\theta) \rightarrow (0, 0, 0)$; 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 of a Riemann sum. Therefore, for any choice of sample points, the limit of the Riemann sum (14.13) for the constructed partition is

$$\begin{aligned} \iiint_E f(x, y, z) dV \\ = \iiint_{E'} f(\rho \sin \phi \cos \theta, \rho \sin \phi \sin \theta, \rho \cos \phi) \rho^2 \sin \phi dV'. \end{aligned}$$

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 14.27. 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 solid is shown in Figure 14.30. The boundaries of E impose no restriction on θ , which can therefore be taken over its full range. Hence, the region E' whose image is E admits the following algebraic description:

$$E' = \left\{ (\rho, \phi, \theta) \mid 0 \leq \rho \leq 2 \cos \phi, (\phi, \theta) \in [0, \pi/4] \times [0, 2\pi] \right\}.$$

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

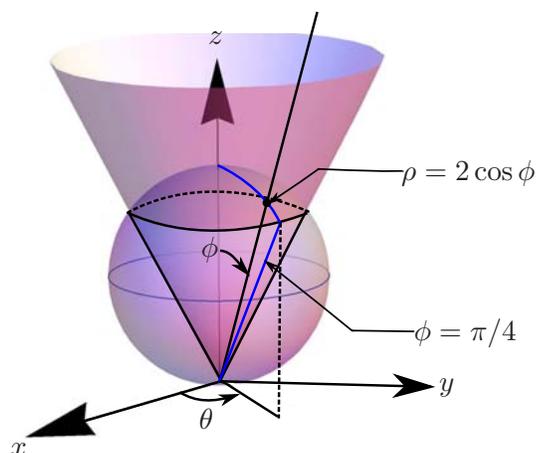


FIGURE 14.30. An illustration to Example 14.27. Any ray in space is defined as the intersection of a cone with an angle ϕ and a half-plane with an angle θ . To find E' whose image is the depicted solid E , note that any such ray intersects E along a *single* straight line segment if $0 \leq \phi \leq \pi/4$, where the cone $\phi = \pi/4$ is a part of the boundary of E . Due to the axial symmetry of E , there is no restriction on the range of θ , that is, $0 \leq \theta \leq 2\pi$ in E' . The range of ρ is determined by the length of the segment of intersection of the ray at fixed ϕ and θ with E : $0 \leq \rho \leq 2 \cos \phi$, where $\rho = 2 \cos \phi$ is the equation of the top boundary of E in spherical coordinates.

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. \square

104.5. Exercises.

(1) Sketch the solid E onto which the specified region E' is mapped by the transformation $(r, \theta, z) \rightarrow (x, y, z)$:

- (i) $0 \leq r \leq 3$, $\pi/4 \leq \theta \leq 5\pi/4$, $0 \leq z \leq 1$
- (ii) $0 \leq r \leq 1$, $0 \leq \theta \leq 2\pi$, $r - 1 \leq z \leq 1 - r$

(iii) $0 \leq r \leq 2, 0 \leq \theta \leq \pi/2, 0 \leq z \leq 4 - r^2$

(2) Given the solid E , find the region E' whose image is E under the transformation to cylindrical coordinates:

- (i) E is bounded by the cylinder $x^2 + y^2 = 1$, the paraboloid $z = x^2 + y^2$, and the plane $z = 0$.
- (ii) E is bounded by the cone $(z - 1)^2 = x^2 + y^2$ and the cylinder $x^2 + y^2 = 1$.
- (iii) E is bounded by the paraboloid $z = x^2 + y^2$, the cylinder $x^2 + y^2 = 2x$, and the plane $z = 0$.
- (iv) E is the part of the ball $x^2 + y^2 + z^2 \leq a^2$ in the first octant.

(3) Evaluate the triple integral by converting it to cylindrical coordinates:

- (i) $\iiint_E |z| dV$, where E is bounded by the sphere $x^2 + y^2 + z^2 = 4$ and the cylinder $x^2 + y^2 = 1$
- (ii) $\iiint_E (x^2y + y^3) dV$, where E lies beneath the paraboloid $z = 1 - x^2 - y^2$ in the first octant
- (iii) $\iiint_E y dV$, where E is enclosed by the planes $z = 0, x + y - z = -5$ and by the cylinders $x^2 + y^2 = 1, x^2 + y^2 = 4$
- (iv) $\iiint_E dV$, where E is enclosed by the cylinder $x^2 + y^2 = 2x$, by the plane $z = 0$, and by the cone $z = \sqrt{x^2 + y^2}$
- (v) $\iiint_E yz dV$, where E lies beneath the paraboloid $z = a^2 - x^2 - y^2$ in the first octant
- (vi) $\iiint_E (x^2 + y^2) dV$, where E is bounded by the surfaces $x^2 + y^2 = 2z$ and $z = 2$
- (vii) $\iiint_E xyz dV$, where E lies in the positive octant and is bounded by the surfaces $x^2 + y^2 = az, x^2 + y^2 = bz, xy = c^2, xy = k^2, y = \alpha x, y = \beta x$, and $0 < a < b, 0 < \alpha < \beta, 0 < c < k$

(4) Sketch the solid E onto which the specified region E' is mapped by the transformation $(\rho, \phi, \theta) \rightarrow (x, y, z)$:

- (i) $0 \leq \rho \leq 1, 0 \leq \phi \leq \pi/2, 0 \leq \theta \leq \pi/4$
- (ii) $1 \leq \rho \leq 2, 0 \leq \phi \leq \pi/4, 0 \leq \theta \leq \pi/2$
- (iii) $\frac{1}{\cos \phi} \leq \rho \leq 2, 0 \leq \phi \leq \pi/6, 0 \leq \theta \leq \pi$

(5) Given the solid E , find the region E' whose image is E under the transformation to spherical coordinates:

- (i) E lies between two spheres $x^2 + y^2 + z^2 = 1$ and $x^2 + y^2 + z^2 = 4$ in the first octant.
- (ii) E is defined by the inequalities $z^2 \leq 3(x^2 + y^2)$ and $x^2 + y^2 + z^2 \leq a^2$.

- (iii) E bounded by the sphere $x^2 + y^2 + z^2 = a^2$ and by the half-planes $y = \sqrt{3}x$, $y = x/\sqrt{3}$ where $x \geq 0$.

(6) Evaluate the triple integral by converting it to spherical coordinates:

- (i) $\iiint_E (x^2 + y^2 + z^2)^3 dV$, where E is the ball of radius a centered at the origin
- (ii) $\iiint_E y^2 dV$, where E is bounded by the yz plane and the hemispheres $x = \sqrt{1 - y^2 - z^2}$ and $x = \sqrt{4 - y^2 - z^2}$
- (iii) $\iiint_E xyz dV$, where E is enclosed by the cone $z = \sqrt{3}\sqrt{x^2 + y^2}$ and the spheres $x^2 + y^2 + z^2 = a^2$, $a = 1, 2$
- (iv) $\iiint_E z dV$, where E is the part of the ball $x^2 + y^2 + z^2 \leq 1$ that lies below the cone $z = \sqrt{3x^2 + 3y^2}$
- (v) $\iiint_E z dV$, where E lies in the first octant between the planes $y = 0$ and $x = \sqrt{3}y$ and is also bounded by the surfaces $z = \sqrt{x^2 + y^2}$ and $x^2 + y^2 + z^2 = 4$
- (vi) $\iiint_E \sqrt{x^2 + y^2 + z^2} dV$, where E is bounded by the sphere $x^2 + y^2 + z^2 = z$

(7) Sketch the region of integration, write the triple integral in spherical coordinates, and then evaluate it:

- (i) $\int_0^1 \int_0^{\sqrt{1-x^2}} \int_{\sqrt{x^2+y^2}}^1 z dz dy dx$
- (ii) $\int_0^1 \int_0^{\sqrt{1-x^2}} \int_{\sqrt{x^2+y^2}}^{\sqrt{2-x^2-y^2}} z^2 dz dy dx$

(8) Sketch the solid whose volume is given by the iterated integral in the spherical coordinates:

$$\int_0^{\pi/2} \int_0^{\pi/4} \int_0^{2/\cos\phi} \rho^2 \sin\phi d\rho d\phi d\theta$$

Write the integral in the cylindrical coordinates and then compute it.

(9) Sketch the domain of integration, write the triple integral in cylindrical coordinates, and then evaluate it:

$$\int_{-1}^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \int_0^{1-x^2-y^2} z dz dy dx$$

(10) Convert the triple integral $\iiint_E f(x^2 + y^2 + z^2) dV$ to iterated integrals in cylindrical and spherical coordinates if E is bounded by the surfaces:

- (i) $z = x^2 + y^2$, $y = x$, $x = 1$, $y = 0$, $z = 0$
- (ii) $z^2 = x^2 + y^2$, $x^2 + y^2 + z^2 = 2z$, $x = y/\sqrt{3}$, $x = y\sqrt{3}$, where $x \geq 0$ and $y \geq 0$

(11) Use spherical coordinates to find the volume of a solid bounded by the surfaces:

- (i) $x^2 + y^2 + z^2 = a^2$, $x^2 + y^2 + z^2 = b^2$, $z = \sqrt{x^2 + y^2}$

- (ii) $(x^2 + y^2 + z^2)^3 = a^6 z^2 / (x^2 + y^2)$
 (iii) $(x^2 + y^2 + z^2)^2 = a^2(x^2 + y^2 - z^2)$
 (iv) $(x^2 + y^2 + z^2)^3 = 3xyz$

(12) Find the volume of a solid bounded by the surfaces $x^2 + z^2 = a^2$, $x^2 + z^2 = b^2$, $x^2 + y^2 = z^2$, where $x > 0$.

105. Change of Variables in Triple Integrals

Consider a transformation of an open region E' in space into a region E defined by $x = x(u, v, w)$, $y = y(u, v, w)$, 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 $w = w(x, y, z)$. Suppose that these functions have continuous partial derivatives so that the gradient of these functions does not vanish. Then, as shown in Section 103.1, the equations $u(x, y, z) = u_0$, $v(x, y, z) = v_0$, and $w(x, y, z) = w_0$ define smooth surfaces, called *coordinate surfaces* of the new variables. 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 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 point (u_0, v_0, w_0) in E' is mapped to \mathbf{r}_0 by the coordinate transformation.

DEFINITION 14.15. (Jacobian of a Transformation).

Suppose that a one-to-one transformation 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 transformation*.

If the determinant is expanded over the first column, then it can also be written as the triple product:

$$\frac{\partial(x, y, z)}{\partial(u, v, w)} = \nabla x \cdot (\nabla y \times \nabla z).$$

The technical details are left to the reader as an exercise.

DEFINITION 14.16. (Change of Variables).

Let a transformation of an open set E' onto E have continuous partial derivatives. It is called a change of variables (or a change of coordinates) if its Jacobian does not vanish in E' .

The inverse function theorem (Theorem 14.10) holds for a transformation $(u, v, w) \rightarrow (x, y, z)$. If the Jacobian of the transformation does not vanish on E' , then the inverse transformation $(x, y, z) \rightarrow (u, v, w)$ exists and has continuous partial derivatives.

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 rectangular box, 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 $(x, y, z) = \mathbf{r}$ so that the change of variables is written as

$$(14.20) \quad \mathbf{r} = (x(\mathbf{r}'), y(\mathbf{r}'), z(\mathbf{r}')) \quad \text{or} \quad \mathbf{r}' = (u(\mathbf{r}), v(\mathbf{r}), w(\mathbf{r})).$$

Let E'_0 be a rectangular box in E' , $u \in [u_0, u_0 + \Delta u]$, $v \in [v_0, v_0 + \Delta v]$, and $w \in [w_0, w_0 + \Delta w]$. Under the transformation $(u, v, w) \rightarrow (x, y, z)$, its image E_0 is bounded by smooth surfaces if the transformation is a change of variables. If the values of Δu , Δv , and Δw are infinitesimally small, that is, they can be viewed as differentials of the new variables, then the boundary surfaces of E_0 can be well approximated by tangent planes to them, and the volume of E_0 is then approximated by the volume of the polyhedron bounded by these planes. This implies, in particular, that when calculating the volume, only terms linear in Δu , Δv , and Δw are to be retained, while their higher powers are neglected. Therefore, the volumes of E_0 and E'_0 must be proportional:

$$\Delta V = J \Delta V', \quad \Delta V' = \Delta u \Delta v \Delta w.$$

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 rectangular box E'_0 is taken. The derivation of J is fully analogous to the two-variable case.

An infinitesimal rectangular box E'_0 and its image under the coordinate transformation are shown in Figure 14.31. Let O' , A' , B' , and

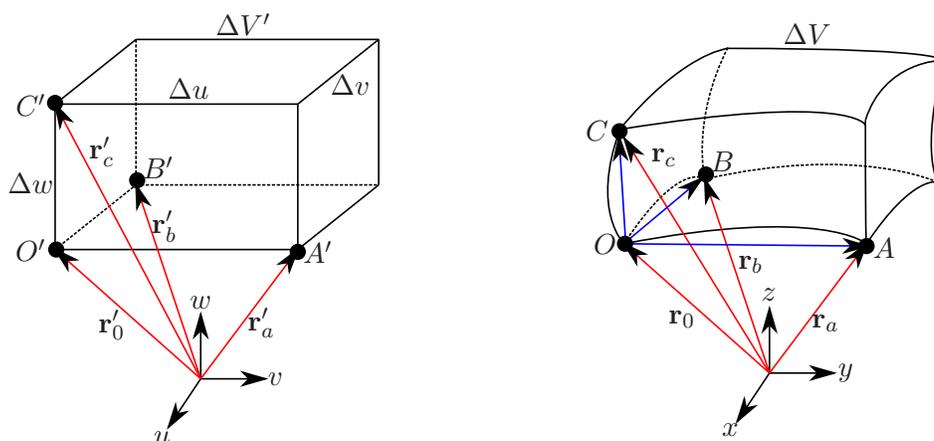


FIGURE 14.31. **Left:** A rectangular box in the region E'_0 with infinitesimal sides $du = \Delta u$, $dv = \Delta v$, $dw = \Delta w$ so that its volume $\Delta V' = du dv dw$. **Right:** The image of the rectangular box under a change of variables. The position vectors \mathbf{r}_p , where $P = 0, a, b, c$, are images of the position vectors \mathbf{r}'_p . The volume ΔV of the image is approximated by the volume of the parallelepiped with adjacent sides OA , OB , and OC . It is computed by linearization of ΔV in du , dv , and dw so that $\Delta V = J du dv dw = J \Delta V'$, where $J > 0$ is the Jacobian of the change of variables.

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 rectangular box E'_0 . Let O , A , B , and C be the images of O' , A' , B' , and C' in the region E . Owing to the smoothness of the boundaries of E_0 , the volume ΔV of E_0 can be approximated by the volume of the parallelepiped with adjacent sides $\mathbf{a} = \overrightarrow{OA}$, $\mathbf{b} = \overrightarrow{OB}$, and $\mathbf{c} = \overrightarrow{OC}$. Then

$$\begin{aligned}\mathbf{a} &= \left(x(\mathbf{r}'_a) - x(\mathbf{r}'_0), y(\mathbf{r}'_a) - y(\mathbf{r}'_0), z(\mathbf{r}'_a) - z(\mathbf{r}'_0) \right) = (x'_u, y'_u, z'_u) \Delta u, \\ \mathbf{b} &= \left(x(\mathbf{r}'_b) - x(\mathbf{r}'_0), y(\mathbf{r}'_b) - y(\mathbf{r}'_0), z(\mathbf{r}'_b) - z(\mathbf{r}'_0) \right) = (x'_v, y'_v, z'_v) \Delta v, \\ \mathbf{c} &= \left(x(\mathbf{r}'_c) - x(\mathbf{r}'_0), y(\mathbf{r}'_c) - y(\mathbf{r}'_0), z(\mathbf{r}'_c) - z(\mathbf{r}'_0) \right) = (x'_w, y'_w, z'_w) \Delta w,\end{aligned}$$

where all the differences have been linearized, for instance, $x(\mathbf{r}'_a) - x(\mathbf{r}'_0) = x(\mathbf{r}'_0 + \hat{\mathbf{e}}_1 \Delta u) - x(\mathbf{r}'_0) = x'_u(\mathbf{r}'_0) \Delta u$. Because of differentiability of the functions $x(\mathbf{r}')$, $y(\mathbf{r}')$, and $z(\mathbf{r}')$, the error of this approximation decreases to 0 faster than Δu , Δv , Δw as the latter approach zero values. This justifies the approach based on retaining only terms linear in Δu , Δv , Δw when calculating the volume. The volume of the parallelepiped is given by the absolute value of the triple product:

$$(14.21) \quad \Delta V = |\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})| = \left| \det \begin{pmatrix} x'_u & y'_u & z'_u \\ x'_v & y'_v & z'_v \\ x'_w & y'_w & z'_w \end{pmatrix} \right| \Delta u \Delta v \Delta w = J \Delta V',$$

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. If the Jacobian of the transformation does not vanish, then by the inverse function theorem (Theorem 14.10) there exists an inverse transformation, and, similarly to the two-dimensional case (compare with (14.11)), it can be proved that

$$(14.22) \quad J = \left| \frac{\partial(x, y, z)}{\partial(u, v, w)} \right| = \frac{1}{\left| \frac{\partial(u, v, w)}{\partial(x, y, z)} \right|} = \left| \det \begin{pmatrix} u'_x & u'_y & u'_z \\ v'_x & v'_y & v'_z \\ w'_x & w'_y & w'_z \end{pmatrix} \right|^{-1} \\ = \left| \nabla u \cdot (\nabla v \times \nabla w) \right|^{-1}.$$

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 rectangular box $u \in [u_i, u_{i+1}]$, $v_j \in [v_j, v_{j+1}]$, 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 $(\Delta u, \Delta v, \Delta w) \rightarrow (0, 0, 0)$. The volume ΔV_{ijk} of E_{ijk} is related to the volume of the rectangle E'_{ijk} by (14.21). By the 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' . The above qualitative consideration suggests that the following theorem holds.

THEOREM 14.15. (Change of Variables in a Triple Integral).

Let a transformation $E' \rightarrow E$ defined by functions $(u, v, w) \rightarrow (x, y, z)$ with continuous partial derivatives have a nonvanishing 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 14.28. (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: $V = (4\pi/3)R^3$.

EXAMPLE 14.29. Let \mathbf{a} , \mathbf{b} , and \mathbf{c} be non-coplanar vectors. Find the volume of a solid E bounded by the surface $(\mathbf{a} \cdot \mathbf{r})^2 + (\mathbf{b} \cdot \mathbf{r})^2 + (\mathbf{c} \cdot \mathbf{r})^2 = R^2$, where $\mathbf{r} = (x, y, z)$.

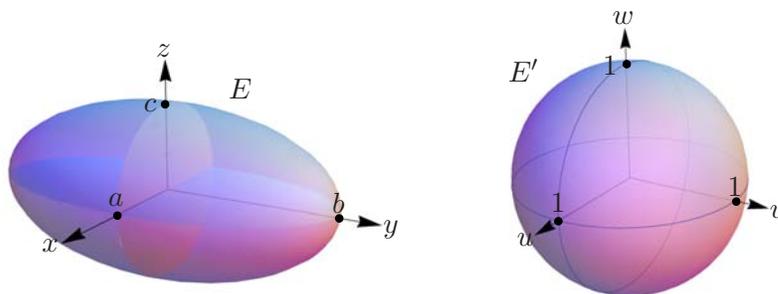


FIGURE 14.32. An illustration to Example 14.28. The ellipsoidal region $x^2/a^2 + y^2/b^2 + z^2/c^2 \leq 1$ is mapped onto the ball $u^2 + v^2 + w^2 \leq 1$ by the coordinate transformation $u = x/a$, $v = y/b$, $w = z/c$ with the Jacobian $J = abc$.

SOLUTION: Define new variables by the transformation $u = \mathbf{a} \cdot \mathbf{r}$, $v = \mathbf{b} \cdot \mathbf{r}$, $w = \mathbf{c} \cdot \mathbf{r}$. The Jacobian of this transformation is obtained by (14.22):

$$\begin{aligned} \frac{\partial(x, y, z)}{\partial(u, v, w)} &= \left(\frac{\partial(u, v, w)}{\partial(x, y, z)} \right)^{-1} = \left(\nabla u \cdot (\nabla v \times \nabla w) \right)^{-1} \\ &= \left(\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) \right)^{-1}. \end{aligned}$$

The vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} are non-coplanar, and hence their triple product is not 0. So the transformation, is a genuine change of variables. Under this transformation, the boundary of E becomes a sphere $u^2 + v^2 + w^2 = R^2$. So

$$\begin{aligned} V(E) &= \iiint_E dV = \iiint_{E'} J dV' = \frac{1}{|\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})|} \iiint_{E'} dV' \\ &= \frac{V(E')}{|\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})|} = \frac{4\pi R^3}{3|\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})|}, \end{aligned}$$

where $V(E') = 4\pi R^3/3$ is the volume of a ball of radius R . \square

105.3. Study Problems.

Problem 14.8. (Volume of a Tetrahedron).

A tetrahedron is a solid with four vertices and four triangular faces. Let the vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} be three adjacent sides of the tetrahedron. Find its volume.

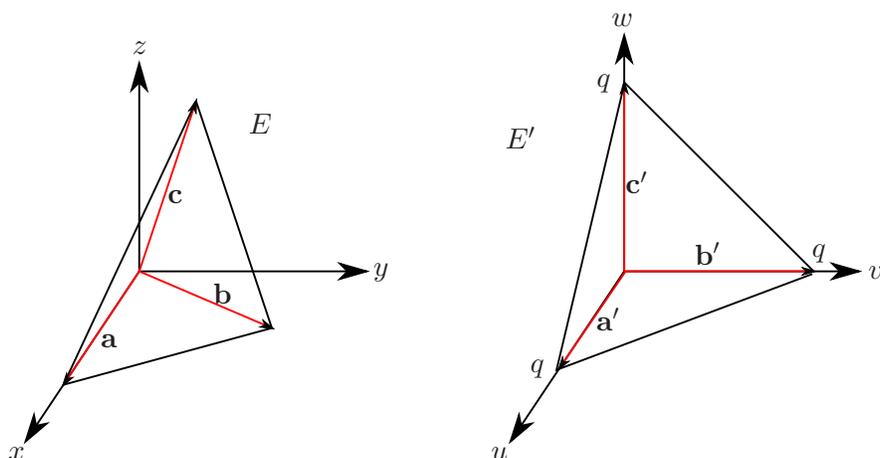


FIGURE 14.33. An illustration to Study Problem 14.8. A general tetrahedron is transformed to a tetrahedron whose faces lie in the coordinate planes by a change of variables.

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 \mathbf{a} and \mathbf{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:

$$\begin{aligned} \mathbf{n} \cdot \mathbf{r} = 0 & \quad \text{or} \quad n_1x + n_2y + n_3z = 0, & \quad \mathbf{n} = \mathbf{a} \times \mathbf{b}, \\ \mathbf{l} \cdot \mathbf{r} = 0 & \quad \text{or} \quad l_1x + l_2y + l_3z = 0, & \quad \mathbf{l} = \mathbf{c} \times \mathbf{a}, \\ \mathbf{m} \cdot \mathbf{r} = 0 & \quad \text{or} \quad m_1x + m_2y + m_3z = 0, & \quad \mathbf{m} = \mathbf{b} \times \mathbf{c}, \end{aligned}$$

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 straightforward 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. \square

105.4. Exercises.

(1) Find the Jacobian of the following transformations:

- (i) $x = u/v, y = v/w, z = w/u$
- (ii) $x = v + w^2, y = w + u^2, z = u + v^2$
- (iii) $x = uv \cos w, y = uv \sin w, z = (u^2 - v^2)/2$ (these coordinates are called *parabolic* coordinates)
- (iv) $x + y + z = u, y + z = uv, z = uvw$

(2) Find the region E' whose image E under the transformation defined in exercise 1, part (iv), is bounded by the coordinate planes and by the plane $x + y + z = 1$. In particular, investigate the image of those points in E' at which the Jacobian of the transformation vanishes.

(3) Let E be the solid region in the first octant defined by the inequality $\sqrt{x} + \sqrt{y} + \sqrt{z} \leq a$, where $a > 0$. Find its volume using the triple integral in the new variables $u = \sqrt{x}, v = \sqrt{y}, w = \sqrt{z}$.

(4) Use a suitable change of variables in the triple integral to find the volume of a solid bounded by the surfaces:

- (i) $(x/a)^{2/3} + (y/b)^{2/3} + (z/c)^{2/3} = 1$
 (ii) $(x/a)^{1/3} + (y/b)^{1/3} + (z/c)^{1/3} = 1$, where $x \geq 0, y \geq 0, z \geq 0$
 (iii) $x = 0, y = 0, z = 0$, and $(x/a)^n + (y/b)^m + (z/c)^k = 1$, where the numbers n, m , and k are positive
 (iv) $(x + y + z)^2 = ax + by$, where (x, y, z) lie in the first octant and a and b are positive
 (v) $(x + y)^2 + z^2 = R^2$, where (x, y, z) lie in the first octant

(5) Evaluate the triple integral $\iiint_E z \, dV$, where E lies above the cone $z = c\sqrt{x^2/a^2 + y^2/b^2}$ and is bounded from above by the ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$.

(6) Evaluate the triple integral $\iiint_E (4x^2 - 9y^2) \, dV$, where E is enclosed by the paraboloid $z = x^2/9 + y^2/4$ and the plane $z = 10$.

(7) Consider a linear transformation of the coordinates $x = \mathbf{a} \cdot \mathbf{r}'$, $y = \mathbf{b} \cdot \mathbf{r}'$, $z = \mathbf{c} \cdot \mathbf{r}'$, where $\mathbf{r}' = (u, v, w)$ and the vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} have constant components. Show that this transformation is *volume preserving* if $|\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})| = 1$. The transformation is said to be volume preserving if the image E' of any E has the same volume as E , that is, $V(E') = V(E)$.

(8) If \mathbf{a} , \mathbf{b} , and \mathbf{c} are constant vectors, $\mathbf{r} = (x, y, z)$, and E is given by the inequalities $0 \leq \mathbf{a} \cdot \mathbf{r} \leq \alpha$, $0 \leq \mathbf{b} \cdot \mathbf{r} \leq \beta$, and $0 \leq \mathbf{c} \cdot \mathbf{r} \leq \gamma$, show that

$$\iiint_E (\mathbf{a} \cdot \mathbf{r})(\mathbf{b} \cdot \mathbf{r})(\mathbf{c} \cdot \mathbf{r}) \, dV = \frac{1}{8}(\alpha\beta\gamma)^2 / |\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})|.$$

(9) Consider parabolic coordinates $x = uv \cos w$, $y = uv \sin w$, and $z = (u^2 - v^2)$. Show that $2z = (x^2 + y^2)/v^2 - v^2$, $2z = -(x^2 + y^2)/u^2 + u^2$, and $\tan w = y/x$. Use these relations to sketch the coordinate surfaces $u(x, y, z) = u_0$, $v(x, y, z) = v_0$, and $w(x, y, z) = w_0$. Evaluate the triple integral of $f(x, y, z) = xyz$ over the region E that lies in the first octant beneath the paraboloid $2z - 1 = -(x^2 + y^2)$ and above the paraboloid $2z + 1 = x^2 + v^2$ by converting to parabolic coordinates.

(10) Use a suitable change of variables to find the volume of a solid that is bounded by the surface

$$\left(\frac{x^2}{a^2} + \frac{y^2}{b^2}\right)^n + \frac{z^{2n}}{c^{2n}} = \frac{z}{h} \left(\frac{x^2}{a^2} + \frac{y^2}{b^2}\right)^{n-2}, \quad n > 1.$$

(11) (Generalized Spherical Coordinates) Generalized spherical coordinates (ρ, ϕ, θ) are defined by the equations

$$x = a\rho \sin^n \phi \cos^m \theta, \quad y = b\rho \sin^n \phi \sin^m \theta, \quad z = c\rho \cos^n \phi,$$

where $0 \leq \rho < \infty$, $0 \leq \theta < 2\pi$, $0 \leq \phi \leq \pi$, and a, b, c, n , and m are parameters. Find the Jacobian of the generalized spherical coordinates.

(12) Use generalized spherical coordinates with a suitable choice of parameters to find the volume of a solid bounded by the surfaces:

- (i) $[(x/a)^2 + (y/b)^2 + (z/c)^2]^2 = (x/a)^2 + (y/b)^2$
- (ii) $[(x/a)^2 + (y/b)^2 + (z/c)^2]^2 = (x/a)^2 + (y/b)^2 - (z/c)^2$
- (iii) $(x/a)^2 + (y/b)^2 + (z/c)^4 = 1$
- (iv) $[(x/a)^2 + (y/b)^2]^2 + (z/c)^4 = 1$

(13) (Dirichlet's Integral) Let n , m , p , and s be positive integers. Use the transformation defined by $x + y + z = u$, $y + z = uv$, $z = uvw$ to show that

$$\iiint_E x^n y^m z^p (1 - x - y - z)^s dV = \frac{n! m! p! s!}{(n + m + p + s + 3)!},$$

where E is the tetrahedron bounded by the coordinate planes and the plane $x + y + z = 1$.

(14) (Orthogonal Curvilinear Coordinates) Curvilinear coordinates (u, v, w) are called *orthogonal* if the normals to their coordinate surfaces are mutually orthogonal at any point of their intersection. In other words, the gradients $\nabla u(x, y, z)$, $\nabla v(x, y, z)$, and $\nabla w(x, y, z)$ are mutually orthogonal. One can define unit vectors orthogonal to the coordinate surfaces:

$$(14.23) \quad \hat{\mathbf{e}}_u = \frac{\nabla u}{\|\nabla u\|}, \quad \hat{\mathbf{e}}_v = \frac{\nabla v}{\|\nabla v\|}, \quad \hat{\mathbf{e}}_w = \frac{\nabla w}{\|\nabla w\|}.$$

Note that the Jacobian of a change of variables does not vanish and the relation (14.22) guarantees that these unit vectors are not coplanar and form a basis in space (any vector can be uniquely expanded into a linear combination of them).

(i) Show that

$$(14.24) \quad \|\nabla r\| = 1, \quad \|\nabla \theta\| = \frac{1}{r}, \quad \|\nabla z\| = 1,$$

$$(14.25) \quad \|\nabla \rho\| = 1, \quad \|\nabla \phi\| = \frac{1}{\rho}, \quad \|\nabla \theta\| = \frac{1}{\rho \sin \phi}$$

for the cylindrical (r, θ, z) and spherical (ρ, ϕ, θ) coordinates.

(ii) Show that the spherical and cylindrical coordinates are orthogonal coordinates and, in particular,

$$(14.26) \quad \hat{\mathbf{e}}_r = (\cos \theta, \sin \theta, 0), \quad \hat{\mathbf{e}}_\theta = (-\sin \theta, \cos \theta, 0), \quad \hat{\mathbf{e}}_z = (0, 0, 1)$$

for the cylindrical coordinates, and

$$(14.27) \quad \begin{aligned} \hat{\mathbf{e}}_r &= (\sin \phi \cos \theta, \sin \phi \sin \theta, \cos \phi), \\ \hat{\mathbf{e}}_\phi &= (\cos \phi \cos \theta, \cos \phi \sin \theta, -\sin \phi), \\ \hat{\mathbf{e}}_\theta &= (-\sin \theta, \cos \theta, 0) \end{aligned}$$

for the spherical coordinates.

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,

$$(14.28) \quad \int_0^1 \frac{dx}{x^\nu} = \lim_{a \rightarrow 0} \int_a^1 \frac{dx}{x^\nu} = \lim_{a \rightarrow 0} \frac{1 - a^{1-\nu}}{1-\nu} = \frac{1}{1-\nu}, \quad \nu < 1,$$

or

$$\int_0^\infty \frac{1}{1+x^2} dx = \lim_{a \rightarrow \infty} \int_0^a \frac{1}{1+x^2} dx = \lim_{a \rightarrow \infty} \tan^{-1} a = \frac{\pi}{2}.$$

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 $\mathbf{r} = (x, y) \in D$. For definiteness, the three-dimensional case is considered, while the two-dimensional case can be treated analogously. If, in any small ball 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 . If a closed bounded region E contains singular points of a function f , then the upper and lower sums cannot be defined because, for partition rectangles containing a singular point, $\sup f$ or $\inf f$ or both do not exist, and neither is defined a multiple integral of f .

Let B_ε be an open ball of radius ε centered at a point \mathbf{r}_0 . Suppose that the function f is singular at \mathbf{r}_0 . Define the region E_ε by removing all points of E that also lie in a ball B_ε . Suppose that f is integrable on E_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 is *defined* as the limit

$$(14.29) \quad \iiint_E f dV = \lim_{\varepsilon \rightarrow 0} \iiint_{E_\varepsilon} f dV \quad \text{or} \quad \iint_D f dA = \lim_{\varepsilon \rightarrow 0} \iint_{D_\varepsilon} f dA,$$

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 E_ε is obtained by removing S_ε from E . The regularization procedure in two dimensions is illustrated in Figure 14.34.

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

$$(14.30) \quad f(x, y) = \frac{y^2 - x^2}{(x^2 + y^2)^2}$$

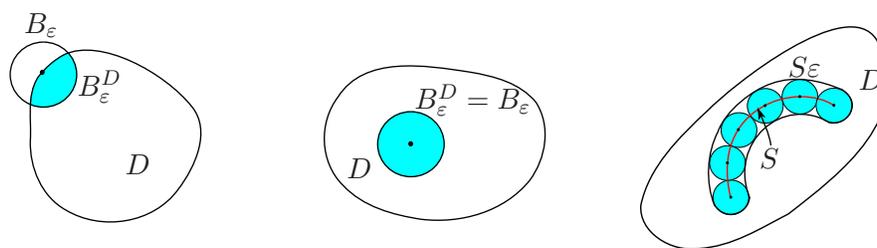


FIGURE 14.34. A regularization of an improper integral. **Left:** B_ε is a ball centered at a singular point of the integrand. B_ε^D is the intersection of B_ε with D . The integration is carried out over the region D with B_ε removed. Then the limit $\varepsilon \rightarrow 0$ is taken. **Middle:** The same regularization procedure when the singular point is an interior point of D . **Right:** A regularization procedure when singular points form a curve S . By removing the set S_ε from D , the region D_ε is obtained. The distance between any point of D_ε and the set S is no less than ε .

is to be integrated over the sector $0 \leq \theta \leq \theta_0$ of a disk $x^2 + y^2 \leq 1$, where θ is the polar angle. If the definition (14.29) is applied, then D_ε is the portion of the ring $\varepsilon^2 \leq x^2 + y^2 \leq 1$ corresponding to $0 \leq \theta \leq \theta_0$. 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 $\varepsilon \rightarrow 0$ does not exist for all θ_0 such that $\sin(2\theta_0) \neq 0$, 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, $(x, y) \rightarrow (y, x)$, $f(y, x) = -f(x, y)$, 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.

Furthermore, if D in the above example is the sector $x^2 + y^2 \leq 1$, $x \geq 0$, $y \geq 0$ or, in polar coordinates, $0 \leq r \leq 1$, $0 \leq \theta \leq \pi/2$, the improper integral could also be regularized by reducing the integration region to $\varepsilon \leq r \leq 1$, $0 \leq \theta \leq \theta_0$ with the subsequent *two-variable* limit $(\varepsilon, \theta_0) \rightarrow (0, \pi/2)$. Evidently, this limit does not exist. Recall that even though the two-variable limit does not exist, the limit along a particular curve may still exist. For example, if the limit $\theta_0 \rightarrow \pi/2$ is taken first, then the limit is 0, whereas the limit is infinite if the limit

$\varepsilon \rightarrow 0$ is taken first. This observation suggests that *the value of the improper integral may depend on the way a regularization is introduced.*

Integrability of Unbounded Functions. Let E be a region in space (possibly unbounded). An *exhaustion* of E is a sequence of bounded simple regions E_k , $k = 1, 2, \dots$, such that $E_1 \subset E_2 \subset \dots \subset E$ and the union of all E_k coincides with E . If the function f defined on E is singular at a limit point \mathbf{r}_0 of E , then one can construct an exhaustion of E such that none of E_k contain \mathbf{r}_0 . For example, one can take E_k to be the regions obtained from E by removing balls centered at \mathbf{r}_0 of radii $\varepsilon = 1/k$. It can be proved that *the sequence of volumes $V(E_k)$ converges to the volume of E .* Owing to the observation that the value of an improper integral may depend on the regularization, the following definition is adopted.

DEFINITION 14.17. (Integrability of an Unbounded Function).

Let E_k be an exhaustion of E . Suppose that a function f on E is integrable on each E_k . Then the function f is integrable on E if the limit $\lim_{k \rightarrow \infty} \iiint_{E_k} f \, dV$ exists and is independent of the choice of E_k . The value of the limit is called an improper integral of f over E .

An improper double integral is defined in the same way. The condition that the limit should not depend on the choice of an exhaustion means that the improper integral should not depend on its regularization. According to this definition, the function (14.30) is not integrable on any region containing the origin because the limit depends on the way the regularization is imposed. Although Definition 14.17 eliminates a potential ambiguity of the relation (14.29) noted above, it is rather difficult to use. A simplification useful in practice is achieved with the help of the concept of *absolute integrability*.

THEOREM 14.16. Let E_k and E'_k be two exhaustions of E . Let f be a function on E such that $|f|$ is integrable on each E_k and each E'_k . Then

$$\lim_{k \rightarrow \infty} \iiint_{E_k} |f| \, dV = \lim_{k \rightarrow \infty} \iiint_{E'_k} |f| \, dV,$$

where the limit may be $+\infty$.

In other words, *the value of the improper integral $\iiint_E |f| \, dV < \infty$, if it exists, is independent of the regularization.* The same statement holds for double integrals.

DEFINITION 14.18. (Absolute Integrability).

If the improper integral of the absolute value $|f|$ over E exists, then f is called absolutely integrable on E .

THEOREM 14.17. (Sufficient Condition for Integrability).

Let f be a continuous function on E . If f is absolutely integrable on E , then it is integrable on E .

This theorem implies that if the limit (14.29) exists for the absolute value $|f|$, then the improper integral of a continuous function f exists and can be calculated by the rule (14.29). The latter comprises a practical way to treat improper integrals.

EXAMPLE 14.30. 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 and continuous elsewhere. Let the restricted region E_ε lie between two spheres: $\varepsilon^2 \leq x^2 + y^2 + z^2 \leq R^2$. Since $|f| = f > 0$ in E , the convergence of the integral over E_ε as $\varepsilon \rightarrow 0$ also implies the absolute integrability of f and hence the existence of the improper integral (Theorem 14.17). By making use of the spherical coordinates, one obtains

$$\iiint_{E_\varepsilon} \frac{dV}{x^2 + y^2 + z^2} = \int_0^{2\pi} \int_0^\pi \int_\varepsilon^R \frac{\rho^2 \sin \theta}{\rho^2} d\rho d\phi d\theta = 4\pi(R - \varepsilon) \rightarrow 4\pi R$$

as $\varepsilon \rightarrow 0$. So the improper integral exists and equals $4\pi R$. \square

The following theorem is useful to assess the integrability.

THEOREM 14.18. (Absolute Integrability Test).

If $|f(\mathbf{r})| \leq g(\mathbf{r})$ for all \mathbf{r} in E and $g(\mathbf{r})$ is integrable on E , then f is absolutely integrable on E .

EXAMPLE 14.31. Investigate the 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 that is 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 $|x| \leq r$ and hence $|f| \leq r/r^\nu = r^{1-\nu} = g$. In the polar coordinates, the improper integral (14.29) of g over a disk of unit radius is

$$\int_0^{2\pi} d\theta \int_\varepsilon^1 g(r)r dr = 2\pi \int_\varepsilon^1 r^{2-\nu} dr = 2\pi \begin{cases} -\ln \varepsilon, & \nu = 3 \\ 1 - \frac{\varepsilon^{3-\nu}}{3-\nu}, & \nu \neq 3 \end{cases}.$$

The limit $\varepsilon \rightarrow 0$ is finite if $\nu < 3$. By the integrability test (Theorem 14.18), the function f is absolutely integrable if $\nu < 3$. For $\nu < 3$ and D being the part of the unit disk in the first quadrant, one infers that

$$\lim_{\varepsilon \rightarrow 0} \iint_{D_\varepsilon} f \, dA = \lim_{\varepsilon \rightarrow 0} \int_0^{\pi/2} \int_\varepsilon^1 \frac{r \cos \theta}{r^\nu} r \, dr \, d\theta = \lim_{\varepsilon \rightarrow 0} \int_\varepsilon^1 r^{2-\nu} \, dr = 1.$$

□

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 14.19. *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 $|f(\mathbf{r})| \leq M \|\mathbf{r} - \mathbf{r}_0\|^{-\nu}$ for all \mathbf{r} in D such that $0 < \|\mathbf{r} - \mathbf{r}_0\| < R$ 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 $\mathbf{r} \rightarrow \mathbf{r} - \mathbf{r}_0$. 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 (compare with Figure 14.34 with $\varepsilon = R$). For $n = 1$, the integrability follows from (14.28). In the two-variable case, the use of the polar coordinates yields $dA = r \, dr \, d\theta$, $\|\mathbf{r}\| = r$, and

$$\iint_{B_R^D} |f| \, dA \leq M \iint_{B_R^D} \frac{dA}{\|\mathbf{r}\|^\nu} \leq M \iint_{B_R} \frac{dA}{\|\mathbf{r}\|^\nu} = 2\pi M \int_0^R \frac{dr}{r^{\nu-1}},$$

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. Iterated Integrals and Integrability. If a function is not integrable, its iterated integrals may still be well defined. However, the value of the iterated integral depends on the order of integration, and, in particular, *Fubini's theorem does not hold*. For example, consider the func-

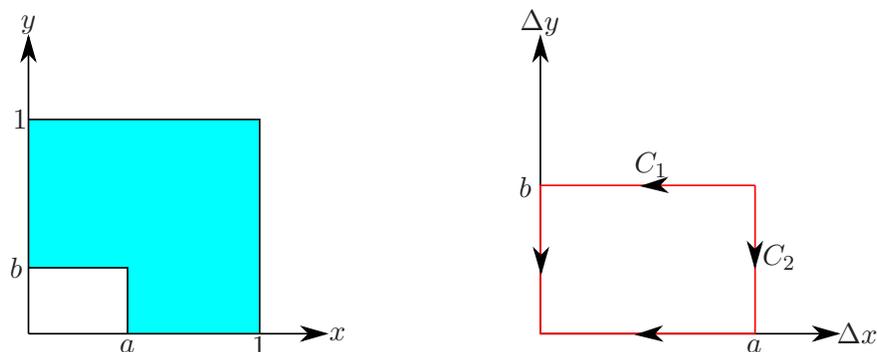


FIGURE 14.35. **Left:** The rectangle $D = [0, 1] \times [0, 1]$ with one partition rectangle $[0, a] \times [0, b]$ removed, where $\Delta x = a$ and $\Delta y = b$. In the limit $(\Delta x, \Delta y) \rightarrow (0, 0)$, the region becomes the rectangle $[0, 1] \times [0, 1]$. **Right:** The limit can be taken along two particular paths C_1 and C_2 . In the former case (C_1), Δx is taken to 0 first. The Riemann sum becomes an iterated integral in which the integration with respect to x is carried out first. When Δy is taken to 0 first, then the Riemann sum becomes an integrated integral in which the integration with respect to y is carried first.

tion (14.30) over the rectangle $D = [0, 1] \times [0, 1]$. It is not integrable as argued. On the other hand, consider a rectangular partition of D where each partition rectangle has the area $\Delta x \Delta y$. The Riemann sum is regularized by removing the rectangle $[0, \Delta x] \times [0, \Delta y]$ as shown in Figure 14.35. The limit of the Riemann sum is the *two-variable* limit $(\Delta x, \Delta y) \rightarrow (0, 0)$. By taking first $\Delta y \rightarrow 0$ and then $\Delta x \rightarrow 0$, one obtains an iterated integral in which the integration with respect to y is carried out first:

$$\begin{aligned} \lim_{a \rightarrow 0} \int_a^1 \lim_{b \rightarrow 0} \int_b^1 \frac{x^2 - y^2}{(x^2 + y^2)^2} dy dx &= \lim_{a \rightarrow 0} \int_a^1 \lim_{b \rightarrow 0} \int_b^1 \frac{\partial}{\partial y} \frac{y}{x^2 + y^2} dy dx \\ &= \lim_{a \rightarrow 0} \int_a^1 \lim_{b \rightarrow 0} \left(\frac{1}{1 + x^2} - \frac{b}{x^2 + b^2} \right) dx \\ &= \lim_{a \rightarrow 0} \int_a^1 \frac{dx}{1 + x^2} = \int_0^1 \frac{dx}{1 + x^2} = \frac{\pi}{4}. \end{aligned}$$

Here $(a, b) = (\Delta x, \Delta y)$. Alternatively, the limit $\Delta x \rightarrow 0$ can be taken first and then $\Delta y \rightarrow 0$, which results in the iterated integral in the reverse order:

$$\begin{aligned}
\lim_{b \rightarrow 0} \int_b^1 \lim_{a \rightarrow 0} \int_a^1 \frac{x^2 - y^2}{(x^2 + y^2)^2} dx dy &= - \lim_{b \rightarrow 0} \int_b^1 \lim_{a \rightarrow 0} \int_a^1 \frac{\partial}{\partial x} \frac{x}{x^2 + y^2} dy dx \\
&= - \lim_{b \rightarrow 0} \int_b^1 \lim_{a \rightarrow 0} \left(\frac{1}{1 + y^2} - \frac{a}{y^2 + a^2} \right) dy \\
&= - \lim_{b \rightarrow 0} \int_b^1 \frac{dy}{1 + y^2} \\
&= - \int_0^1 \frac{dy}{1 + y^2} = -\frac{\pi}{4}.
\end{aligned}$$

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 (the function is not integrable).

106.3. Multiple Integrals Over Unbounded Regions. The treatment of multiple integrals over unbounded regions follows the same steps introduced when discussing the integrability of unbounded functions.

DEFINITION 14.19. *Let E be an unbounded region and let E_k be an exhaustion of E where each E_k is bounded. Suppose that f is integrable on each E_k . Then*

$$\iiint_E f dV = \lim_{k \rightarrow \infty} \iiint_{E_k} f dV$$

if the limit exists and is independent of the choice of E_k .

Double integrals over unbounded regions are defined in the same way. Theorems 14.16, 14.17, and 14.18 hold for unbounded regions.

The following practical approach may be used to evaluate improper integrals over unbounded regions. 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. Let f be a continuous function that is absolutely integrable on E . The integral of f over D (or E) is evaluated by the rule

$$\iint_D f(\mathbf{r}) dA = \lim_{R \rightarrow \infty} \iint_{D_R} f(\mathbf{r}) dA$$

or

$$\iiint_E f(\mathbf{r}) dV = \lim_{R \rightarrow \infty} \iiint_{E_R} f(\mathbf{r}) dV$$

The absolute integrability of f means that these limits exist and are finite for the absolute value $|f|$. The asymptotic behavior of a function sufficient for absolute integrability on an unbounded region is stated in the following theorem, which is an analog of Theorem 14.19.

THEOREM 14.20. *Suppose f is a continuous function on an unbounded region D of a Euclidean space such that $|f(\mathbf{r})| \leq M\|\mathbf{r}\|^{-\nu}$ for all $\|\mathbf{r}\| \geq R$ in D and some $R > 0$ and $M \geq 0$. 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:

$$\int_R^\infty \frac{dx}{x^\nu} = \lim_{a \rightarrow \infty} \int_R^a \frac{dx}{x^\nu} = \lim_{a \rightarrow \infty} \left. \frac{x^{1-\nu}}{1-\nu} \right|_R^a = -\frac{R^{1-\nu}}{1-\nu} + \lim_{a \rightarrow \infty} \frac{a^{1-\nu}}{1-\nu}$$

if $\nu \neq 1$. 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'_R be the part of the space outside B_R (see Figure 14.36, left panel). Note that B'_R includes D'_R . In the two-variable case, the use of the polar coordinates gives

$$\begin{aligned} \iint_{D'_R} |f| dA &\leq \iint_{B'_R} |f| dA \leq \iint_{B'_R} \frac{M dA}{\|\mathbf{r}\|^\nu} = M \int_0^{2\pi} d\theta \int_R^\infty \frac{r dr}{r^\nu} \\ &= 2\pi M \int_R^\infty \frac{dr}{r^{\nu-1}}, \end{aligned}$$

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 $0 \leq \phi \leq \pi$ and $0 \leq \theta \leq 2\pi$ for the region B'_R so that

$$\begin{aligned} \iiint_{D'_R} |f| dV &\leq \iiint_{B'_R} |f| dV \leq \iiint_{B'_R} \frac{M dV}{\|\mathbf{r}\|^\nu} = 4\pi M \int_R^\infty \frac{\rho^2 d\rho}{\rho^\nu} \\ &= 4\pi M \int_R^\infty \frac{d\rho}{\rho^{\nu-2}}, \end{aligned}$$

which converges if $\nu > 3$. □

EXAMPLE 14.32. *Evaluate the double integral of $f(x, y) = \exp(-x^2 - y^2)$ over the entire plane.*

SOLUTION: In polar coordinates, $|f| = e^{-r^2}$. So, as $r \rightarrow \infty$, $|f|$ decreases faster than any inverse power r^{-n} , $n > 0$, and by virtue of Theorem 14.20, f is absolutely integrable on the plane. By making use

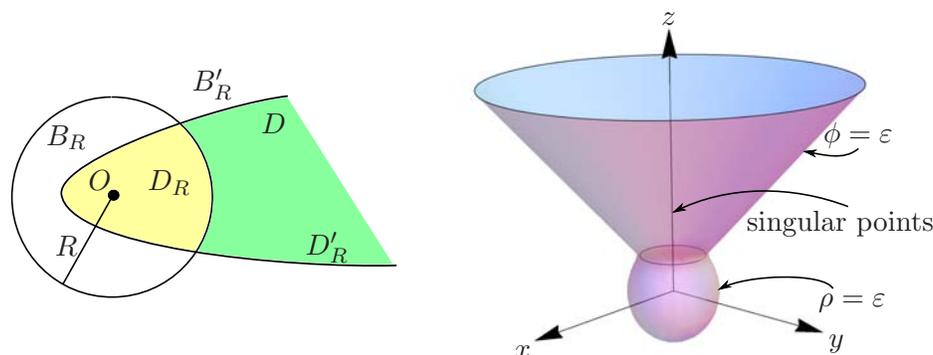


FIGURE 14.36. **Left:** An unbounded region D is split into two parts: D_R lies inside the ball B_R of radius R , and D'_R is the part of D that lies outside the ball B_R . The region B'_R is the entire space with the ball B_R removed. The region D'_R is contained in B'_R . **Right:** A regularization procedure for the integral in Study Problem 14.9. The integration region E contains singular points along the z axis. The integral is regularized by removing the ball $\rho < \varepsilon$ and the solid cone $\phi < \varepsilon$ from E . After the evaluation of the integral, the limit $\varepsilon \rightarrow 0$ is taken.

of the polar coordinates,

$$\begin{aligned} \iint_D e^{-x^2-y^2} dA &= \lim_{R \rightarrow \infty} \int_0^{2\pi} \int_0^R e^{-r^2} r dr d\theta = \pi \lim_{R \rightarrow \infty} \int_0^{R^2} e^{-u} du \\ &= \pi \lim_{R \rightarrow \infty} (1 - e^{-R^2}) = \pi, \end{aligned}$$

where the substitution $u = r^2$ has been made. \square

It is interesting to observe the following. As the function is absolutely integrable, the double integral can also be evaluated by Fubini's theorem in rectangular coordinates:

$$\begin{aligned} \iint_D e^{-x^2-y^2} dA &= \int_{-\infty}^{\infty} e^{-x^2} dx \int_{-\infty}^{\infty} e^{-y^2} dy = I^2, \\ I &= \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi} \end{aligned}$$

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 functions.

106.4. Study Problem.

Problem 14.9. 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 solid cone $\phi < \varepsilon$ and a ball $\rho < \varepsilon$, 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 coordinates is a rectangle (i.e., it is bounded). Hence,

$$\lim_{\varepsilon \rightarrow 0} \iiint_{E_\varepsilon} f dV = \lim_{\varepsilon \rightarrow 0} \iiint_{E'_\varepsilon} d\rho d\phi d\theta = \int_0^{2\pi} d\theta \int_0^{\pi/4} d\phi \int_0^1 d\rho = \frac{\pi^2}{2}.$$

So the Jacobian cancels out all the singularities of the function. \square

106.5. Exercises.

(1) Let the function $g(x, y)$ be bounded so that $0 < m \leq g(x, y) \leq M$ for all (x, y) . Investigate the convergence of the following double integrals:

- (i) $\iint_D g(x, y)(x^2 + y^2)^{-1} dA$, where D is defined by the conditions $|y| \leq x^2$, $x^2 + y^2 \leq 1$
- (ii) $\iint_D g(x, y)(|x|^p + |y|^q)^{-1} dA$, $p > 0$, $q > 0$, where D is defined by the condition $|x| + |y| \leq 1$
- (iii) $\iint_D g(x, y)(1 - x^2 - y^2)^{-p} dA$, where D is defined by the condition $x^2 + y^2 \leq 1$
- (iv) $\iint_D g(x, y)|x - y|^{-p} dA$, where D is the square $[0, a] \times [0, a]$
- (v) $\iint_D e^{-(x+y)} dA$, where D is defined by $0 \leq x \leq y$

(2) Let the function $g(x, y, z)$ be bounded so that $0 < m \leq g(x, y, z) \leq M$ for all (x, y, z) . Investigate the convergence of the following triple integrals:

- (i) $\iiint_E g(x, y, z)(x^2 + y^2 + z^2)^{-\nu} dV$, where E is defined by $x^2 + y^2 + z^2 \geq 1$
- (ii) $\iiint_E g(x, y, z)(x^2 + y^2 + z^2)^{-\nu} dV$, where E is defined by $x^2 + y^2 + z^2 \leq 1$

- (iii) $\iiint_E g(x, y, z)(|x|^p + |y|^q + |z|^s)^{-1} dV$, where p , q , and s are positive numbers and E is defined by $|x| + |y| + |z| \geq 1$
- (iv) $\iiint_E g(x, y, z)|x + y - z|^{-\nu} dV$, where $E = [-1, 1] \times [-1, 1] \times [-1, 1]$

(3) Evaluate the improper integral if it exists. Use appropriate coordinates when needed.

- (i) $\iiint_E (x^2 + y^2 + z^2)^{-1/2} (x^2 + y^2)^{-1/2} dV$, where E is the region in the first octant bounded from above by the sphere $x^2 + y^2 + z^2 = 2z$ and from below by the cone $z = \sqrt{3}\sqrt{x^2 + y^2}$
- (ii) $\iiint_E z(x^2 + y^2)^{-1/2} dV$, where E is in the first octant and bounded from above by the cone $z = 2 - \sqrt{x^2 + y^2}$ and from below by the paraboloid $z = x^2 + y^2$
- (iii) $\iiint_E xy(x^2 + y^2)^{-1} (x^2 + y^2 + z^2)^{-1} dV$, where E is the portion of the ball $x^2 + y^2 + z^2 \leq a^2$ above the plane $z = 0$
- (iv) $\iiint_E e^{-x^2 - y^2 - z^2} (x^2 + y^2 + z^2)^{-1/2} dV$, where E is the entire space
- (v) $\iint_D (x^2 + y^2)^{-1/2} dA$, where D lies between the two circles $x^2 + y^2 = 4$ and $(x - 1)^2 + y^2 = 1$ in the first quadrant, $x, y \geq 0$
- (vi) $\iint_D \ln(x^2 + y^2) dA$, where D is the disk $x^2 + y^2 \leq a^2$
- (vii) $\iiint_E (x^2 + y^2 + z^2)^\nu \ln(x^2 + y^2 + z^2) dV$, where E is the ball $x^2 + y^2 + z^2 \leq a^2$ and ν is real. Does the integral exist for all ν ?
- (viii) $\iint_D (x^2 + y^2)^\nu \ln(x^2 + y^2) dA$, where D is defined by $x^2 + y^2 \geq a^2 > 0$ and ν is real. Does the integral exist for all ν ?
- (ix) $\iiint_E (x^2 + y^2 + z^2)^\nu \ln(x^2 + y^2 + z^2) dV$, where E is defined by $x^2 + y^2 + z^2 \geq a^2 > 0$ and ν is real. Does the integral exist for all ν ?
- (x) $\iint_D [(a - x)(x - y)]^{-1/2} dA$, where D is the triangle bounded by the lines $y = 0$, $y = x$, and $x = a$
- (xi) $\iint_D \ln \sin(x - y) dA$, where D is bounded by the lines $y = 0$, $y = x$, and $x = \pi$
- (xii) $\iint_D (x^2 + y^2)^{-1} dA$, where D is defined by $x^2 + y^2 \leq x$
- (xiii) $\iiint_E x^{-p} y^{-q} z^{-s} dV$, where $E = [0, 1] \times [0, 1] \times [0, 1]$
- (xiv) $\iiint_E (x^2 + y^2 + z^2)^{-3} dV$, where E is defined by $x^2 + y^2 + z^2 \geq 1$
- (xv) $\iiint_E (1 - x^2 - y^2 - z^2)^{-\nu} dV$, where E is defined by $x^2 + y^2 + z^2 \leq 1$
- (xvi) $\iiint_E e^{-x^2 - y^2 - z^2} dV$, where E is the entire space
- (xvii) $\iint_D e^{-x^2 - y^2} \sin(x^2 + y^2) dA$, where D is the entire plane
- (xviii) $\iint_D e^{-(x/a)^2 - (y/b)^2} dA$, where D is the entire plane
- (xix) $\iint_D e^{ax^2 + 2bxy + cy^2} dA$, where $a < 0$, $ac - b^2 > 0$, and D is the entire plane (*Hint*: Find a rotation that transforms x and y so

that in the new variables the bilinear term “ xy ” is absent in the exponential.)

(xx) $\iiint_E e^{-(x/a)^2 - (y/b)^2 - (z/c)^2 + \alpha x + \beta y + \gamma z} dV$, where E is the entire space

(4) Let n be an integer. Show that

$$\lim_{n \rightarrow \infty} \iint_{D_n} \sin(x^2 + y^2) dA = \pi, \quad D_n: |x| \leq n, |y| \leq n,$$

$$\lim_{n \rightarrow \infty} \iint_{D_n} \sin(x^2 + y^2) dA = 0, \quad D_n: x^2 + y^2 \leq 2\pi n.$$

Note that in each case D_n covers the entire plane as $n \rightarrow \infty$. What can be said about the convergence of the integral over the entire plane?

(5) Show that the integral $\iint_D (x^2 - y^2)(x^2 + y^2)^{-2} dA$, where D is defined by $x \geq 1, y \geq 1$, diverges, whereas the iterated integrals in both orders converge.

(6) Show that the following improper integrals converge. Use the geometric series to show that their values are given by the specified convergent series:

(i) $\lim_{a \rightarrow 1^-} \iint_{D_a} (1 - xy)^{-1} dA = \sum_{n=1}^{\infty} \frac{1}{n^2}$, where $D_a = [0, a] \times [0, a]$

(ii) $\lim_{a \rightarrow 1^-} \iiint_{E_a} (1 - xyz)^{-1} dV = \sum_{n=1}^{\infty} \frac{1}{n^3}$, where $E_a = [0, a] \times [0, a] \times [0, a]$

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}):

$$\sigma(\mathbf{r}) = \lim_{\Delta s \rightarrow 0} \frac{\Delta m(\mathbf{r})}{\Delta s}.$$

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 p th segment, then the total mass reads

$$M = \lim_{N \rightarrow \infty} \sum_{p=1}^N \sigma(\mathbf{r}_p^*) \Delta s,$$

where the mass of the p th 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 $\Delta s \rightarrow 0$. 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 , $p = 1, 2, \dots, N$, 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 14.20. (Line Integral of a Function).

The line integral of a function f along a piecewise-smooth curve C is

$$\int_C f(\mathbf{r}) ds = \lim_{N \rightarrow \infty} U(f, N) = \lim_{N \rightarrow \infty} L(f, N),$$

provided the limits of the upper and lower sums exist and coincide. The limit is understood in the sense that $\max \Delta s_p \rightarrow 0$ as $N \rightarrow \infty$ (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:

$$\int_C f(\mathbf{r}) ds = \lim_{N \rightarrow \infty} \sum_{p=1}^N f(\mathbf{r}_p^*) \Delta s_p = \lim_{N \rightarrow \infty} R(f, N).$$

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 by the squeeze principle the limit of the Riemann sum is *independent* of the choice of sample points \mathbf{r}_p^* (see the left panel of Figure 14.37).

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 that contains a smooth curve C . Let E_a be a *neighborhood* of C that is defined as the set of points whose distance (in the sense of Definition 11.14) to C cannot exceed $a > 0$. So, for a small enough, the cross section of E_a by a plane normal to C is a disk of radius a whose area is $\Delta A = \pi a^2$ (see the right panel of Figure 14.37). Then, in the limit $a \rightarrow 0$,

$$(14.31) \quad \frac{1}{\pi a^2} \iiint_{E_a} f(\mathbf{r}) dV \rightarrow \int_C f(\mathbf{r}) ds.$$

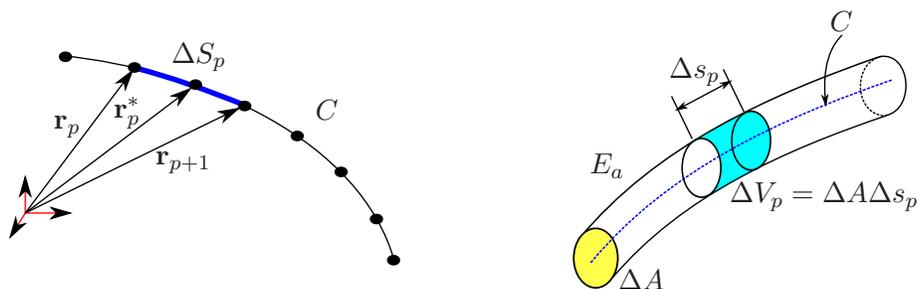


FIGURE 14.37. **Left:** A partition of a smooth curve C by segments of arc length Δs_p used in the definition of the line integral and its Riemann sum. **Right:** The region E_a is a neighborhood of a smooth curve C . It consists of points whose distance to C cannot exceed $a > 0$ (recall Definition 11.14). For a and Δs_p small enough, planes normal to C through the points \mathbf{r}_p partition E_a into elements whose volume is $\Delta V_p = \Delta A \Delta s_p$, where $\Delta A = \pi a^2$ is the area of the cross section of E_a . This partition is used to establish the relation (14.31) between the triple and line integrals.

In other words, line integrals can be viewed as the limiting case of triple integrals when two dimensions of the integration region become infinitesimally small. This follows from (14.13) by taking a partition of E_a by volume elements $\Delta V_p = \Delta A \Delta s_p$ and sample points along the curve C in E_a . In the Riemann sum for the left side of (14.31), the factor $\Delta A = \pi a^2$ in ΔV_p cancels the same factor in the denominator so that the Riemann sum becomes a Riemann sum for the line integral on the right side of (14.31). 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 14.21. (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

$$(14.32) \quad \int_C f(\mathbf{r}) ds = \int_a^b f(\mathbf{r}(t)) \|\mathbf{r}'(t)\| dt.$$

PROOF. Consider a partition of $[a, b]$, $t_p = a + p \Delta t$, $p = 0, 1, 2, \dots, N$, 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 $t \in [t_{p-1}, t_p]$, $p = 1, 2, \dots, N$. 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,

$$\int_C f ds = \lim_{N \rightarrow \infty} \sum_{p=1}^N f(\mathbf{r}(t_p^*)) \|\mathbf{r}'(t_p^*)\| \Delta t = \int_a^b f(\mathbf{r}(t)) \|\mathbf{r}'(t)\| dt.$$

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 $[a, b]$. \square

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 $\|\mathbf{r}'(t)\|$. 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 (14.32).

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 (14.32) 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 14.33. 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 $(0, a)$). The 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,

$$\int_C x^2 y \, ds = \int_0^{2\pi} (R \cos t)^2 (a + R \sin t) R \, dt = R^2 a \int_0^{2\pi} \cos^2 t \, dt = \pi R^2 a,$$

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 $\cos^2 t = (1 + \cos(2t))/2$. \square

EXAMPLE 14.34. 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 $x + y + z = 0$.

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,

$$\int_C f \, ds = \int_0^{2\pi} \sqrt{2 - \sin(2t)} \sqrt{2 - \sin(2t)} \, dt = \int_0^{2\pi} (2 - \sin(2t)) \, dt = 4\pi.$$

\square

107.3. Exercises.

(1) Evaluate the line integral:

- (i) $\int_C xy^2 \, ds$, where C is the right half of the circle $x^2 + y^2 = 4$
- (ii) $\int_C x \sin y \, ds$, where C is the line segment from $(0, a)$ to $(b, 0)$
- (iii) $\int_C xyz \, ds$, where C is the helix $x = 2 \cos t$, $y = t$, $z = -2 \sin t$, $0 \leq t \leq \pi$
- (1) (iv) $\int_C (2x + 9z) \, ds$, where C is the curve $x = t$, $y = t^2$, $z = t^3$ from $(0, 0, 0)$ to $(1, 1, 1)$

- (v) $\int_C z \, ds$, where C is the intersection of the paraboloid $z = x^2 + y^2$ and the plane $z = 4$
- (vi) $\int_C y \, ds$, where C is the part of the graph $y = e^x$ for $0 \leq x \leq 1$
- (vii) $\int_C (x + y) \, ds$, where C is the triangle with vertices $(0, 0)$, $(1, 0)$, and $(0, 1)$
- (viii) $\int_C y^2 \, ds$, where C is an arc of the cycloid $x = R(t - \sin t)$, $y = R(1 - \cos t)$ from $(0, 0)$ to $(2\pi R, 0)$
- (ix) $\int_C xy \, ds$, where C is an arc of the hyperbola $x = a \cosh t$, $y = a \sinh t$ for $0 \leq t \leq T$
- (x) $\int_C (x^{4/3} + y^{4/3}) \, ds$, where C is the astroid $x^{2/3} + y^{2/3} = a^{2/3}$
- (xi) $\int_C x \, ds$, where C is the part of the spiral $r = ae^\theta$ that lies in the disk $r \leq a$; here (r, θ) are polar coordinates
- (xii) $\int_C \sqrt{x^2 + y^2} \, ds$, where C is the circle $x^2 + y^2 = ax$
- (xiii) $\int_C y^{-2} \, ds$, where C is $y = a \cosh(x/a)$
- (xiv) $\int_C (x^2 + y^2 + z^2) \, ds$, where C is one turn of the helix $x = R \cos t$, $y = R \sin t$, $z = ht$ ($0 \leq t \leq 2\pi$)
- (xv) $\int_C y^2 \, ds$, where C is the circle $x^2 + y^2 + z^2 = 1$, $x + y + z = 0$
- (xvi) $\int_C z \, ds$, where C is the conic helix $x = t \cos t$, $y = t \sin t$, $z = t$ ($0 \leq t \leq T$)
- (xvii) $\int_C z \, ds$, where C is the curve of intersection of the surfaces $x^2 + y^2 = z^2$ and $y^2 = ax$ from the origin to the point $(a, a, a\sqrt{2})$
- (2) Find the mass of an arc of the parabola $y^2 = 2ax$, $0 \leq x \leq a/2$, if its linear mass density is $\sigma(x, y) = |y|$.
- (3) Find the mass of the curve $x = at$, $y = at^2/2$, $z = at^2/3$, where $0 \leq t \leq 1$, if its linear mass density is $\sigma(x, y, z) = \sqrt{2y/a}$.

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, $x^2 + y^2 \leq 1$. 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} . Suppose that f has continuous partial derivatives, and hence its linearization at a point in D defines a tangent plane to the graph.

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 deviation of f from its linearization tends to 0 faster than $\sqrt{(\Delta x)^2 + (\Delta y)^2}$ as $(\Delta x, \Delta y) \rightarrow (0, 0)$. Therefore, in this limit, Δx and Δy can be viewed as the differentials dx and dy , and, when calculating ΔS_{ij} , only terms linear in dx and dy must be retained. Therefore, the areas ΔS_{ij} and $\Delta A = \Delta x \Delta y$ have to be proportional:

$$\Delta S_{ij} = J_{ij} \Delta A.$$

The coefficient J_{ij} 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} (see the left panel of Figure 14.38). 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 $\mathbf{a} = \overrightarrow{OA}$ and $\mathbf{b} = \overrightarrow{OB}$ 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 $\mathbf{a} = (\Delta x, 0, f'_x(x_i^*, y_j^*) \Delta x)$. Similarly, $\mathbf{b} = (0, \Delta y, f'_y(x_i^*, y_j^*) \Delta y)$. Hence,

$$\begin{aligned} \mathbf{a} \times \mathbf{b} &= (-f'_x, -f'_y, 1) \Delta x \Delta y, \\ \Delta S_{ij} = \|\mathbf{a} \times \mathbf{b}\| &= \sqrt{1 + (f'_x)^2 + (f'_y)^2} \Delta A = J(x_i^*, y_j^*) \Delta A, \end{aligned}$$

where $J(x, y) = \sqrt{1 + (f'_x)^2 + (f'_y)^2}$. Thus, the surface area is given by

$$A(S) = \lim_{(\Delta x, \Delta y) \rightarrow (0, 0)} \sum_{ij} J(x_i^*, y_j^*) \Delta A.$$

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 .

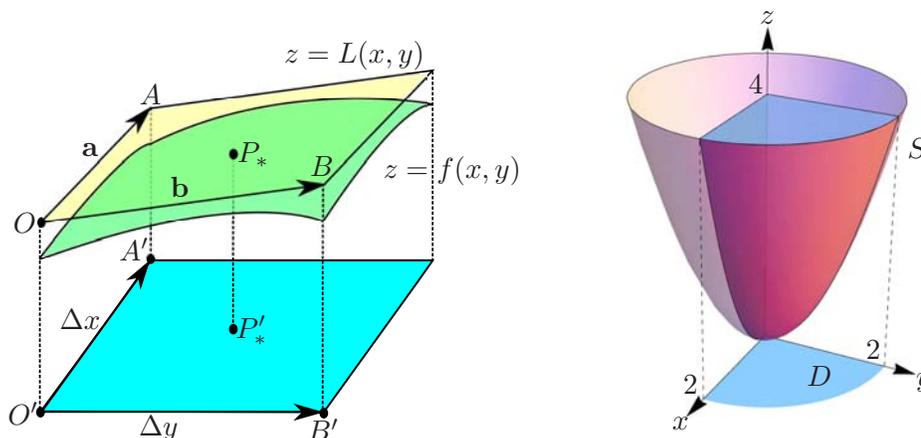


FIGURE 14.38. **Left:** The rectangle with adjacent sides $O'A'$ and $O'B'$ is an element of a rectangular partition of D and P'_* is a sample point. The point P_* is the point on the graph $z = f(x, y)$ for $(x, y) = P'_*$. The linearization of f at P_* defines the tangent plane $z = L(x, y)$ to the graph through P_* . The surface of the portion of the graph above the partition rectangle is approximated by the area of the portion of the tangent plane above the partition rectangle, which is the area of the parallelogram with adjacent sides OA and OB . It equals $\|\mathbf{a} \times \mathbf{b}\|$. **Right:** An illustration to Example 14.36. The part of the paraboloid whose area is to be evaluated is obtained by restricting (x, y) to the part D of the disk of radius 2 that lies in the first quadrant.

DEFINITION 14.21. (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 14.35. Show 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 $x^2 + y^2 \leq R^2$ 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,

$$\begin{aligned} A(S) &= 2R \iint_D \frac{dA}{\sqrt{R^2 - x^2 - y^2}} = 2R \int_0^{2\pi} d\theta \int_0^R \frac{r dr}{\sqrt{R^2 - r^2}} \\ &= 4\pi R \int_0^R \frac{r dr}{\sqrt{R^2 - r^2}} = 2\pi R \int_0^{R^2} \frac{du}{\sqrt{u}} = 4\pi R^2, \end{aligned}$$

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. \square

EXAMPLE 14.36. 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. \square

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 are 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 an open region 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 , $p = 1, 2, \dots, N$, which have surface area ΔS_p . Suppose that S is defined as a level surface $g(x, y, z) = k$ of a function g that has continuous partial derivatives on E and whose gradient does not vanish. It was shown in Section 93.2 that given a point P on S there is a function of two variables whose graph coincides with S in a neighborhood of P . So the surface area ΔS_p of a sufficiently small partition element S_p can be found by Definition 14.21. 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 partition elements become *uniformly* smaller.

DEFINITION 14.22. (Surface Integral of a Function).

The surface integral of a bounded function f over a surface S is

$$\iint_S f(\mathbf{r}) dS = \lim_{N \rightarrow \infty} U(f, N) = \lim_{N \rightarrow \infty} L(f, N),$$

provided the limits of the upper and lower sums exist and coincide. The limit is understood in the sense $R_N \rightarrow 0$ as $N \rightarrow \infty$.

The surface integral can also be represented by the limit of a Riemann sum:

$$(14.33) \quad \iint_S f(\mathbf{r}) dS = \lim_{N \rightarrow \infty} \sum_{p=1}^N f(\mathbf{r}_p^*) \Delta S_p = \lim_{N \rightarrow \infty} R(f, N).$$

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 by the

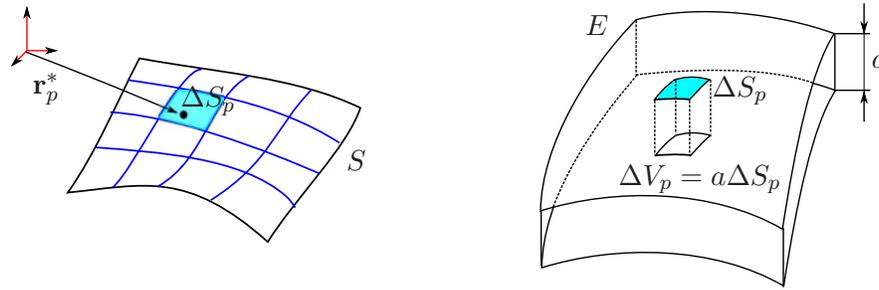


FIGURE 14.39. **Left:** A partition of a surface S by elements with surface area ΔS_p . It is used in the definition of the surface integral and also to construct its Riemann sums. **Right:** A neighborhood E_a of a smooth surface S defined as the set of points whose distance to S cannot exceed $a > 0$. For sufficiently fine partition of S and small a , the region E_a is partitioned by elements of volume $\Delta V_p = a \Delta S_p$.

squeeze principle 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. Consider a neighborhood E_a of a smooth surface S that is defined as the set of points whose distance to S cannot exceed $a/2 > 0$ (in the sense of Definition 11.14). The region E_a looks like a shell with thickness a (see the right panel of Figure 14.39). Suppose that f is integrable on E_a . Then, in the limit $a \rightarrow 0$,

$$(14.34) \quad \frac{1}{a} \iiint_{E_a} f(\mathbf{r}) dV \rightarrow \iint_S f(\mathbf{r}) dS.$$

This relation can be understood by considering the Riemann sum (14.13) for the triple integral in (14.34) in which E_a is partitioned by volume elements $\Delta V_p = a \Delta S_p$ with sample points taken on S . Partition elements are cylinders of height a along the normal to the surface and with the area of the cross section ΔS_p defined by the tangent plane. The factor $1/a$ on the right side of (14.34) cancels the common factor a in ΔV_p , and the Riemann sum turns into a Riemann sum for a surface integral. 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 14.22. (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 an open region that contains D . Then

$$(14.35) \quad \iint_S f(x, y, z) \, dS = \iint_D f(x, y, g(x, y)) \sqrt{1 + (g'_x)^2 + (g'_y)^2} \, dA.$$

Consider a partition of D by elements D_p of area ΔA_p , $p = 1, 2, \dots, N$. 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 (14.33), 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 14.22, 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 (14.35).

EXAMPLE 14.37. 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

$$\begin{aligned} \iint_S z \, dS &= \iint_D xy \sqrt{1 + x^2 + y^2} \, dA \\ &= \int_0^{\pi/2} \cos \theta \sin \theta \, d\theta \int_0^1 r^2 \sqrt{1 + r^2} \, r \, dr \\ &= \frac{\sin^2 \theta}{2} \Big|_0^{\pi/2} \frac{1}{2} \int_1^2 (u - 1) \sqrt{u} \, du \\ &= \frac{1}{2} \left(\frac{u^{5/2}}{5} - \frac{u^{3/2}}{3} \right) \Big|_1^2 = \frac{2(4\sqrt{2} + 1)}{15}, \end{aligned}$$

where the double integral has been converted to polar coordinates and the last integral is evaluated by the substitution $u = 1 + r^2$. \square

108.4. Parametric Equations of a Surface. The graph $z = g(x, y)$ of a continuous function g , 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 . For every pair (u, v) , the rule $\mathbf{r}(u, v) = (u, v, g(u, v))$ defines a vector in space, which is the position vector of a point on the surface. One can make a continuous transformation from D to D' by changing variables $(u, v) \rightarrow (u', v')$. Then the components of position vectors of points of S become general continuous functions of the new variables (u', v') . This observation suggests that a surface in space can be defined by specifying three continuous functions of two variables, $x(u, v)$, $y(u, v)$, and $z(u, v)$, on a region D that are viewed as components of the position vector $\mathbf{r}(u, v) = (x(u, v), y(u, v), z(u, v))$. A mapping of D into space defined by this rule is called a *vector function* on D . The range of this mapping is called a *parametric surface in space*, and the equations $x = x(u, v)$, $y = y(u, v)$, and $z = z(u, v)$ are called *parametric equations* of the surface.

For example, the equations

$$(14.36) \quad x = R \cos v \sin u, \quad y = R \sin v \sin u, \quad z = R \cos u$$

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 to be represented as the union of graphs. For example, the whole sphere is described by the single vector-valued

function (14.36) of two variables instead of the union of two graphs $z = \pm\sqrt{R^2 - x^2 - y^2}$.

DEFINITION 14.23. 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 $\mathbf{0}$.

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 14.38. Find the parametric equations of the double cone $z^2 = x^2 + y^2$.

SOLUTION: Suppose $z \neq 0$. Then $(x/z)^2 + (y/z)^2 = 1$. A 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 $t = t(u, v)$. \square

EXAMPLE 14.39. 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 $v \in [0, 2\pi]$. Since all points $(x_0, 0, z_0)$ are on the circle $z^2 + (x - a)^2 = R^2$, they can be parameterized as $x_0 - a = R \cos u$, $z_0 = R \sin u$, where $u \in [0, 2\pi]$. Thus, the parametric equations of a torus are

(14.37)

$$x = (R + a \cos u) \cos v, \quad y = (R + a \cos u) \sin v, \quad z = R \sin u,$$

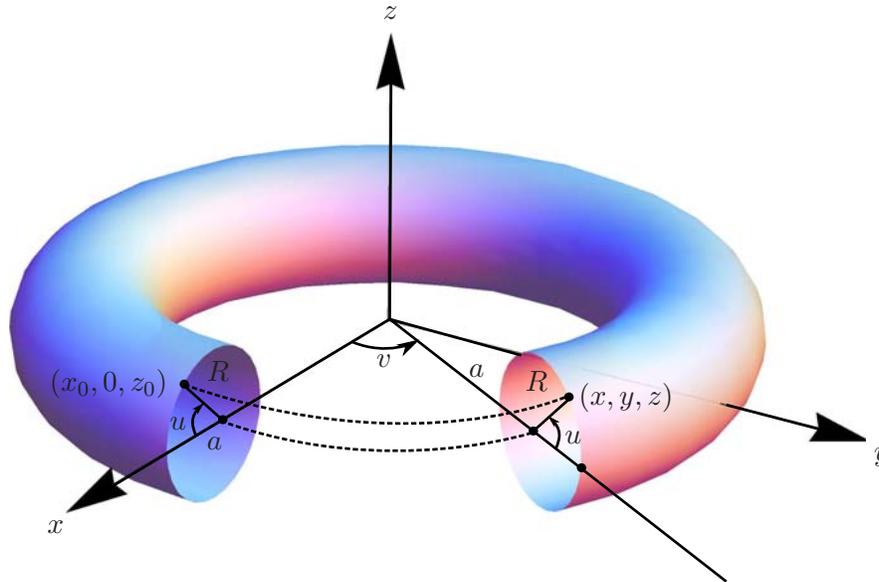


FIGURE 14.40. A torus. Consider a circle of radius R in the xz plane whose center is positioned on the positive x axis at a distance $a > R$. Any point $(x_0, 0, z_0)$ on the circle is obtained from the point $(a + R, 0, 0)$ by rotation about the center of the circle through an angle $0 \leq u \leq 2\pi$ so that $x_0 = a + R \cos u$ and $z_0 = R \sin u$. A torus is a surface swept by the circle when the xz plane is rotated about the z axis. A generic point (x, y, z) on the torus is obtained from $(x_0, 0, z_0)$ by rotating the latter about the z axis through an angle $0 \leq v \leq 2\pi$. Under this rotation, z_0 does not change and $z = z_0$, while the pair $(x_0, 0)$ in the xy plane changes to $(x, y) = (x_0 \cos v, x_0 \sin v)$. Parametric equations of a torus are $x = (a + R \cos u) \cos v$, $y = (a + R \cos u) \sin v$, $z = R \sin u$, where (u, v) ranges over the rectangle $[0, 2\pi] \times [0, 2\pi]$.

where $(u, v) \in [0, 2\pi] \times [0, 2\pi]$. An alternative (geometrical) derivation of these parametric equations is given in the caption of Figure 14.40. \square

A Tangent Plane to a Parametric Surface. The line $v = v_0$ in D is mapped onto the curve $\mathbf{r} = \mathbf{r}(u, v_0)$ in S (see Figure 14.41). The derivative $\mathbf{r}'_u(u, v_0)$ is tangent to the curve. Similarly, the line $u = u_0$ in D is mapped to the curve $\mathbf{r} = \mathbf{r}(u_0, v)$ in S , and the derivative $\mathbf{r}'_v(u_0, v)$ is tangent to it. If the cross product $\mathbf{r}'_u \times \mathbf{r}'_v$ does not vanish in D , then one can define a plane normal to the cross product at any point of S . Furthermore, if $\mathbf{r}'_u \times \mathbf{r}'_v \neq \mathbf{0}$ in a neighborhood of (u_0, v_0) , then,

without loss of generality, one can assume that, say, the z component of the cross product is not 0: $x'_u y'_v - x'_v y'_u = \partial(x, y)/\partial(u, v) \neq 0$. This shows that the transformation $x = x(u, v)$, $y = y(u, v)$ with continuous partial derivatives has a nonvanishing Jacobian. By the inverse function theorem (Theorem 14.10), there exists an inverse transformation $u = u(x, y)$, $v = v(x, y)$ that also has continuous partial derivatives. So the vector function $\mathbf{r}(u, v)$ can be written in the new variables (x, y) as

$$\mathbf{R}(x, y) = \mathbf{r}(u(x, y), v(x, y)) = (x, y, z(u(x, y), v(x, y))) = (x, y, g(x, y)),$$

which is a vector function that traces out the graph $z = g(x, y)$. Thus, a smooth parametric surface near any of its points can always be represented as the graph of a function of two variables. By the chain rule, the function g has continuous partial derivatives. Therefore, its linearization near $(x_0, y_0) = (x(u_0, v_0), y(u_0, v_0))$ defines the tangent plane to the graph and hence to the parametric surface at the point $\mathbf{r}_0 = \mathbf{r}(u_0, v_0)$. In particular, the vectors \mathbf{r}'_v and \mathbf{r}'_u must lie in this plane as they are tangent to two curves in the graph. Thus, the vector $\mathbf{r}'_u \times \mathbf{r}'_v$ is normal to the tangent plane. So Definition 14.23 of a smooth parametric surface agrees with the notion of a smooth surface introduced in Section 103.1 and the following theorem holds.

THEOREM 14.23. (Normal to a Smooth Parametric Surface).

Let $\mathbf{r} = \mathbf{r}(u, v)$ be a smooth parametric surface. Then the vector $\mathbf{n} = \mathbf{r}'_u \times \mathbf{r}'_v$ is normal to the surface.

Area of a Smooth Parametric Surface. Owing to the definition of the surface area element of the graph and the established relation between graphs and smooth parametric surfaces, the area a smooth surface can be found using the tangent planes to it (see Figure 14.38, left panel). Let a region D spanned by the parameters (u, v) be partitioned by rectangles of area $\Delta A = \Delta u \Delta v$. Then the vector function $\mathbf{r}(u, v)$ defines a partition of the surface (a partition element of the surface is the image of a partition rectangle in D). Consider a rectangle $[u_0, u_0 + \Delta u] \times [v_0, v_0 + \Delta v] = R_0$. 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 $\mathbf{r}_b = \mathbf{r}(u_0, v_0 + \Delta v)$, respectively. The area ΔS of the image of the rectangle R_0 can be

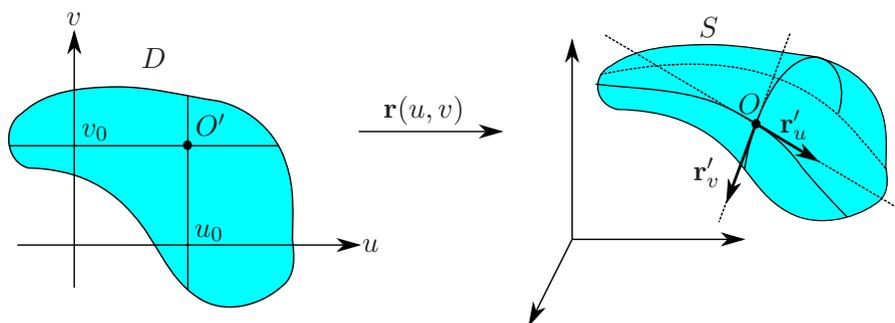


FIGURE 14.41. The lines $u = u_0$ and $v = v_0$ in D are mapped onto the curves in S that are traced out by the vector functions $\mathbf{r} = \mathbf{r}(u_0, v)$ and $\mathbf{r} = \mathbf{r}(u, v_0)$, respectively. The curves intersect at the point O with the position vector $\mathbf{r}(u_0, v_0)$. The derivatives $\mathbf{r}'_v(u_0, v_0)$ and $\mathbf{r}'_u(u_0, v_0)$ are tangential to the curves. If they do not vanish and are not parallel, then their cross product is normal to the plane through O that contains \mathbf{r}'_u and \mathbf{r}'_v . If the parametric surface is smooth, this plane is tangent to it.

approximated by the area of the parallelogram $\|\mathbf{a} \times \mathbf{b}\|$ with adjacent sides:

$$\begin{aligned}\mathbf{a} &= \overrightarrow{OA} = \mathbf{r}_a - \mathbf{r}_0 = \mathbf{r}(u_0 + \Delta u, v_0) - \mathbf{r}(u_0, v_0) = \mathbf{r}'_u(u_0, v_0) \Delta u, \\ \mathbf{b} &= \overrightarrow{OB} = \mathbf{r}_b - \mathbf{r}_0 = \mathbf{r}(u_0, v_0 + \Delta v) - \mathbf{r}(u_0, v_0) = \mathbf{r}'_v(u_0, v_0) \Delta v.\end{aligned}$$

The last equalities are obtained by the linearization of the components of $\mathbf{r}(u, v)$ near (u_0, v_0) , which is justified because the surface has a tangent plane at any point. The area transformation law is now easy to find: $\Delta S = \|\mathbf{a} \times \mathbf{b}\| = \|\mathbf{r}'_u \times \mathbf{r}'_v\| \Delta A$. Thus, the surface area of a smooth parametric surface is given by the double integral

$$A(S) = \iint_D \|\mathbf{r}'_u \times \mathbf{r}'_v\| dA.$$

Accordingly, the surface integral of a function $f(\mathbf{r})$ over a smooth parametric surface is

$$\iint_S f(\mathbf{r}) dS = \iint_D f(\mathbf{r}(u, v)) \|\mathbf{r}'_u \times \mathbf{r}'_v\| dA.$$

EXAMPLE 14.40. Find the surface area of the torus (14.37).

SOLUTION: To shorten the notation, put $w = R + a \cos u$. One has

$$\begin{aligned}\mathbf{r}'_u &= (-a \sin u \cos v, -a \sin u \sin v, R \cos u), \\ \mathbf{r}'_v &= -(R + a \cos u) \sin v, (R + a \cos u) \cos v, 0) \\ &= w(-\sin v, \cos v, 0), \\ \mathbf{n} &= \mathbf{r}'_u \times \mathbf{r}'_v = w(-a \cos v \cos u, -a \cos v \sin u, -a \sin u), \\ J &= \|\mathbf{r}'_u \times \mathbf{r}'_v\| = aw = a(R + a \cos u).\end{aligned}$$

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 14.41. 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 (14.36), one finds

$$\begin{aligned}\mathbf{r}'_u &= (R \cos v \cos u, R \sin v \cos u, -R \sin u), \\ \mathbf{r}'_v &= (-R \sin v \sin u, R \cos v \sin u, 0) \\ &= R \sin u(-\sin v, \cos v, 0), \\ \mathbf{n} &= \mathbf{r}'_u \times \mathbf{r}'_v = R \sin u(R \sin u \cos v, R \sin u \sin v, R \cos u) \\ &= R \sin u \mathbf{r}(u, v), \\ J &= \|\mathbf{r}'_u \times \mathbf{r}'_v\| = R \sin u \|\mathbf{r}(u, v)\| = R^2 \sin u, \\ f(\mathbf{r}(u, v)) &= (R \cos u)^2 R^2 \sin^2 u = R^4 \cos^2 u (1 - \cos^2 u).\end{aligned}$$

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. □

108.5. Exercises.**(1)** Find the surface area of the specified surface:

- (i) The part of the plane in the first octant that intersects the coordinate axes at $(a, 0, 0)$, $(0, b, 0)$, and $(0, 0, c)$, where a , b , and c are positive numbers
- (ii) The part of the plane $3x + 2y + z = 1$ that lies inside the cylinder $x^2 + y^2 = 4$
- (iii) The part of the hyperboloid $z = y^2 - x^2$ that lies between the cylinders $x^2 + y^2 = 1$ and $x^2 + y^2 = 4$
- (iv) The part of the paraboloid $z = x^2 + y^2$ that lies between two planes $z = 1$ and $z = 9$
- (v) The part of the surface $y = 4x + z^2$ that lies between the planes $x = 0$, $x = 1$, $z = 0$, and $z = 1$

(2) Evaluate the integral over the specified surface:

- (i) $\iint_S yz \, dS$, where S is the part of the plane $x + y + z = 1$ that lies in the first octant
- (ii) $\iint_S x^2 z^2 \, dS$, where S is the part of the cone $z^2 = x^2 + y^2$ that lies between the planes $z = 1$ and $z = 2$
- (iii) $\iint_S xz \, dS$, where S is the boundary of the solid region enclosed by the cylinder $y^2 + z^2 = 1$ and the planes $x = 0$ and $x + y = 3$ (*Hint:* Use the additivity of the surface integral.)
- (iv) $\iint_S z \, dS$, where S is the part of the sphere $x^2 + y^2 + z^2 = 2$ that lies above the plane $z = 1$
- (v) $\iint_S z(\sin(x^2) - \sin(y^2)) \, dS$, where S is the part of the paraboloid $z = 1 - x^2 - y^2$ that lies in the first octant (*Hint:* Use the symmetry.)

(3) Suppose that $f(\mathbf{r}) = g(\|\mathbf{r}\|)$, where $\mathbf{r} = (x, y, z)$. If $g(a) = 2$, use the geometrical interpretation of the surface integral to find $\iint_S f \, dS$, where S is the sphere of radius a centered at the origin.**(4)** Identify and sketch the parametric surface:

- (i) $\mathbf{r}(u, v) = (u + v, 2 - v, 2 - 2v + 3u)$
- (ii) $\mathbf{r}(u, v) = (a \cos u, b \sin u, v)$

(5) For the given parametric surface, sketch the curves $\mathbf{r}(u, v_0)$ for several fixed values $v = v_0$ and the curves $\mathbf{r}(u_0, v)$ for several fixed values $u = u_0$. Use them to visualize the parametric surface if

- (i) $\mathbf{r}(u, v) = (\sin v, u \sin v, \sin u \sin(2v))$
- (ii) $\mathbf{r}(u, v) = (u \cos v \sin \theta, u \sin v \sin \theta, u \cos \theta)$, where $0 \leq \theta \leq \pi/2$ is a parameter

(6) Find a parametric representation for the following surfaces

- (i) The plane through \mathbf{r}_0 that contains two nonzero and nonparallel vectors \mathbf{a} and \mathbf{b}
- (ii) The elliptic cylinder $y^2/a^2 + z^2/b^2 = 1$
- (iii) The part of the sphere $x^2 + y^2 + z^2 = a^2$ that lies below the cone $z = \sqrt{x^2 + y^2}$
- (iv) The ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$

(7) Find an equation of the tangent plane to the given parametric surface at the specified point P :

- (i) $\mathbf{r}(u, v) = (u^2, u - v, u + v)$ at $P = (1, -1, 3)$
- (ii) $\mathbf{r}(u, v) = (\sin v, u \sin v, \sin u \sin(2v))$ at $P = (1, \pi/2, 0)$

(8) Evaluate the surface integral over the specified parametric surface:

- (i) $\iint_S z^2 dS$, where S is the torus (14.37) with $R = 1$ and $a = 2$
- (ii) $\iint_S (1 + x^2 + y^2)^{1/2} dS$, where S is the *helicoid* with parametric equations $\mathbf{r}(u, v) = (u \cos v, u \sin v, v)$ and $(u, v) \in [0, 1] \times [0, \pi]$
- (iii) $\iint_S z dS$, where S is the part of the helicoid $\mathbf{r}(u, v) = (u \cos v, u \sin v, v)$, $(u, v) \in [0, a] \times [0, 2\pi]$
- (iv) $\iint_S z^2 dS$, where S is the part of the cone $x = u \cos v \sin \theta$, $y = u \sin v \sin \theta$, $z = u \cos \theta$, and $(u, v) \in [0, a] \times [0, 2\pi]$, and $0 \leq \theta \leq \pi/2$ is a parameter

(9) Evaluate the surface integral. If necessary, use suitable parametric equations of the surface:

- (i) $\iint_S (x^2 + y^2 + z^2) dS$, where S is the sphere $x^2 + y^2 + z^2 = R^2$
- (ii) $\iint_S (x^2 + y^2 + z^2) dS$, where S is the surface $|x| + |y| + |z| = R$; compare the result with the previous exercise
- (iii) $\iint_S (x^2 + y^2) dS$, where S is the boundary of the solid $\sqrt{x^2 + y^2} \leq z \leq 1$
- (iv) $\iint_S (1 + x + y)^{-2} dS$, where S is the boundary of the tetrahedron bounded by the coordinate planes and by the plane $x + y + z = 1$
- (v) $\iint_S |xyz| dS$, where S is the part of the paraboloid $z = x^2 + y^2$ below the plane $z = 1$
- (vi) $\iint_S (1/h) dS$, where S is an ellipsoid $(x/a)^2 + (y/b)^2 + (z/c)^2 = 1$ and h is the distance from the origin to the plane tangent to the ellipsoid at the point where the surface area element dS is taken
- (vii) $\iint_S (xy + yz + zx) dS$, where S is the part of the cone $z = \sqrt{x^2 + y^2}$ cut out by the cylinder $x^2 + y^2 = 2ax$

(10) Prove the *Poisson formula*

$$\iint_S f(ax + by + cz) dS = 2\pi \int_{-1}^1 f(u\sqrt{a^2 + b^2 + c^2}) du,$$

where S is the sphere $x^2 + y^2 + z^2 = 1$.

(11) Evaluate $F(a, b, c, t) = \iint_S f(x, y, z) dS$, where S is the sphere $(x - a)^2 + (y - b)^2 + (z - c)^2 = t^2$, $f(x, y, z) = 1$ if $x^2 + y^2 + z^2 < R^2$ and $f(x, y, z) = 0$ elsewhere. Assume that $\sqrt{a^2 + b^2 + c^2} > R > 0$.

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; K_{rot} is determined by *moments of inertia* discussed later. 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-fall 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 $x > x_c$.

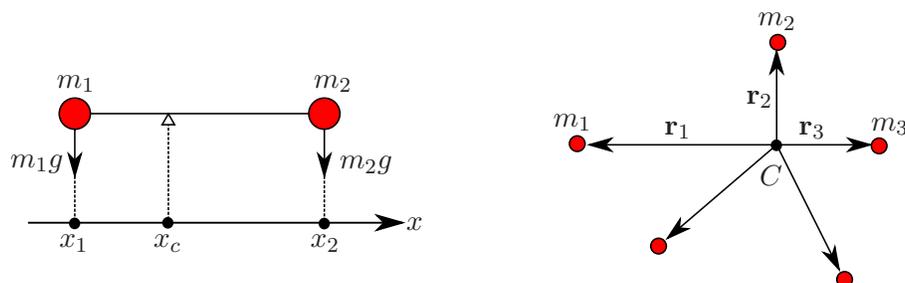


FIGURE 14.42. **Left:** Two masses connected by a rigid massless rod (or its mass is much smaller than the masses m_1 and m_2) are positioned at x_1 and x_2 . The gravitational force is perpendicular to the rod. The center of mass x_c is determined by the condition that the system does not rotate about x_c under the gravitational forces. **Right:** An extended object consisting of point masses with fixed distances between them. If the position vectors of the masses relative to the center of mass C are \mathbf{r}_i , then $m_1\mathbf{r}_1 + m_2\mathbf{r}_2 + \cdots + m_N\mathbf{r}_N = \mathbf{0}$.

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. It is shown in the left panel of Figure 14.42. 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

$$m_1(x_1 - x_c) + m_2(x_2 - x_c) = 0 \quad \implies \quad x_c = \frac{m_1x_1 + m_2x_2}{m_1 + m_2}.$$

The center of mass (x_c, y_c) of point masses m_i , $i = 1, 2, \dots, N$, 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 (a line) is put underneath the plane, then due to an uneven 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:

$$\sum_{i=1}^N (x_i - x_c)m_i = 0 \quad \Longrightarrow \quad x_c = \frac{1}{m} \sum_{i=1}^N m_i x_i = \frac{M_y}{m}, \quad m = \sum_{i=1}^N m_i,$$

$$\sum_{i=1}^N (y_i - y_c)m_i = 0 \quad \Longrightarrow \quad y_c = \frac{1}{m} \sum_{i=1}^N m_i y_i = \frac{M_x}{m},$$

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 $y = 0$).

Consider an extended object that is a collection of point masses shown in the right panel of Figure 14.42. Its center of mass is defined similarly by assuming that the total moments about any 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 14.24. (Center of Mass).

Suppose that an extended object consists of N point masses m_i , $i = 1, 2, \dots, N$, whose position vectors are \mathbf{r}_i . Then its center of mass is a point with the position vector

$$(14.38) \quad \mathbf{r}_c = \frac{1}{m} \sum_{i=1}^N m_i \mathbf{r}_i, \quad m = \sum_{i=1}^N m_i,$$

where m is the total mass of the object. The quantities

$$M_{yz} = \sum_{i=1}^N m_i x_i, \quad M_{xz} = \sum_{i=1}^N m_i y_i, \quad M_{xy} = \sum_{i=1}^N m_i z_i$$

are called the moments about the coordinate planes.

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 i th partition piece lies. Although all the partition pieces are small, they still have finite sizes R_i , and the definition (14.38) 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 $N \rightarrow \infty$ in the sense that all the partition sizes tend to 0 uniformly, $R_i \leq \max_i R_i = R_N \rightarrow 0$ as $N \rightarrow \infty$. 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 i th partition element and

$$\mathbf{r}_c = \frac{1}{m} \lim_{N \rightarrow \infty} \sum_{i=1}^N \mathbf{r}_i^* \Delta m_i = \frac{1}{m} \iiint_E \mathbf{r} \sigma(\mathbf{r}) dV, \quad m = \iiint_E \sigma(\mathbf{r}) dV.$$

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 (14.34), or to a line integral, according to (14.31). For two- and one-dimensional extended objects, the center of mass can be written as, respectively,

$$\begin{aligned} \mathbf{r}_c &= \frac{1}{m} \iint_S \mathbf{r} \sigma(\mathbf{r}) dS, & m &= \iint_S \sigma(\mathbf{r}) dS, \\ \mathbf{r}_c &= \frac{1}{m} \int_C \mathbf{r} \sigma(\mathbf{r}) ds, & m &= \int_C \sigma(\mathbf{r}) ds, \end{aligned}$$

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 14.42. 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

$$\begin{aligned} m &= \iint_D \sigma dA = k \iint_D y dA = k \int_0^\pi \int_0^R r \sin \theta r dr d\theta \\ &= 2k \int_0^R r^2 dr = \frac{2kR^3}{3}, \end{aligned}$$

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 $y_c = M_x/m = 3\pi R/16$. \square

EXAMPLE 14.43. 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 $\sigma = k = \text{const}$. 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

$$\begin{aligned} M_{xy} &= \iiint_E z\sigma dV = k \int_0^{2\pi} \int_0^{\pi/3} \int_a^b \rho \cos \phi \rho^2 \sin \phi d\rho d\phi d\theta \\ &= \frac{3\pi k}{16} (b^4 - a^4). \end{aligned}$$

So $z_c = M_{xy}/m = (9/16)(a + b)(a^2 + b^2)/(a^2 + ab + b^2)$. \square

Centroid. The center of mass of an extended object with a *constant* mass density is called the *centroid*. The centroid of a region depends only on the shape of the region. In this sense, the centroid is an intrinsic (geometrical) characteristic of the region.

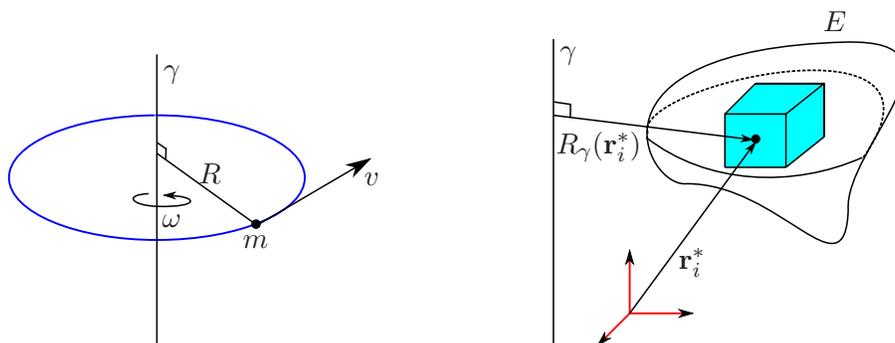


FIGURE 14.43. **Left:** The moment of inertia of a point mass about an axis γ . A point mass rotates about an axis γ with the rate ω , called the angular velocity. Its linear velocity is $v = \omega R$, where R is the distance from γ . So the kinetic energy of the rotational motion is $mv^2/2 = mR^2\omega^2/2 = I_\gamma\omega^2/2$, where $I_\gamma = mR^2$ is the moment of inertia. **Right:** The moment of inertia of an extended object E about an axis γ is defined as the sum of moments of inertia of partition elements of E : $\Delta I_i = \Delta m(\mathbf{r}_i^*)R_\gamma^2(\mathbf{r}_i^*)$, where $R_\gamma(\mathbf{r}_i^*)$ is the distance to the axis γ and $\Delta m(\mathbf{r}_i^*)$ is the mass of the partition element.

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*). The system is shown in Figure 14.43 (left panel). 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

$$K_{\text{rot}} = \frac{mv^2}{2} = \frac{mR^2\omega^2}{2} = \frac{I_\gamma\omega^2}{2}.$$

The constant I_γ is called the *moment of inertia* of the point mass m about the axis γ . Similarly, consider an extended solid object consisting of N point masses. The distances between the masses do not change when the object moves (the object is *solid*). 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 $K_{\text{rot}} = I_\gamma \omega^2 / 2$, 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 $\sigma(\mathbf{r})$ (see the right panel of Figure 14.43). 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 $\Delta m_i = \sigma(\mathbf{r}_i^*) \Delta V_i$ 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

$$I_\gamma = \lim_{N \rightarrow \infty} \sum_{i=1}^N R_\gamma^2(\mathbf{r}_i^*) \sigma(\mathbf{r}_i^*) \Delta V_i = \iiint_E R_\gamma^2(\mathbf{r}) \sigma(\mathbf{r}) dV$$

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

$$\begin{aligned} I_x &= \iiint_E (y^2 + z^2) \sigma dV, & I_y &= \iiint_E (x^2 + z^2) \sigma dV, \\ I_z &= \iiint_E (x^2 + y^2) \sigma dV. \end{aligned}$$

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,

$$\begin{aligned} R_\gamma^2(\mathbf{r}) &= \|\hat{\mathbf{u}} \times \mathbf{r}\|^2 = (\hat{\mathbf{u}} \times \mathbf{r}) \cdot (\hat{\mathbf{u}} \times \mathbf{r}) = \hat{\mathbf{u}} \cdot (\mathbf{r} \times (\hat{\mathbf{u}} \times \mathbf{r})) \\ (14.39) \quad &= \mathbf{r}^2 - (\hat{\mathbf{u}} \cdot \mathbf{r})^2, \end{aligned}$$

where the *bac* – *cab* rule (11.9) 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 (14.34) or (14.31); that is, for two- or one-dimensional objects, the moment of inertia becomes, respectively,

$$I_\gamma = \iint_S R_\gamma^2(\mathbf{r}) \sigma(\mathbf{r}) dS, \quad I_\gamma = \int_C R_\gamma^2(\mathbf{r}) \sigma(\mathbf{r}) ds,$$

where σ is either the surface or linear mass density.

EXAMPLE 14.44. *A rocket tip is made of thin plates with a constant surface mass density $\sigma = k$. It has a circular conic shape with base*

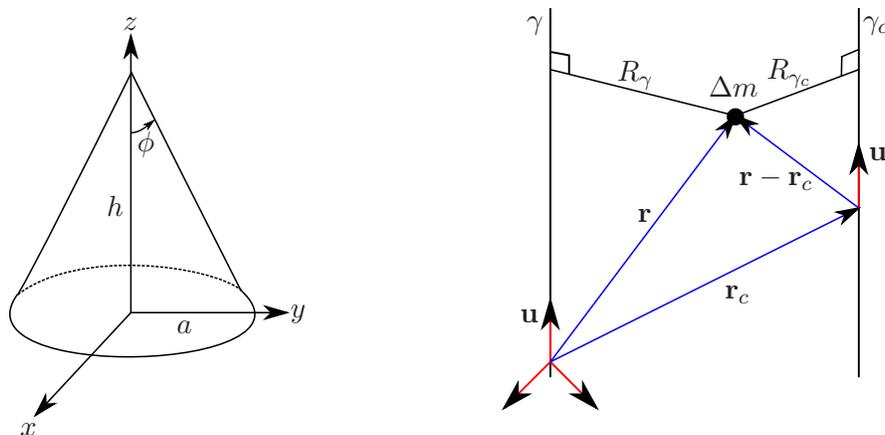


FIGURE 14.44. **Left:** An illustration to Example 14.44. **Right:** An illustration to the proof of the parallel axis theorem for moments of inertia (Study Problem 14.11). The axis γ_c is parallel to γ and goes through the center of mass with the position vector \mathbf{r}_c . The vectors \mathbf{r} and $\mathbf{r} - \mathbf{r}_c$ are position vectors of a partition element of mass Δm relative to the origin and the center of mass, respectively.

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 $D: x^2 + y^2 \leq a^2$. 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 $g'_y = (hy/a)(x^2 + y^2)^{-1/2}$ 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 14.45. 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

$$\begin{aligned} I_z &= \iiint_E (x^2 + y^2) \sigma \, dV = \frac{3m}{4\pi a^3} \int_0^{2\pi} \int_0^\pi \int_0^a (\rho \sin \phi)^2 \rho^2 \sin \phi \, d\rho \, d\phi \, d\theta \\ &= \frac{3}{10} m a^2 \int_0^\pi \sin^3 \phi \, d\phi = \frac{3}{10} m a^2 \int_{-1}^1 (1 - u^2) \, du = \frac{2}{5} m a^2, \end{aligned}$$

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:

$$I_z = \frac{1}{3} \sigma \iiint_E 2(x^2 + y^2 + z^2) \, dV = \frac{1}{3} \sigma 8\pi \int_0^a \rho^4 \, d\rho = \frac{2}{5} m a^2.$$

□

EXAMPLE 14.46. Find the center of mass and the moment of inertia of a homogeneous rod of mass m bent into a half-circle of radius R about the line through the endpoints of the rod.

SOLUTION: Set up the coordinate system so the half-circle lies above the x axis: $x^2 + y^2 = R^2$, $y \geq 0$. The linear mass density is constant $\sigma = m/(\pi R)$, where πR is the length of the rod. By the symmetry of the mass distribution, the center of mass lies on the y axis, $x_c = 0$, and $y_c = (1/m) \int_C y \sigma \, ds$. To evaluate the line integral, choose the following parametric equation of the half-circle $\mathbf{r}(t) = (R \cos t, R \sin t)$, $0 \leq t \leq \pi$. Then $\mathbf{r}'(t) = (-R \sin t, R \cos t)$ and $ds = \|\mathbf{r}'(t)\| dt = R \, dt$. Therefore,

$$y_c = \frac{1}{m} \int_C y \sigma \, ds = \frac{1}{\pi R} \int_0^\pi R \sin t R \, dt = \frac{2R}{\pi}.$$

If R_γ is the distance from the line connecting the endpoint of the rod to its particular point, then, in the chosen coordinate system, $R_\gamma = y$. Therefore, the moment of inertia in question is

$$I_\gamma = \int_C R_\gamma^2 \sigma \, ds = \frac{m}{\pi R} \int_C y^2 \, ds = \frac{mR^2}{\pi} \int_0^\pi \sin^2 t \, dt = \frac{mR^2}{2}.$$

□

109.3. Study Problems.

Problem 14.10. Find the center of mass of the shell described in Example 14.44.

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 14.44, 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

$$\begin{aligned} M_{xy} &= \iint_S z\sigma \, dS = k \iint_D (h/a)\sqrt{x^2 + y^2} J \, dA = \frac{kJh}{a} \iint_D \sqrt{x^2 + y^2} \, dA \\ &= \frac{kJh}{a} \int_0^{2\pi} \int_0^a r^2 \, dr \, d\theta = \frac{2\pi kha}{3} \sqrt{h^2 + a^2}. \end{aligned}$$

Thus, the center of mass is at the distance $z_c = M_{xy}/m = 2h/3$ from the tip of the cone. \square

Problem 14.11. (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 (see the right panel of Figure 14.44). 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 $R_{\gamma_c}^2(\mathbf{r})$ is obtained from $R_\gamma^2(\mathbf{r})$ (see (14.39)) 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, $R_\gamma^2(\mathbf{r}_c) = R_c^2$ by the definition of the function $R_\gamma(\mathbf{r})$. Hence, by (14.39),

$$\begin{aligned} R_\gamma^2(\mathbf{r}) - R_{\gamma_c}^2(\mathbf{r}) &= R_\gamma^2(\mathbf{r}) - R_\gamma^2(\mathbf{r} - \mathbf{r}_c) \\ &= 2\mathbf{r}_c \cdot \mathbf{r} - \mathbf{r}_c^2 - (\hat{\mathbf{u}} \cdot \mathbf{r}_c)(2\hat{\mathbf{u}} \cdot \mathbf{r} - \hat{\mathbf{u}} \cdot \mathbf{r}_c) \\ &= \mathbf{r}_c^2 - (\hat{\mathbf{u}} \cdot \mathbf{r}_c)^2 + 2\mathbf{r}_c \cdot (\mathbf{r} - \mathbf{r}_c) - 2(\hat{\mathbf{u}} \cdot \mathbf{r}_c)\hat{\mathbf{u}} \cdot (\mathbf{r} - \mathbf{r}_c) \\ &= R_c^2 - 2\mathbf{a} \cdot (\mathbf{r} - \mathbf{r}_c), \end{aligned}$$

where $\mathbf{a} = \mathbf{r}_c - (\hat{\mathbf{u}} \cdot \mathbf{r}_c)\hat{\mathbf{u}}$. Therefore,

$$\begin{aligned} I_\gamma - I_{\gamma_c} &= \iiint_E \left(R_\gamma^2(\mathbf{r}) - R_{\gamma_c}^2(\mathbf{r}) \right) \sigma(\mathbf{r}) dV \\ &= R_c^2 \iiint_E \sigma(\mathbf{r}) dV - 2\mathbf{a} \cdot \iiint_E (\mathbf{r} - \mathbf{r}_c) \sigma(\mathbf{r}) dV = R_c^2 m, \end{aligned}$$

where the second integral vanishes by the definition of the center of mass. \square

Problem 14.12. 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 14.45. By the parallel axis theorem, for any axis γ at a distance R from the center of mass, $I_\gamma = I_{\gamma_c} + mR^2 = m(R^2 + 2a^2/5)$. \square

109.4. Exercises.

- (1) Find the center of mass of the specified extended object:
 - (i) A homogeneous thin rod of length L
 - (ii) A homogeneous thin wire that occupies the part of a circle of radius R that lies in the first quadrant
 - (iii) A homogeneous thin wire bent into one turn of the helix of radius R that rises by the distance h per each turn
 - (iv) A homogeneous thin shell that occupies a hemisphere of radius R
 - (v) A homogeneous thin disk of radius R that has a circular hole of radius $a < R/2$ and whose center is at the distance $R/2$ from the disk center
 - (vi) A homogeneous solid enclosed by the ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$ that has a square box cavity $[0, h] \times [0, h] \times [0, h]$
 - (vii) The part of the ball $x^2 + y^2 + z^2 \leq 4$ that lies above the cone $z\sqrt{3} = \sqrt{x^2 + y^2}$ and whose mass density at any point is proportional to its distance from the origin
 - (viii) The part of the spherical shell $a^2 \leq x^2 + y^2 + z^2 \leq b^2$ that lies above the xy plane and whose mass density at any point is proportional to its distance from the z axis
 - (ix) The part of the disk $x^2 + y^2 = a^2$ in the first quadrant bounded by the lines $y = x$ and $y = \sqrt{3}x$ if the mass density at any point is proportional to its distance from the origin

- (x) The part of the solid enclosed by the paraboloid $z = 2 - x^2 - y^2$ and the cone $z = \sqrt{x^2 + y^2}$ that lies in the first octant and whose mass density at any point is proportional to its distance from the z axis
- (xi) A homogeneous surface cut from the cone $z = \sqrt{x^2 + y^2}$ by the cylinder $x^2 + y^2 = ax$
- (xii) The part of a homogeneous sphere defined by $z = \sqrt{a^2 - x^2 - y^2}$, $x \geq 0$, $y \geq 0$, $x + y \leq a$
- (xiii) The arc of the homogeneous cycloid $x = a(t - \sin t)$, $y = a(1 - \cos t)$, $0 \leq t \leq \pi$
- (xiv) The arc of the homogeneous curve $y = a \cosh(x/a)$ from the point $(0, a)$ to the point (b, h)
- (xv) The arc of the homogeneous astroid $x^{2/3} + y^{2/3} = a^{2/3}$ in the first quadrant
- (xvi) The homogeneous lamina bounded by the curves $\sqrt{x} + \sqrt{y} = \sqrt{a}$, $x = 0$, $y = 0$
- (xvii) The part of the homogeneous lamina bounded by the curve $x^{2/3} + y^{2/3} = a^{2/3}$ in the first quadrant
- (xviii) The homogeneous solid bounded by the surfaces $x^2 + y^2 = 2z$, $x + y = z$
- (xix) The homogeneous solid bounded by the surfaces $z = x^2 + y^2$, $2z = x^2 + y^2$, $x + y = \pm 1$, $x - y = \pm 1$

(2) Show that the centroid of a triangle is the point of intersection of its medians (the lines joining each vertex with the midpoint of the opposite side).

(3) Show that the centroid of a pyramid is located on the line segment that connects the apex to the centroid of the base and is $1/4$ the distance from the base to the apex.

(4) Find the specified moment of inertia of the given extended object:

- (i) The smaller wedge cut out from a ball of radius R by two planes that intersect along the diameter of the ball at an angle $0 < \theta_0 \leq \pi$. The wedge is homogeneous and has mass m . Find the moment of inertia about the diameter.
- (ii) The moment of inertia about the z axis of the solid that is enclosed by the cylinder $x^2 + y^2 \leq 1$ and the planes $z = 0$, $y + z = 5$ and has a mass density of $\sigma(x, y, z) = 10 - 2z$.
- (iii) A thin homogeneous shell in the shape of the torus with radii R and $a > R$ that has mass m . The moment of inertia about the symmetry axis of the torus.

- (iv) The moments of inertia I_x and I_y of the part of the disk of radius a that lies in the first quadrant and whose mass density at any point is proportional to its distance from the y axis.
- (v) The solid homogeneous cone with height h and radius of the base a . The moments of inertia about its symmetry axis, the axis through its vertex and perpendicular to the symmetry axis, and an axis that contains a diameter of the base.
- (vi) The part of the homogeneous plane $x + y + z = a$; the moments of inertia about the coordinate axes.
- (vii) The homogeneous triangle of mass m whose vertices in polar coordinates are $(r, \theta) = (a, 0), (a, 2\pi/3), (a, 4\pi/3)$; the polar moment of inertia $I_0 = I_x + I_y$.
- (viii) The homogeneous solid cylinder $x^2 + y^2 \leq a^2, -h \leq z \leq h$ of mass m ; the moment of inertia about the line parallel to the z axis through the point $(a, 0, 0)$;
- (ix) The homogeneous solid of mass density σ_0 bounded by the surface $(x^2 + y^2 + z^2)^2 = a^2(x^2 + y^2)$; the sum of the moments of inertia $I_x + I_y + I_z$.
- (x) The lamina with a constant mass density σ_0 bounded by the circle $(x - a)^2 + (y - a)^2 = a^2$ and by the segments $0 \leq y \leq a, 0 \leq x \leq a$; the moments of inertia about the coordinate axes.
- (xi) The lamina with a constant mass density σ_0 bounded by the curves $xy = a^2, xy = 2a^2, x = 2y, 2x = y$; the moments of inertia about the coordinate axes.
- (xii) The solid that has a constant mass density σ_0 and is bounded by the ellipsoid $(x/a)^2 + (y/b)^2 + (z/c)^2 = 1$; moments of inertia about the coordinate axes.
- (xiii) The spherical homogeneous shell of mass m and radius R ; the moment of inertia about a diameter of the sphere.

CHAPTER 15

Vector Calculus

110. Line Integrals of a Vector Field

110.1. Vector Fields. Consider an airflow 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:

$$\mathbf{F}(\mathbf{r}) = \nabla f(\mathbf{r}) \iff F_1(\mathbf{r}) = f'_x(\mathbf{r}), \quad F_2(\mathbf{r}) = f'_y(\mathbf{r}), \quad F_3(\mathbf{r}) = f'_z(\mathbf{r}).$$

Many physical quantities are described by vector fields. Electric and magnetic fields are vector fields. All modern communication devices (radio, TV, cell phones, etc.) use electromagnetic waves. Visible light is also electromagnetic waves. 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. If one thinks about a planet as a homogeneous ball of mass M , then the gravitational force exerted by it on a point mass m depends on the position of the point mass relative to the planet's center according to Newton's law of gravity:

$$\mathbf{F}(\mathbf{r}) = -\frac{GMm}{r^3} \mathbf{r} = \left(-GMm \frac{x}{r^3}, -GMm \frac{y}{r^3}, -GMm \frac{z}{r^3} \right),$$

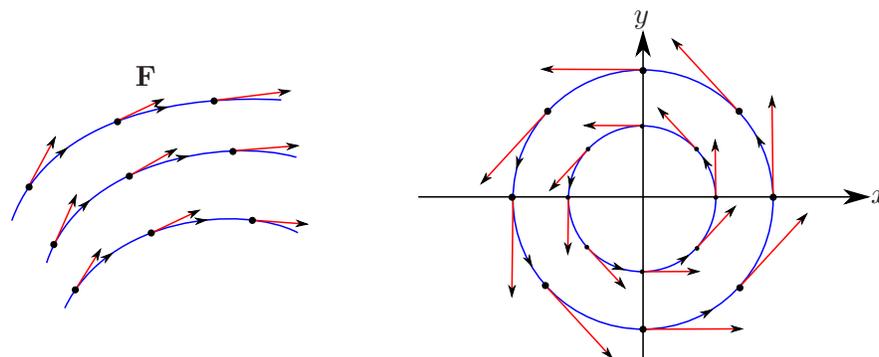


FIGURE 15.1. **Left:** Flow lines of a vector field \mathbf{F} are curves to which the vector field is tangential. The flow lines are oriented by the direction of the vector field. **Right:** Flow lines of the vector field $\mathbf{F} = (-y, x, 0)$ in Example 15.1 are concentric circles oriented counterclockwise. The magnitude $\|\mathbf{F}\| = \sqrt{x^2 + y^2}$ is constant along the flow lines and linearly increases with increasing distance from the origin.

where G is Newton's gravitational constant, \mathbf{r} is the position vector relative to the planet's center, and $r = \|\mathbf{r}\|$ is its length. The force is proportional to the position vector and hence parallel to it at each point. 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. The direction of a tangent vector \mathbf{F} is shown by arrows on the flow lines as depicted in the left panel of Figure 15.1. 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 $\mathbf{F} = \nabla f \neq \mathbf{0}$ are normal to level surfaces of the function f and oriented in the direction in which f increases most rapidly (Theorem 13.16). They are the

curves of steepest ascent of the function f . Flow lines of the velocity vector field of the air are often shown in weather forecasts to indicate the wind direction over large areas. For example, flow lines of the air velocity in 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, $\mathbf{r}'(t) = \mathbf{F}(\mathbf{r}(t))$. To find a flow line through a particular point \mathbf{r}_0 , the differential equations must be supplemented by *initial* conditions, for example, $\mathbf{r}(t_0) = \mathbf{r}_0$. If the equations have a unique solution, then the flow through \mathbf{r}_0 exists and is given by the solution.

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. The flow lines are shown in the right panel of Figure 15.1. \square

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 (Section 73.5). 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 is evidently of great practical significance. To answer it, the concept of the line integral of a vector field was developed.

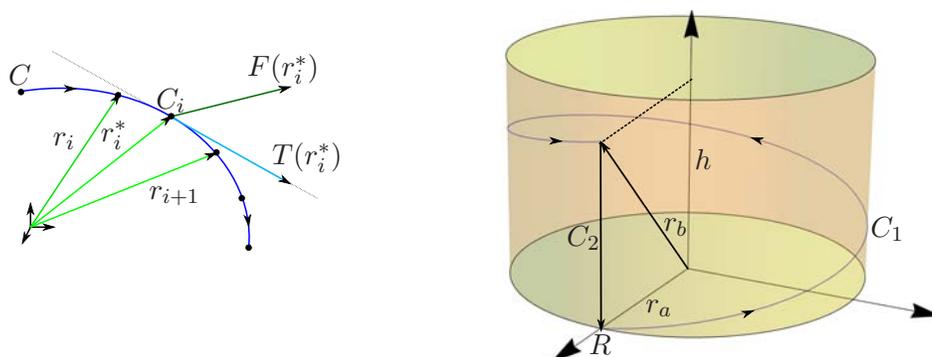


FIGURE 15.2. **Left:** To calculate the work done by a continuous force $\mathbf{F}(\mathbf{r})$ in moving a point object along a smooth curve C , the latter is partitioned into segments C_i with arc length Δs . The work done by the force along a partition segment is $\mathbf{F}(\mathbf{r}_i^*) \cdot \mathbf{d}_i$, where the displacement vector is approximated by the oriented segment of length Δs that is tangent to the curve at a sample point \mathbf{r}_i^* , that is, $\mathbf{d}_i = \hat{\mathbf{T}}(\mathbf{r}_i^*) \Delta s$, where $\hat{\mathbf{T}}$ is the unit tangent vector along the curve. **Right:** An illustration to Example 15.2. The closed contour of integration in the line integral consists of two smooth pieces, one turn of the helix C_1 and the straight line segment C_2 . The line integral is the sum of line integrals along C_1 and C_2 .

Let C be a smooth curve that goes from a point \mathbf{r}_a to a point \mathbf{r}_b and has length L . Consider a partition of C by segments C_i , $i = 1, 2, \dots, N$, of length $\Delta s = L/N$. Following the discussion of smooth curves in Sections 80.3 and 84.3, 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 $\mathbf{r}_i^* \in C_i$ (see the left panel of Figure 15.2). 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 $W = \Delta W_1 + \Delta W_2 + \dots + \Delta W_N$. The actual work should not depend on the choice of sample points. This problem is resolved by the usual trick of integral calculus by refining a partition, finding the low and upper sums, and taking their limits. If these limits exist and coincide, the limiting value should not depend on the choice of sample points and is the sought-after 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 of f along C . So, if the function f

is integrable on the curve C , then the work is the line integral 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 14.20). In particular, the line integral of a *continuous* vector field over a smooth curve of a finite length always exists.

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 only once so that $\mathbf{r}(a) = \mathbf{r}_a$ and $\mathbf{r}(b) = \mathbf{r}_b$. Then

$$\begin{aligned} \int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} &= \int_C \mathbf{F} \cdot \hat{\mathbf{T}} ds = \int_a^b \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) dt \\ (15.1) \quad &= \int_a^b \left(F_1(\mathbf{r}(t))x'(t) + F_2(\mathbf{r}(t))y'(t) + F_3(\mathbf{r}(t))z'(t) \right) dt. \end{aligned}$$

PROOF. The unit tangent vector reads $\hat{\mathbf{T}} = \mathbf{r}'/\|\mathbf{r}'\|$ and $ds = \|\mathbf{r}'\| dt$. Therefore, $\hat{\mathbf{T}} ds = \mathbf{r}'(t) dt$. For a smooth curve, $\mathbf{r}'(t)$ is continuous on $[a, b]$. Therefore, by the continuity of the vector field, 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 14.21. \square

Equation (15.1) 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. Owing to the representation (15.1) and the relations $dx = x'dt$, $dy = y'dt$, and $dz = z'dt$, the line integral is often written in the form:

$$(15.2) \quad \int_C \mathbf{F} \cdot d\mathbf{r} = \int_C F_1 dx + F_2 dy + F_3 dz.$$

For a smooth curve traversed by a vector function $\mathbf{r}(t)$, the differential $d\mathbf{r}(t)$ is tangent to the curve.

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 $a \leq t \leq b$, 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:

$$(15.3) \quad \int_{-C} \mathbf{F} \cdot d\mathbf{r} = - \int_C \mathbf{F} \cdot d\mathbf{r}$$

because the direction of the derivative $\mathbf{r}'(t)$ is reversed for all t . 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:

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

Line Integral Along a Parametric Curve. A parametric curve is defined by a vector function $\mathbf{r}(t)$ on $[a, b]$ (recall Definition 12.4). The vector function $\mathbf{r}(t)$ may trace its range (as a point set in space) or some parts of it several times as t changes from a to b . Furthermore, two *different* vector functions $\mathbf{r}_1(t)$ and $\mathbf{r}_2(t)$ on $[a, b]$ may have the same range. For example, $\mathbf{r}_1 = (\cos t, \sin t, 0)$ and $\mathbf{r}_2(t) = (\cos(2t), \sin(2t), 0)$ 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. The line integral over a parametric curve is *defined* by the relation (15.1). A parametric curve is much like the trajectory of a particle that can pass through the same points multiple times. So the relation (15.1) defines the work done by a nonconstant force \mathbf{F} along a particle's trajectory $\mathbf{r} = \mathbf{r}(t)$.

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 orientation opposite to the required one, then it can still be used according to (15.3).

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 (15.1).

EXAMPLE 15.2. Evaluate the line integral of $\mathbf{F} = (-y, x, z^2)$ along a closed contour C that consists of two parts. The first part is one turn of a helix of radius R , which winds about the z axis counterclockwise as viewed from the top of the z axis, starting from the point $\mathbf{r}_a = (R, 0, 0)$ and ending at the point $\mathbf{r}_b = (R, 0, 2\pi h)$. The second part is 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 . Note the positive signs at $\cos t$ and $\sin t$ in the parametric equations that are necessary to make the helix winding about the z axis counterclockwise (see Study Problem 12.1). The range of t has to be restricted to $[0, 2\pi]$. Then $\mathbf{r}'(t) = (-R \sin t, R \cos t, h)$. Therefore,

$$\begin{aligned} \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) &= (-R \sin t, R \cos t, h^2 t^2) \cdot (-R \sin t, R \cos t, h) \\ &= R^2 + h^3 t^2, \\ \int_{C_1} \mathbf{F} \cdot d\mathbf{r} &= \int_0^{2\pi} \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) dt = \int_0^{2\pi} (R^2 + h^3 t^2) dt \\ &= 2\pi R^2 + \frac{(2\pi h)^3}{3}. \end{aligned}$$

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 $t \in [0, 1]$. These parametric equations describe the curve $-C_2$. One has $\mathbf{r}'(t) = (0, 0, 2\pi h)$ and hence

$$\begin{aligned} \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) &= (0, R, (2\pi h)^2 t^2) \cdot (0, 0, 2\pi h) = (2\pi h)^3 t^2, \\ \int_{C_2} \mathbf{F} \cdot d\mathbf{r} &= - \int_{-C_2} \mathbf{F} \cdot d\mathbf{r} = -(2\pi h)^3 \int_0^1 t^2 dt = -\frac{(2\pi h)^3}{3}. \end{aligned}$$

The line integral along C is the sum of these integrals, which is equal to $2\pi R^2$. \square

110.5. Study Problem.

Problem 15.1. Find the work done by the force $\mathbf{F} = (2x, 3y^2, 4z^3)$ along any smooth curve originating from the point $(0, 0, 0)$ and ending at the point $(1, 1, 1)$.

SOLUTION: For any infinitesimal part of the curve, the work is

$$\mathbf{F} \cdot d\mathbf{r} = 2x dx + 3y^2 dy + 4z^3 dz = d(x^2 + y^3 + z^4).$$

If $\mathbf{r}(t) = (x(t), y(t), z(t))$ is a parametric equation of a smooth curve, $a \leq t \leq b$, such that $\mathbf{r}(a) = (0, 0, 0)$ and $\mathbf{r}(b) = (1, 1, 1)$, then the total work is

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_a^b d(x^2(t) + y^3(t) + z^4(t)) = (x^2(t) + y^3(t) + z^4(t)) \Big|_a^b = 3.$$

□

110.6. Exercises.

(1) Sketch flow lines of the given planar vector field:

- (i) $\mathbf{F} = (ax, by)$, where a and b are positive
- (ii) $\mathbf{F} = (ay, bx)$, where a and b are positive
- (iii) $\mathbf{F} = (ay, bx)$, where a and b have different signs
- (iv) $\mathbf{F} = \nabla u$, $u = \tan^{-1}(y/x)$
- (v) $\mathbf{F} = \nabla u$, $u = \ln[(x^2 + y^2)^{-1/2}]$
- (vi) $\mathbf{F} = \nabla u$, $u = \ln[(x - a)^2 + (y - b)^2]$

(2) Sketch flow lines of the given vector field in space:

- (i) $\mathbf{F} = (ax, by, cz)$, where a , b , and c are positive
- (ii) $\mathbf{F} = (ax, by, cz)$, where a and b are positive, while c is negative
- (iii) $\mathbf{F} = (y, -x, a)$, where a is a constant
- (iv) $\mathbf{F} = \nabla \|\mathbf{r}\|$, $\mathbf{r} = (x, y, z)$
- (v) $\mathbf{F} = \nabla \|\mathbf{r}\|^{-1}$, $\mathbf{r} = (x, y, z)$
- (vi) $\mathbf{F} = \nabla u$, $u = (x/a)^2 + (y/b)^2 + (z/c)^2$;
- (vii) $\mathbf{F} = \nabla u$, $u = \sqrt{x^2 + y^2 + (z + c)^2} + \sqrt{x^2 + y^2 + (z - c)^2}$, where c is positive
- (viii) $\mathbf{F} = \mathbf{a} \times \mathbf{r}$, where \mathbf{a} is a constant vector and $\mathbf{r} = (x, y, z)$
- (ix) $\mathbf{F} = \nabla u$, $u = z/\sqrt{x^2 + y^2 + z^2}$

(3) A ball rotates at a constant rate ω about its diameter parallel to a unit vector \mathbf{n} . If the origin of the coordinate system is set at the center of the ball, find the velocity vector field as a function of the position vector \mathbf{r} of a point of the ball.

(4) Evaluate the line integral $\int_C \mathbf{F} \cdot d\mathbf{r}$ for the given vector field \mathbf{F} and the specified curve C :

- (i) $\mathbf{F} = (y, xy, 0)$ and C is the parametric curve $\mathbf{r}(t) = (t^2, t^3, 0)$ for $t \in [0, 1]$
- (ii) $\mathbf{F} = (z, yx, zy)$ and C is the ellipse $x^2/a^2 + y^2/b^2 = 1$ oriented clockwise
- (iii) $\mathbf{F} = (z, yx, zy)$ and C is the parametric curve $\mathbf{r}(t) = (2t, t + t^2, 1 + t^3)$ from the point $(-2, 0, 0)$ to the point $(2, 2, 2)$
- (iv) $\mathbf{F} = (-y, x, z)$ and C is the boundary of the part of the paraboloid $z = a^2 - x^2 - y^2$ that lies in the first octant; C is oriented counterclockwise as viewed from the top of the z axis
- (v) $\mathbf{F} = (-z, 0, x)$ and C is the boundary of the part of the sphere $x^2 + y^2 + z^2 = a^2$ that lies in the first octant; C is oriented clockwise as viewed from the top of the z axis
- (vi) $\mathbf{F} = \mathbf{a} \times \mathbf{r}$, where \mathbf{a} is a constant vector and $\mathbf{r} = (x, y, z)$; C is the straight line segment from \mathbf{r}_1 to \mathbf{r}_2
- (vii) $\mathbf{F} = (y \sin z, z \sin x, x \sin y)$ and C is the parametric curve $\mathbf{r} = (\cos t, \sin t, \sin(5t))$ for $t \in [0, \pi]$
- (viii) $\mathbf{F} = (y, -xz, y(x^2 + z^2))$ and C is the intersection of the cylinder $x^2 + z^2 = 1$ with the plane $x + y + z = 1$ that is oriented counterclockwise as viewed from the top of the y axis
- (ix) $\mathbf{F} = (-y \sin(\pi z^2), x \cos(\pi z^2), e^{xyz})$ and C is the intersection of the cone $z = \sqrt{x^2 + y^2}$ and the sphere $x^2 + y^2 + z^2 = 2$; C is oriented counterclockwise as viewed from the top of the z axis
- (x) $\mathbf{F} = (e^{\sqrt{y}}, e^x, 0)$ and C is the parabola in the xy plane from the origin to the point $(1, 1)$
- (xi) $\mathbf{F} = (x, y, z)$ and C is an elliptic helix $\mathbf{r}(t) = (a \cos t, b \sin t, ct)$, $0 \leq t \leq 2\pi$
- (xii) $\mathbf{F} = (y^{-1}, z^{-1}, x^{-1})$ and C is the straight line segment from the point $(1, 1, 1)$ to the point $(2, 4, 8)$
- (xiii) $\mathbf{F} = (e^{y-z}, e^{z-x}, e^{x-y})$ and C is the straight line segment from the origin to the point $(1, 3, 5)$
- (xiv) $\mathbf{F} = (y+z, 2+x, x+y)$ and C is the shortest arc on the sphere $x^2 + y^2 + z^2 = 25$ from the point $(3, 4, 0)$ to the point $(0, 0, 5)$
- (5) Find the work done by the constant force \mathbf{F} in moving a point object along a smooth path from a point \mathbf{r}_a to a point \mathbf{r}_b .
- (6) Find the work done by the force $\mathbf{F} = f'(r)\mathbf{r}/r$ in moving a point object along a smooth path from a point \mathbf{r}_a to a point \mathbf{r}_b , where the derivative f' of f is a continuous function of $r = \|\mathbf{r}\|$.
- (7) Find the work done by the force $\mathbf{F} = (-y, x, c)$, where c is a constant, in moving a point object along:
- (i) The circle $x^2 + y^2 = 1, z = 0$
- (ii) The circle $(x - 2)^2 + y^2 = 1, z = 0$

(8) The force acting on a charged particle that moves in a magnetic field \mathbf{B} and an electric field \mathbf{E} is $\mathbf{F} = e\mathbf{E} + (e/c)\mathbf{v} \times \mathbf{B}$, where \mathbf{v} is the velocity of the particle, e is its electric charge, and c is the speed of light in a vacuum. Find the work done by the force along a trajectory originating from a point \mathbf{r}_a and ending at the point \mathbf{r}_b if

- (i) The electric and magnetic fields are constant.
- (ii) The electric field vanishes, but the magnetic field is a continuous function of the position vector, $\mathbf{B} = \mathbf{B}(\mathbf{r})$.

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 13.14) 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. Continuous 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

$$(15.4) \quad \int_C \nabla f \cdot d\mathbf{r} = f(\mathbf{r}_b) - f(\mathbf{r}_a).$$

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 (15.1) 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 partial derivatives of f and $\mathbf{r}'(t)$ for a smooth curve. \square

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 continuous 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 $\mathbf{F} = \nabla f$:

$$\mathbf{F} = \nabla f \iff \oint_C \mathbf{F} \cdot d\mathbf{r} = 0.$$

The symbol \oint_C is often used to denote line integrals along a closed curve.

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

$$(15.5) \quad f(\mathbf{r}) = \int_C \mathbf{F} \cdot d\mathbf{r}$$

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 (15.4), 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 (15.3), the line integrals along C_1 and C_2 coincide. To calculate the derivative $f'_x(\mathbf{r}) = \lim_{h \rightarrow 0} (f(\mathbf{r} + h\hat{\mathbf{e}}_1) - f(\mathbf{r}))/h$, 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 $x \leq t \leq x + h$. Therefore, $\mathbf{r}'(t) = \hat{\mathbf{e}}_1$ and $\mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = F_1(t, y, z)$. Thus,

$$\begin{aligned} f'_x(\mathbf{r}) &= \lim_{h \rightarrow 0} \frac{1}{h} \int_x^{x+h} F_1(t, y, z) dt = \lim_{h \rightarrow 0} \frac{1}{h} \left(\int_a^{x+h} - \int_a^x \right) F_1(t, y, z) dt \\ &= \frac{\partial}{\partial x} \int_a^x F_1(t, y, z) dt = F_1(x, y, z) = F_1(\mathbf{r}) \end{aligned}$$

by the continuity of F_1 . The equalities $f'_y = F_2$ and $f'_z = F_3$ are established similarly. The details are omitted. \square

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 (15.5) gives a practical method of finding a potential if the vector field is found to be conservative (see Study Problem 15.3).

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 $\mathbf{sa} = (sa_1, sa_2, sa_3)$. 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 :

$$\nabla f = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right).$$

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

$$\text{curl } \mathbf{F} = \nabla \times \mathbf{F}.$$

The curl of a vector field is a vector field whose components can be computed according to the definition of the cross product:

$$\begin{aligned} \nabla \times \mathbf{F} &= \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_1 & F_2 & F_3 \end{pmatrix} \\ &= \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z} \right) \hat{\mathbf{e}}_1 + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x} \right) \hat{\mathbf{e}}_2 + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) \hat{\mathbf{e}}_3. \end{aligned}$$

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.

EXAMPLE 15.3. Find the curl of the vector field $\mathbf{F} = (yz, xyz, x^2)$.

SOLUTION:

$$\begin{aligned} \nabla \times \mathbf{F} &= \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ yz & xyz & x^2 \end{pmatrix} \\ &= \left((x^2)'_y - (xyz)'_z, -(x^2)'_x + (yz)'_z, (xyz)'_x - (yz)'_y \right) \\ &= (-xy, y - 2x, yz - z). \end{aligned}$$

□

The geometrical significance of the curl of a vector field will be discussed in Section 114.4. Here the curl is used to formulate sufficient conditions for a vector field to be conservative.

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.*

111.4. Test for a Vector Field to Be Conservative. A conservative vector field with continuous partial derivatives in a region E has been shown to have the vanishing curl:

$$\mathbf{F} = \nabla f \quad \implies \quad \operatorname{curl} \mathbf{F} = \mathbf{0}.$$

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 nonintersecting (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.

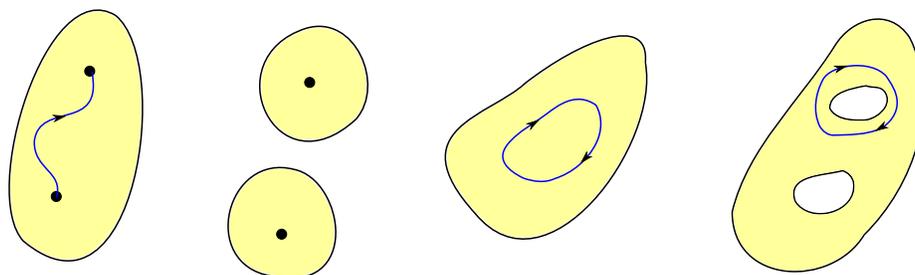


FIGURE 15.3. From left to right: A planar connected region (any two points in it can be connected by a continuous curve that lies in the region); a planar disconnected region (there are points in it that cannot be connected by a continuous curve that lies in the region); a planar simply connected region (every simple closed curve in it can be continuously shrunk to a point in it while remaining in the region throughout the deformation); a planar region that is not simply connected (it has holes).

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 \mathbf{F} is a vector field whose components have continuous partial derivatives 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 :

$$\operatorname{curl} \mathbf{F} = \mathbf{0} \text{ on simply connected } E \iff \mathbf{F} = \nabla f \text{ on } E.$$

This theorem follows from Stokes’ theorem discussed later in Section 114 and has two useful consequences. First, *the test for the path independence of line integrals*:

$$\operatorname{curl} \mathbf{F} = \mathbf{0} \text{ on simply connected } E \iff \int_{C_1} \mathbf{F} \cdot d\mathbf{r} = \int_{C_2} \mathbf{F} \cdot d\mathbf{r}$$

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 $\mathbf{r}_b \in E$. It follows from Theorem 15.3

for the curve C that is the union of C_1 and $-C_2$. Second, the test for vanishing line integrals along closed paths:

$$\operatorname{curl} \mathbf{F} = \mathbf{0} \text{ on simply connected } E \iff \oint_C \mathbf{F} \cdot d\mathbf{r} = 0,$$

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 Study Problem 15.2.

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 , which 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 in values of its potential U (also called *potential energy*) at the initial and terminal points of the path. More generally, since the work done by a force equals the change in kinetic energy (see Section 73.5), the motion under a conservative force $\mathbf{F} = -\nabla U$ has the fundamental property that *the sum of kinetic and potential energies, $m\mathbf{v}^2/2 + U(\mathbf{r})$, is conserved along a trajectory of the motion* (recall Study Problem 13.14).

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 \mathbf{F} is conservative before evaluating the line integral using the parametric equations of C . First, note that the components of \mathbf{F} are polynomials and hence have continuous partial derivatives in the entire space. Therefore, if its curl vanishes, then \mathbf{F} is conservative in the entire space by Theorem 15.4 as the entire space is simply connected:

$$\begin{aligned} \nabla \times \mathbf{F} &= \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_1 & F_2 & F_3 \end{pmatrix} = \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ yz & xz+z+2y & xy+y+2z \end{pmatrix} \\ &= ((F_3)'_y - (F_2)'_z, -(F_3)'_x + (F_1)'_z, (F_2)'_x - (F_1)'_y) \\ &= (x+1 - (x+1), -y+y, z-z) = (0, 0, 0). \end{aligned}$$

Thus, \mathbf{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 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 \mathbf{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,

$$f'_x = F_1 = yz \implies f(x, y, z) = xyz + g(y, z),$$

where $g(y, z)$ is arbitrary. The substitution of f into the second equation $f'_y = F_2$ yields

$$xz + g'_y(y, z) = xz + z + 2y \implies g(y, z) = y^2 + zy + h(z),$$

where $h(z)$ is arbitrary. The substitution of $f = xyz + y^2 + zy + h(z)$ into the third equation $f'_z = F_3$ yields

$$xy + y + h'(z) = xy + y + 2z \implies h(z) = z^2 + c,$$

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. \square

111.5. Study Problems.

Problem 15.2. *Verify that*

$$\mathbf{F} = \nabla f = \left(-\frac{y}{x^2 + y^2}, \frac{x}{x^2 + y^2}, 2z \right), \quad f(x, y, z) = \tan^{-1}(y/x) + z^2,$$

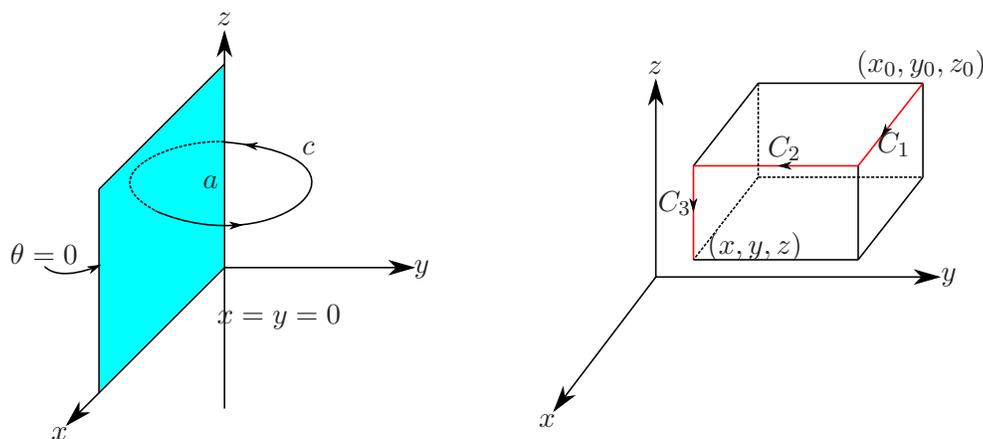


FIGURE 15.4. **Left:** An illustration to Study Problem 15.2. **Right:** An illustration to Study Problem 15.3. To find the potential of a conservative vector field, one can evaluate its line integral from any point (x_0, y_0, z_0) to a generic point (x, y, z) along the rectangular contour C that is the union of the straight line segments C_1 , C_2 , and C_3 parallel to the coordinate axes.

and $\text{curl } \mathbf{F} = \mathbf{0}$ 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 the fundamental theorem for line integrals? Explain.

SOLUTION: A straightforward differentiation of f shows that indeed $\nabla f = \mathbf{F}$, and therefore $\text{curl } \mathbf{F} = \mathbf{0}$ 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 $\text{curl } \mathbf{F} = \mathbf{0}$ 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. \square

Problem 15.3. *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 (15.5), assume C consists of three straight line segments, $(x_0, y_0, z_0) \rightarrow (x, y_0, z_0) \rightarrow (x, y, z_0) \rightarrow (x, y, z)$, as depicted in the right panel of Figure 15.4. The parametric equation of the first line C_1 is $\mathbf{r}(t) = (t, y_0, z_0)$, where $x_0 \leq t \leq x$. 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 $y_0 \leq t \leq y$, 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 $z_0 \leq t \leq z$. 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. \square

Problem 15.4. (Operator ∇ in Curvilinear Coordinates).

Let the transformation $(u, v, w) \rightarrow (x, y, z)$ be a change of variables. If $\hat{\mathbf{e}}_u$, $\hat{\mathbf{e}}_v$, and $\hat{\mathbf{e}}_w$ are unit vectors normal to the coordinate surfaces (see exercise 14 in Section 105.4), show that

$$\nabla = \|\nabla u\| \hat{\mathbf{e}}_u \frac{\partial}{\partial u} + \|\nabla v\| \hat{\mathbf{e}}_v \frac{\partial}{\partial v} + \|\nabla w\| \hat{\mathbf{e}}_w \frac{\partial}{\partial w}.$$

In particular, find the ∇ operator in the cylindrical and spherical coordinates.

SOLUTION: By the chain rule,

$$\frac{\partial}{\partial x} = \frac{\partial u}{\partial x} \frac{\partial}{\partial u} + \frac{\partial v}{\partial x} \frac{\partial}{\partial v} + \frac{\partial w}{\partial x} \frac{\partial}{\partial w}$$

and similarly for $\partial/\partial y$ and $\partial/\partial z$. Then

$$\begin{aligned} \nabla &= \hat{\mathbf{e}}_1 \frac{\partial}{\partial x} + \hat{\mathbf{e}}_2 \frac{\partial}{\partial y} + \hat{\mathbf{e}}_3 \frac{\partial}{\partial z} \\ &= \left(\frac{\partial u}{\partial x} \hat{\mathbf{e}}_1 + \frac{\partial u}{\partial y} \hat{\mathbf{e}}_2 + \frac{\partial u}{\partial z} \hat{\mathbf{e}}_3 \right) \frac{\partial}{\partial u} + \left(\frac{\partial v}{\partial x} \hat{\mathbf{e}}_1 + \frac{\partial v}{\partial y} \hat{\mathbf{e}}_2 + \frac{\partial v}{\partial z} \hat{\mathbf{e}}_3 \right) \frac{\partial}{\partial v} \\ &\quad + \left(\frac{\partial w}{\partial x} \hat{\mathbf{e}}_1 + \frac{\partial w}{\partial y} \hat{\mathbf{e}}_2 + \frac{\partial w}{\partial z} \hat{\mathbf{e}}_3 \right) \frac{\partial}{\partial w} \\ &= \nabla u \frac{\partial}{\partial u} + \nabla v \frac{\partial}{\partial v} + \nabla w \frac{\partial}{\partial w} \\ &= \|\nabla u\| \hat{\mathbf{e}}_u \frac{\partial}{\partial u} + \|\nabla v\| \hat{\mathbf{e}}_v \frac{\partial}{\partial v} + \|\nabla w\| \hat{\mathbf{e}}_w \frac{\partial}{\partial w}, \end{aligned}$$

where the unit vectors are defined in (14.23). Making use of (14.24), (14.25), (14.26), and (14.27), the operator ∇ is obtained in the cylindrical and spherical coordinates:

$$\begin{aligned} \nabla &= \hat{\mathbf{e}}_r \frac{\partial}{\partial r} + \frac{1}{r} \hat{\mathbf{e}}_\theta \frac{\partial}{\partial \theta} + \hat{\mathbf{e}}_3 \frac{\partial}{\partial z}, \\ \nabla &= \hat{\mathbf{e}}_\rho \frac{\partial}{\partial \rho} + \frac{1}{\rho} \hat{\mathbf{e}}_\phi \frac{\partial}{\partial \phi} + \frac{1}{\rho \sin \phi} \hat{\mathbf{e}}_\theta \frac{\partial}{\partial \theta}. \end{aligned}$$

\square

111.6. Exercises.

(1) Calculate the curl of the given vector field:

(i) $\mathbf{F} = (xyz, -y^2x, 0)$

(ii) $\mathbf{F} = (\cos(xz), \sin(yz), 2)$

(iii) $\mathbf{F} = (h(x), g(y), h(z))$, where the functions h , g , and h are differentiable

(iv) $\mathbf{F} = (\ln(xyz), \ln(yz), \ln z)$

(v) $\mathbf{F} = \mathbf{a} \times \mathbf{r}$, where \mathbf{a} is a constant vector and $\mathbf{r} = (x, y, z)$

(2) Suppose that a vector field $\mathbf{F}(\mathbf{r})$ and a function $f(\mathbf{r})$ are differentiable. Show that $\nabla \times (f\mathbf{F}) = f(\nabla \times \mathbf{F}) + \nabla f \times \mathbf{F}$.

(3) Find $\nabla(\mathbf{c} \times \mathbf{r}f(r))$, where $r = \|\mathbf{r}\|$, f is differentiable, and \mathbf{c} is a constant vector.

(4) A fluid, filling the entire space, rotates at a constant rate ω about an axis parallel to a unit vector $\hat{\mathbf{n}}$. Find the curl of the velocity vector field at a generic point \mathbf{r} . Assume that the position vector \mathbf{r} originates from a point on the axis of rotation.

(5) Determine whether the vector field is conservative and, if it is, find its potential:

(i) $\mathbf{F} = (2xy, x^2 + 2yz^3, 3z^2y^2 + 1)$

(ii) $\mathbf{F} = (yz, xz + 2y \cos z, xy - y^2 \sin z)$

(iii) $\mathbf{F} = (e^y, xe^y - z^2, -2yz)$

(iv) $\mathbf{F} = (6xy + z^4y, 3x^2 + z^4x, 4z^3xy)$

(v) $\mathbf{F} = (yz(2x + y + z), xz(x + 2y + z), xy(x + y + 2z))$

(vi) $\mathbf{F} = (-y(x^2 + y^2)^{-1} + z, x(x^2 + y^2)^{-1}, x)$

(vii) $\mathbf{F} = (y \cos(xy), x \cos(xy), z + y)$

(viii) $\mathbf{F} = (-yz/x^2, z/x, y/x)$

(6) Determine first whether the vector field is conservative and then evaluate the line integral $\int_C \mathbf{F} \cdot d\mathbf{r}$:

(i) $\mathbf{F}(x, y, z) = (y^2z^2 + 2x + 2y, 2xyz^2 + 2x, 2xy^2z + 1)$ and C consists of three line segments: $(1, 1, 1) \rightarrow (a, b, c) \rightarrow (1, 2, 3)$

(ii) $\mathbf{F} = (zx, yz, z^2)$ and C is the part of the helix $\mathbf{r}(t) = (2 \sin t, -2 \cos t, t)$ that lies inside the ellipsoid $x^2 + y^2 + 2z^2 = 6$

(iii) $\mathbf{F} = (y - z^2, x + \sin z, y \cos z - 2xz)$ and C is one turn of a helix of radius a from $(a, 0, 0)$ to $(a, 0, b)$.

(iv) $\mathbf{F} = g(r^2)\mathbf{r}$, where $\mathbf{r} = (x, y, z)$, $r = \|\mathbf{r}\|$, g is continuous, and C is a smooth curve from a point on the sphere $x^2 + y^2 + z^2 = a^2$ to a point on the sphere $x^2 + y^2 + z^2 = b^2$. What is the work done by the force \mathbf{F} if $g = -1/r^3$?

(v) $\mathbf{F} = (2(y+z)^{1/2}, -x(y+z)^{3/2}, -x(y+z)^{-3/2})$ and C is a smooth curve from the point $(1, 1, 3)$ and $(2, 4, 5)$

(7) Suppose that \mathbf{F} and \mathbf{G} are continuous on E . Show that $\oint_C \mathbf{F} \cdot d\mathbf{r} = \oint_C \mathbf{G} \cdot d\mathbf{r}$ for any smooth closed curve C in E if there is a function f with continuous partial derivatives in E such that $\mathbf{F} - \mathbf{G} = \nabla f$.

(8) Use the properties of the gradient to show that the vectors $\hat{\mathbf{e}}_r = (\cos \theta, \sin \theta)$ and $\hat{\mathbf{e}}_\theta = (-\sin \theta, \cos \theta)$ are unit vectors orthogonal to the coordinate curves $r(x, y) = \text{const}$ and $\theta(x, y) = \text{const}$ of polar coordinates. Given a planar vector field, put $\mathbf{F} = F_r \hat{\mathbf{e}}_r + F_\theta \hat{\mathbf{e}}_\theta$. Use the chain rule to express the curl of a planar vector field $\mathbf{F}(r, \theta)$ in polar coordinates (r, θ) as a linear combination of $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$.

(9) Evaluate the pairwise cross products of the unit vectors (14.27) and the pairwise cross products of the unit vectors (14.26). Use the obtained relations and the result of Study Problem 15.4 to express the curl of a vector field in spherical and cylindrical coordinates:

$$\begin{aligned} \nabla \times \mathbf{F} &= \frac{1}{\rho \sin \phi} \left(\frac{\partial(\sin \phi F_\theta)}{\partial \phi} - \frac{\partial F_\phi}{\partial \theta} \right) \hat{\mathbf{e}}_\rho \\ &\quad + \frac{1}{\rho} \left(\frac{1}{\sin \phi} \frac{\partial F_\rho}{\partial \theta} - \frac{\partial(\rho F_\theta)}{\partial \rho} \right) \hat{\mathbf{e}}_\phi + \frac{1}{\rho} \left(\frac{\partial(\rho F_\phi)}{\partial \rho} - \frac{\partial F_\rho}{\partial \phi} \right) \hat{\mathbf{e}}_\theta, \\ \nabla \times \mathbf{F} &= \left(\frac{1}{r} \frac{\partial F_z}{\partial \theta} - \frac{\partial F_\theta}{\partial z} \right) \hat{\mathbf{e}}_r + \left(\frac{\partial F_r}{\partial z} - \frac{\partial F_z}{\partial r} \right) \hat{\mathbf{e}}_\theta \\ &\quad + \frac{1}{r} \left(\frac{\partial(r F_\theta)}{\partial r} - \frac{\partial F_r}{\partial \theta} \right) \hat{\mathbf{e}}_z, \end{aligned}$$

where the field \mathbf{F} is decomposed over the bases (14.27) and (14.26): $\mathbf{F} = F_\rho \hat{\mathbf{e}}_\rho + F_\phi \hat{\mathbf{e}}_\phi + F_\theta \hat{\mathbf{e}}_\theta$ and $\mathbf{F} = F_r \hat{\mathbf{e}}_r + F_\theta \hat{\mathbf{e}}_\theta + F_z \hat{\mathbf{e}}_z$. *Hint:* Show $\partial \hat{\mathbf{e}}_\rho / \partial \phi = \hat{\mathbf{e}}_\phi$, $\partial \hat{\mathbf{e}}_\rho / \partial \theta = \sin \theta \hat{\mathbf{e}}_\theta$, and similar relations for the partial derivatives of other unit vectors.

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 (see the left panel of Figure 15.5).

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,

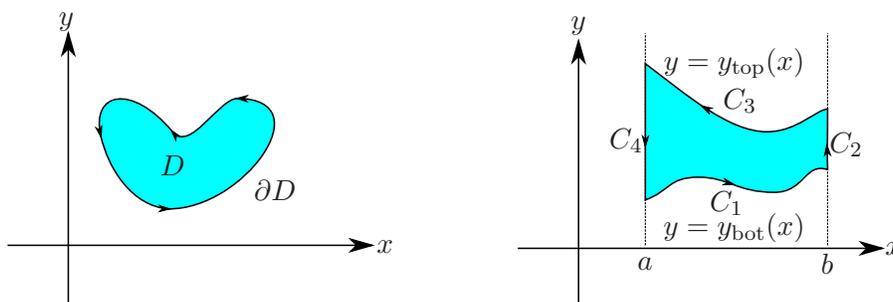


FIGURE 15.5. **Left:** A simple closed planar curve encloses a connected region D in the plane. The positive orientation of the boundary of D means that the boundary curve ∂D is traversed counterclockwise. **Right:** A simple region D is bounded by four smooth curves: two graphs C_1 and C_3 and two vertical lines C_2 ($x = b$) and C_4 ($x = a$). The boundary ∂D is the union of these curves oriented counterclockwise.

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 in an open region that contains D , then

$$\iint_D \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dA = \oint_{\partial D} F_1 dx + F_2 dy.$$

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:

$$(15.6) \quad D = \{(x, y) \mid y_{\text{bot}}(x) \leq y \leq y_{\text{top}}(x), x \in [a, b]\},$$

$$(15.7) \quad D = \{(x, y) \mid x_{\text{bot}}(y) \leq x \leq x_{\text{top}}(y), y \in [c, d]\}.$$

The idea of the proof is to establish the equalities

$$(15.8) \quad \oint_{\partial D} F_1 dx = - \iint_D \frac{\partial F_1}{\partial y} dA, \quad \oint_{\partial D} F_2 dy = \iint_D \frac{\partial F_2}{\partial x} dA$$

using, respectively, (15.6) and (15.7). 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 (see the right panel of Figure 15.5). The curve C_1 is the graph $y = y_{\text{bot}}(x)$ whose parametric equations are $\mathbf{r} = (t, y_{\text{bot}}(t))$, 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 $y = y_{\text{top}}(x)$, and, similarly, its parametric equations $\mathbf{r}(t) = (t, y_{\text{top}}(t))$, where $t \in [a, b]$. This vector function traverses C_3 from left to right. So the orientation of C_3 must be reversed to obtain the corresponding part of ∂D . 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 $y = y_{\text{bot}}(x)$ and $y = y_{\text{top}}(x)$ 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,

$$\begin{aligned} \oint_{\partial D} F_1 dx &= \int_{C_1} F_1 dx + \int_{-C_2} F_1 dx \\ &= \int_a^b \left(F(x, y_{\text{bot}}(x)) - F(x, y_{\text{top}}(x)) \right) dx, \end{aligned}$$

where the property (15.3) has been used. Next, the double integral is transformed into an ordinary integral by converting it to an iterated integral:

$$\begin{aligned} \iint_D \frac{\partial F_1}{\partial y} dA &= \int_a^b \int_{y_{\text{bot}}(x)}^{y_{\text{top}}(x)} \frac{\partial F_1}{\partial y} dy dx \\ &= \int_a^b \left(F(x, y_{\text{top}}(x)) - F(x, y_{\text{bot}}(x)) \right) dx, \end{aligned}$$

where the latter equality follows from the fundamental theorem of calculus and the continuity of F_1 on an open interval that contains $[y_{\text{bot}}(x), y_{\text{top}}(x)]$ 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 (15.8) is established. The

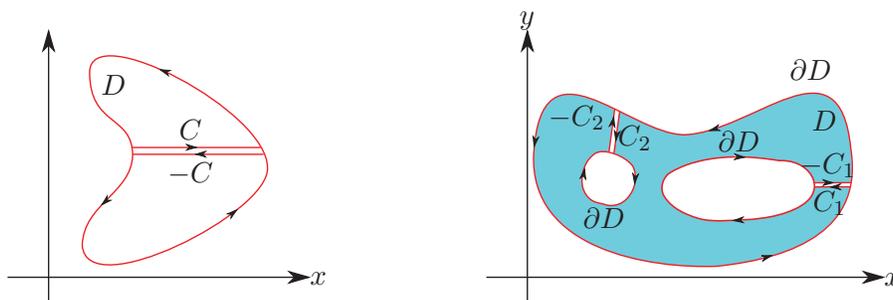


FIGURE 15.6. **Left:** A region D is split into two regions by a curve C . If the boundary of the upper part of D has positive orientation, then the positively oriented boundary of the lower part of D has the curve $-C$. **Right:** Green's theorem holds for nonsimply connected regions. The orientation of the boundaries of "holes" in D is obtained by making cuts along curves C_1 and C_2 so that D becomes simply connected. The positive orientation of the outer boundary of D induces the orientation of the boundaries of the "holes."

second equality in (15.8) is proved analogously by using (15.7). The details are omitted. \square

Suppose that a smooth, oriented curve C divides a region D into two *simple* regions D_1 and D_2 (see the left panel of Figure 15.6). 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

$$\begin{aligned} \oint_{\partial D} \mathbf{F} \cdot d\mathbf{r} &= \oint_{\partial D_1} \mathbf{F} \cdot d\mathbf{r} + \oint_{\partial D_2} \mathbf{F} \cdot d\mathbf{r} \\ &= \iint_{D_1} \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dA + \iint_{D_2} \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dA \\ &= \iint_D \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dA. \end{aligned}$$

The first equality holds because of the cancellation of the line integrals along C and $-C$ according to (15.3). 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.

Green's Theorem for Nonsimply Connected Regions. 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 (see the right panel of Figure 15.6). 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

$$\begin{aligned}
 \iint_D \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dA &= \iint_{D_1} \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dA - \iint_{D_2} \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dA \\
 &= \oint_{\partial D_1} \mathbf{F} \cdot d\mathbf{r} - \oint_{\partial D_2} \mathbf{F} \cdot d\mathbf{r} \\
 &= \oint_{\partial D_1} \mathbf{F} \cdot d\mathbf{r} + \oint_{-\partial D_2} \mathbf{F} \cdot d\mathbf{r} \\
 (15.9) \qquad &= \oint_{\partial D} \mathbf{F} \cdot d\mathbf{r}.
 \end{aligned}$$

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 Green's theorem 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 connect 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 (15.9) follows from the additivity of the double integral.

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.



FIGURE 15.7. **Left:** The integration curve in the line integral discussed in Example 15.5. **Right:** A general polygon. Its area is evaluated in Example 15.7 by representing the area via a line integral.

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 $[1, 2]$, as shown in the left panel of Figure 15.7. 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 to appreciate the following alternative 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 antiderivatives are not expressible in elementary functions).

SOLUTION: The curve C is a simple, piecewise-smooth, closed curve, and the components of \mathbf{F} have continuous partial derivatives everywhere. 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,

$$\begin{aligned} \oint_C \mathbf{F} \cdot d\mathbf{r} &= - \oint_{\partial D} \mathbf{F} \cdot d\mathbf{r} = - \iint_D \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dA = - \iint_D y \, dA \\ &= - \int_0^\pi \int_1^2 r \sin \theta \, r \, dr \, d\theta = - \int_0^\pi \sin \theta \, d\theta \int_1^2 r^2 \, dr = -\frac{14}{3}, \end{aligned}$$

where the double integral has been transformed to polar coordinates. The region D is the image of the rectangle $D' = [1, 2] \times [0, \pi]$ in the polar plane under the transformation $(r, \theta) \rightarrow (x, y)$. \square

Changing the Integration Curve in a Line Integral. If a planar vector field is not conservative, then its line integral along a curve C originating from a point A and terminating at a point B depends on C . If C' is another curve outgoing from A and terminating at B , what is the relation between the line integrals of \mathbf{F} over C and C' ? Green's theorem

allows us to establish such a relation. Suppose that C and C' have no self-intersections and do not intersect. Then their union is a boundary of a simply connected region D . Let us reverse the orientation of one of the curves so that their union is the positively oriented boundary ∂D , where ∂D is the union of C and $-C'$. Then

$$\oint_{\partial D} \mathbf{F} \cdot d\mathbf{r} = \int_C \mathbf{F} \cdot d\mathbf{r} + \int_{-C'} \mathbf{F} \cdot d\mathbf{r} = \int_C \mathbf{F} \cdot d\mathbf{r} - \int_{C'} \mathbf{F} \cdot d\mathbf{r}.$$

By Green's theorem,

$$(15.10) \quad \int_C \mathbf{F} \cdot d\mathbf{r} = \int_{C'} \mathbf{F} \cdot d\mathbf{r} + \iint_D \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dA,$$

which establishes the relation between line integrals of a nonconservative planar vector field over two different curves that have common endpoints.

EXAMPLE 15.6. Evaluate the line integral of the vector field $\mathbf{F} = (2y + \cos(x^2), x^2 + y^3)$ over the curve C , which consists of the line segments $(0, 0) \rightarrow (1, 1)$ and $(1, 1) \rightarrow (0, 2)$.

SOLUTION: Let C' be the line segment $(0, 0) \rightarrow (0, 2)$. Then the union of C and $-C'$ is the boundary ∂D (positively oriented) of the triangular region D with vertices $(0, 0)$, $(1, 1)$, and $(0, 2)$. The relation (15.10) can be applied to evaluate the line integral over C . The parametric equations of C' are $x = 0$, $y = t$, $0 \leq t \leq 2$. Hence, along C' , $\mathbf{F} \cdot d\mathbf{r} = F_2(0, t) dt = t^3 dt$ and

$$\int_{C'} \mathbf{F} \cdot d\mathbf{r} = \int_0^2 t^3 dt = 4.$$

Then $\partial F_2/\partial x = 2x$ and $\partial F_1/\partial y = 2$. The region D admits an algebraic description as a vertically simple region: $x \leq y \leq 2 - x$, $0 \leq x \leq 1$. Hence,

$$\begin{aligned} \iint_D \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dA &= \iint_D (2x - 2) dA = 2 \int_0^1 (x - 1) \int_x^{2-x} dy \\ &= -4 \int_0^1 (x - 1)^2 dx = \frac{4}{3}. \end{aligned}$$

Therefore, by (15.10),

$$\int_C \mathbf{F} \cdot d\mathbf{r} = 4 + \frac{4}{3} = \frac{16}{3}.$$

□

112.2. Area of a Planar Region as a Line Integral. Put $F_2 = x$ and $F_1 = 0$. Then

$$\iint_D \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dA = \iint_D dA = A(D).$$

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:

$$(15.11) \quad A(D) = \oint_{\partial D} x \, dy = - \oint_{\partial D} y \, dx = \frac{1}{2} \oint_{\partial D} x \, dy - y \, dx,$$

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 the gradient of a function whose line integral along a closed curve vanishes owing to the fundamental theorem for line integrals. 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 (15.11) 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 15.7. (Area of a Polygon).

Consider an arbitrary polygon whose vertices in counterclockwise order are $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. 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 as shown in the right panel of Figure 15.7. If C_i is such a segment oriented from (x_i, y_i) to (x_{i+1}, y_{i+1}) for $i = 1, 2, \dots, n-1$, 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

$$\begin{aligned} A &= \frac{1}{2} \oint_{\partial D} x dy - y dx = \frac{1}{2} \sum_{i=1}^n \int_{C_i} x dy - y dx \\ &= \frac{1}{2} \sum_{i=1}^{n-1} \int_0^1 (x_i y_{i+1} - y_i x_{i+1}) dt + \frac{1}{2} \int_0^1 (x_n y_1 - y_n x_1) dt \\ &= \frac{1}{2} \left(\sum_{i=1}^{n-1} (x_i y_{i+1} - y_i x_{i+1}) + (x_n y_1 - y_n x_1) \right). \end{aligned}$$

□

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:

$$\nabla \times \mathbf{F} = \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_1(x, y) & F_2(x, y) & 0 \end{pmatrix} = \hat{\mathbf{e}}_3 \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right).$$

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 $C = \partial D_s$,

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = \iint_{D_s} \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dA = \iint_{D_s} 0 dA = 0$$

for any closed simple curve C in D . By the pathindependence property, the vector field \mathbf{F} is conservative in D .

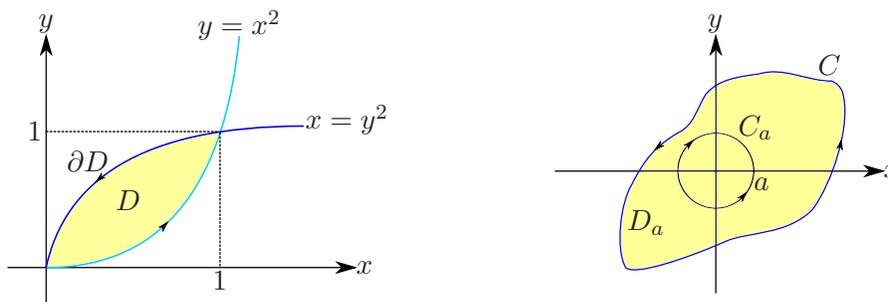


FIGURE 15.8. **Left:** An illustration to Study Problem 15.5. **Right:** An illustration to Study Problem 15.6. The region D_a is bounded by a curve C and the circle C_a .

112.4. Study Problems.

Problem 15.5. 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,

$$\oint_{\partial D} \mathbf{F} \cdot d\mathbf{r} = \iint_D 2 \, dA = 2 \int_0^1 \int_{x^2}^{\sqrt{x}} dy \, dx = 2 \int_0^1 (\sqrt{x} - x^2) \, dx = \frac{2}{3}.$$

The integration region D is shown in the left panel of Figure 5.8. \square

Problem 15.6. Prove that the line integral of the planar vector field

$$\mathbf{F} = \left(-\frac{y}{x^2 + y^2}, \frac{x}{x^2 + y^2} \right)$$

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 15.2) 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 15.2). \square

Problem 15.7. (Volume of Axially Symmetric Solids).

Let D be a region in the upper part of the xy plane ($y \geq 0$). Consider the solid E obtained by rotation of D about the x axis. Show that the volume of the solid is given by

$$V(E) = -\pi \oint_{\partial D} y^2 dx.$$

SOLUTION: Let dA be the area of a partition element of D that contains a point (x, y) . If the partition element is rotated about the x axis, the point (x, y) traverses the circle of radius y (the distance from the point (x, y) to the x axis). The length of the circle is $2\pi y$. Consequently, the volume of the solid ring swept by the partition element is $dV = 2\pi y dA$. Taking the sum over the partition of D , the volume is expressed via the double integral over D :

$$V(E) = 2\pi \iint_D y dA.$$

In Green's theorem, put $\partial F_1/\partial y = 2y$ and $\partial F_2/\partial x = 0$ so that the above double integral is proportional to the left side of Green's equation. In particular, $F_1 = y^2$ and $F_2 = 0$ satisfy these conditions. By Green's theorem,

$$V(E) = \pi \iint_D \frac{\partial F_1}{\partial y} dA = -\pi \oint_{\partial D} F_1 dx = -\pi \oint_{\partial D} y^2 dx$$

as required. \square

112.5. Exercises.

(1) Evaluate the line integral by two methods: (a) directly and (b) using Green's theorem:

- (i) $\oint_C xy^2 dx - y^2x dy$, where C is the triangle with vertices $(0, 0)$, $(1, 0)$, and $(1, 2)$; C is oriented counterclockwise
- (ii) $\oint_C 2yx dx + x^2 dy$, where C consists of the line segments from $(0, 1)$ to $(0, 0)$ and from $(0, 0)$ to $(1, 0)$ and the parabola $y = 1 - x^2$ from $(1, 0)$ to $(0, 1)$

(2) Evaluate the line integral using Green's theorem:

- (i) $\oint x \sin(x^2) dx + (xy^2 - x^8) dy$, where C is the positively oriented boundary of the region between two circles $x^2 + y^2 = 1$ and $x^2 + y^2 = 4$
- (ii) $\oint_C (y^3 dx - x^3 dy)$, where C is the positively oriented circle $x^2 + y^2 = a^2$
- (iii) $\oint_C (\sqrt{x} + y^3) dx + (x^2 + \sqrt{y}) dy$, where C consists of the arc of the curve $y = \cos x$ from $(-\pi/2, 0)$ to $(\pi/2, 0)$ and the line segment from $(\pi/2, 0)$ to $(-\pi/2, 0)$
- (iv) $\oint_C (y^4 - \ln(x^2 + y^2)) dx + 2 \tan^{-1}(y/x) dy$, where C is the positively oriented circle of radius $a > 0$ with center (x_0, y_0) such that $x_0 > a$ and $y_0 > a$
- (v) $\oint_C (x + y)^2 dx - (x^2 + y^2) dy$, where C is a positively oriented triangle with vertices $(1, 1)$, $(3, 2)$, and $(2, 5)$
- (vi) $\oint_C xy^2 dx - x^2y dy$, where C is the negatively oriented circle $x^2 + y^2 = a^2$
- (vii) $\oint_C (x + y) dx - (x - y) dy$, where C is the positively oriented ellipse $(x/a)^2 + (y/b)^2 = 1$
- (viii) $\oint_C e^x [(1 - \cos y) dx - (y - \sin y) dy]$, where C is the positively oriented boundary of the region $0 \leq x \leq \pi$, $0 \leq y \leq \sin x$
- (ix) $\oint_C e^{-x^2+y^2} [\cos(2xy) dx - \sin(2xy) dy]$, where C is the positively oriented circle $x^2 + y^2 = a^2$

(3) Use the contour transformation law (15.10) to solve the following problems:

- (i) Let $\mathbf{F} = ((x + y)^2, -(x - y)^2)$. Find the difference between the line integrals of \mathbf{F} over C_1 , which is the straight line segment $(1, 1) \rightarrow (2, 6)$, and over C_2 which is the parabola with the vertical axis and passing through $(1, 1)$, $(2, 6)$, and $(0, 0)$.
- (ii) $\int_C (e^x \sin y - qx) dx + (e^x \cos y - q) dy$, where q is a constant and C is the upper part of the circle $x^2 + y^2 = ax$ oriented from $(a, 0)$ to $(0, 0)$.
- (iii) $\int_C [g(y)e^x - qy] dx + [g'(y)e^x - q] dy$, where $g(y)$ and $g'(y)$ are continuous functions and C is a smooth curve from the point $P_1 = (x_1, y_1)$ to the point $P_2 = (x_2, y_2)$ such that it and the straight line segment P_1P_2 form a boundary of a region D of the area $A(D)$.

(4) Use Green's theorem to find the work done by the force $\mathbf{F} = (3xy^2 + y^3, y^4)$ in moving a particle along the circle $x^2 + y^2 = a^2$ from $(0, -a)$ to $(0, a)$ counterclockwise.

(5) Use a representation of the area of a planar region by the line integral to find the area of the specified region D :

- (i) D is bounded by an ellipse $x = a \cos t$, $y = b \sin t$, $0 \leq t \leq 2\pi$
(ii) D is under one arc of the cycloid $x = a(t - \sin t)$, $y = a(1 - \cos t)$
(iii) D is the *astroid* enclosed by the curve $x = a \cos^3 t$, $y = a \sin^3 t$
(iv) D is bounded by the curve $x(t) = a \cos^2 t$, $y(t) = b \sin(2t)$ for $t \in [0, \pi]$
(v) D is bounded by the parabola $(x + y)^2 = ax$ and by the x axis, $a > 0$
(vi) D is bounded by one loop of the curve $x^3 + y^3 = 3axy$, $a > 0$ (*Hint*: Put $y = tx$.)
(vii) D is bounded by the curve $(x^2 + y^2)^2 = a^2(x^2 - y^2)$ (*Hint*: Put $y = x \tan t$.)
(viii) D is bounded by $(x/a)^n + (y/b)^n = 1$, $n > 0$ (*Hint*: $x = a \cos^{n/2} t$, $y = b \sin^{n/2} t$.)

(6) Let a curve C have fixed endpoints. Under what condition on the function $g(x, y)$ is the line integral $\int_C g(x, y)(y dx + x dy)$ independent of C ?

(7) Let D be a planar region bounded by a simple closed curve. If A is the area of D , show that the coordinates (x_c, y_c) of the centroid of D are

$$x_c = \frac{1}{2A} \oint_{\partial D} x^2 dy, \quad y_c = -\frac{1}{2A} \oint_{\partial D} y^2 dx.$$

Hint: Use an approach similar to the derivation of (15.11).

(8) Let a lamina with a constant surface mass density σ occupy a planar region D enclosed by a simple piecewise smooth curve. Show that its moments of inertia about the x and y axes are

$$I_x = -\frac{\sigma}{3} \oint_{\partial D} y^3 dx, \quad I_y = \frac{\sigma}{3} \oint_{\partial D} x^3 dy.$$

Hint: Use an approach similar to the derivation of (15.11).

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-sectional

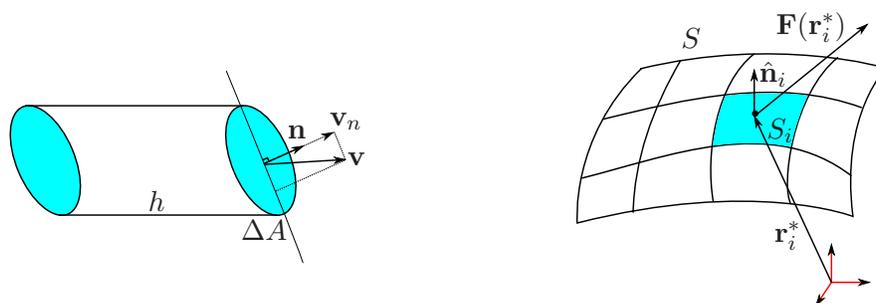


FIGURE 15.9. **Left:** A mass transferred by a homogeneous mass flow with a constant velocity \mathbf{v} across an area element ΔA in time Δt is $\Delta m = \sigma \Delta V$, where $\Delta V = h \Delta A$ is the volume of the cylinder with cross-sectional area ΔA and height $h = \Delta t v_n$; v_n is the scalar projection of \mathbf{v} onto the normal \mathbf{n} . **Right:** A partition of a smooth surface S by elements S_i . If \mathbf{r}_i^* is a sample point in S_i , $\hat{\mathbf{n}}_i$ is the unit normal to S at \mathbf{r}_i^* , and ΔS_i is the surface of the partition element, then the flux of a continuous vector field $\mathbf{F}(\mathbf{r})$ across S_i is approximated by $\Delta \Phi_i = \mathbf{F}(\mathbf{r}_i^*) \cdot \hat{\mathbf{n}}_i \Delta S_i$.

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 (see the left panel of Figure 15.9)

$$(15.12) \quad \Delta \Phi = \sigma v_n \Delta A = \sigma \mathbf{v} \cdot \hat{\mathbf{n}} \Delta A = \mathbf{F} \cdot \hat{\mathbf{n}} \Delta A = F_n \Delta A,$$

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 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 , $i = 1, 2, \dots, N$, whose surface areas are ΔS_i

as shown in the right panel of Figure 15.9. 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 of \mathbf{F} and the normal $\hat{\mathbf{n}}$ within S_i , the flux across S_i can be approximated by (15.12), $\Delta\Phi_i \approx \mathbf{F}(\mathbf{r}_i^*) \cdot \hat{\mathbf{n}}_i \Delta S_i$. The approximation becomes better when $N \rightarrow \infty$ so that the sizes of S_i decrease to 0 uniformly and hence the total flux is

$$\Phi = \lim_{N \rightarrow \infty} \sum_{i=1}^N \Delta\Phi_i = \lim_{N \rightarrow \infty} \sum_{i=1}^N \mathbf{F}(\mathbf{r}_i^*) \cdot \hat{\mathbf{n}}_i \Delta S_i = \lim_{N \rightarrow \infty} \sum_{i=1}^N F_n(\mathbf{r}_i^*) \Delta S_i.$$

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 $\hat{\mathbf{n}}$ as illustrated in the left panel of Figure 15.10. 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

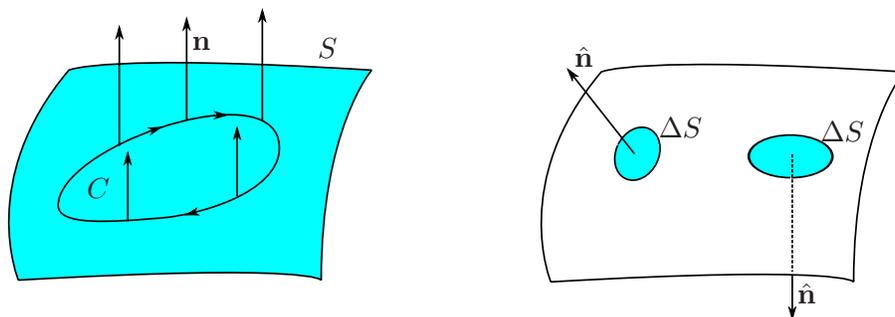


FIGURE 15.10. **Left:** If there is a continuous unit normal vector \mathbf{n} on a surface S , then when \mathbf{n} is transported along a closed curve in S , its initial direction should match the final direction. **Right:** A small patch ΔS of a surface S can be oriented in two different ways according to two possible choices of the unit normal vector $\hat{\mathbf{n}}$ or $-\hat{\mathbf{n}}$. If there is a “one-sided” surface, the face-up patch can be transported along a closed curve in S to the face-down patch at the same position on S . If S has a boundary, then the closed curve is not allowed to cross the boundary.

has the same shape but its normal is $-\hat{\mathbf{n}}$ (face-down patch) as shown in the right panel of Figure 15.10. 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 closed 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 .

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 procedure is shown in Figure 15.11. The resulting surface is the famous *Möbius strip* (named after the German mathematician August Möbius). It is one-sided. All curves winding

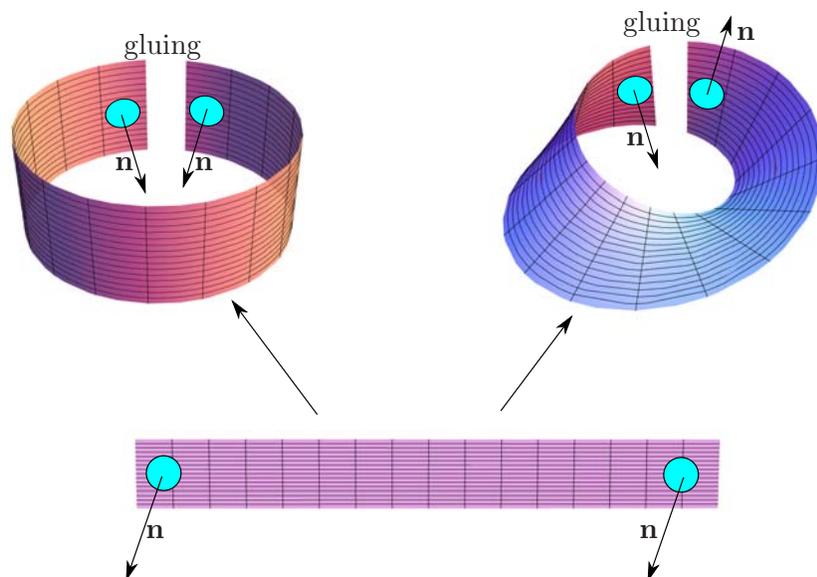


FIGURE 15.11. Construction of a two-sided surface (a portion of a cylinder) from a band by gluing its edges (left). Construction of a one-sided surface (a Möbius strip) from a band by gluing its edges after twisting the band (right).

about it traverse both sides of the glued rectangle without crossing its boundaries (the horizontal edges). A face-up patch can be transported into a face-up patch at the same position along a closed curve in the band.

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.

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.

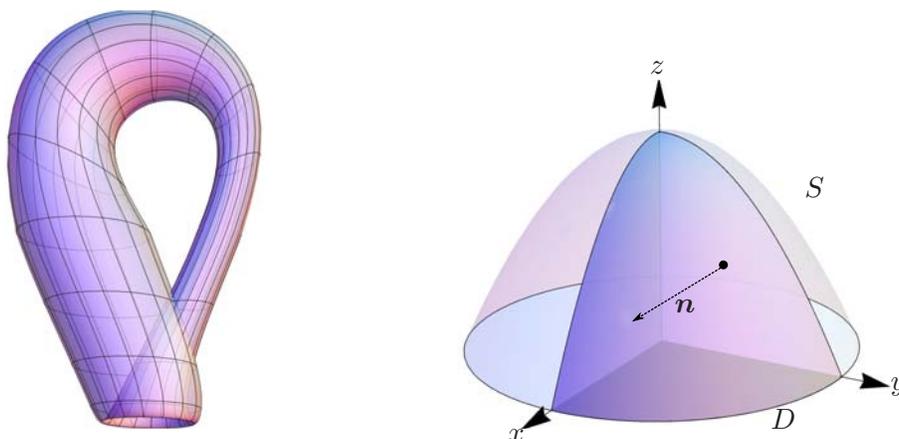


FIGURE 15.12. **Left:** A Klein bottle is an example of a one-sided closed surface (it has no boundaries). **Right:** An illustration to Example 15.8.

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 14.22). In particular, the flux of a continuous vector field across a smooth orientable surface exists.

113.3. Evaluation of the Flux of a Vector Field. Suppose that a surface S is a graph $z = g(x, y)$ and g has continuous partial derivatives in 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 90). 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 Φ_\uparrow (Φ_\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

$$\hat{\mathbf{n}} = \frac{1}{\|\mathbf{n}\|} \mathbf{n} = \frac{1}{J} (-g'_x, g'_y, 1), \quad J = \sqrt{1 + (g'_x)^2 + (g'_y)^2}.$$

In Section 108.1, it was established that the area of the portion of the graph above a planar region of area dA 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 vector field F is continuous, then the dot product $\mathbf{F} \cdot \mathbf{n}$ is a continuous function on D so that 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 a 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). One can think of D as the vertical projection of S onto the xy plane.

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 15.8. 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 . \square

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 Theorem 14.23, 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 15.9. 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 (14.36), and the normal vector is computed in Example 14.41: $\mathbf{n} = \sin(u)\mathbf{r}(u, v)$, 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 Parametric Surfaces. Nonorientable smooth 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

$$(15.13) \quad \mathbf{r}(u, v) = \left([R + u \cos(v/2)] \cos v, [R + u \cos(v/2)] \sin v, u \sin(v/2) \right),$$

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 (15.13) 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

$$\begin{aligned}\mathbf{r}'_u(0, v) &= (\cos(v/2) \cos v, \cos(v/2) \sin v, \sin(v/2)), \\ \mathbf{r}'_v(0, v) &= (-R \sin v, R \cos v, 0).\end{aligned}$$

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.

113.5. Exercises.

(1) Find the flux of a constant vector field $\mathbf{F} = (a, b, c)$ across the specified surface S :

- (i) S is a rectangle of area A in each of the coordinate planes oriented along the coordinate axis orthogonal to the rectangle
- (ii) S is the part of the plane $(x/a) + (y/b) + (z/c) = 1$ in the positive octant oriented outward from the origin and a, b , and c are positive
- (iii) S is the boundary of the pyramid whose base is the square $[-q, q] \times [-q, q]$ in the xy plane and whose vertex is $(0, 0, h)$
- (iv) S is the cylinder $x^2 + y^2 = R^2$, $0 \leq z \leq h$
- (v) S is the surface of a rectangular box oriented outward
- (vi) S is the sphere $x^2 + y^2 + z^2 = R^2$ oriented outward
- (vii) S is a torus oriented inward

(2) Find the flux of the vector field \mathbf{F} across the specified oriented surface S :

- (i) $\mathbf{F} = (xy, zx, xy)$ and S is the part of the paraboloid $z = 1 - x^2 - y^2$ that lies above the square $[0, 1] \times [0, 1]$ and is oriented upward
- (ii) $\mathbf{F} = (y, -x, z^2)$ and S is the part of the paraboloid $z = 1 - x^2 - y^2$ that lies above the xy plane and is oriented downward

- (iii) $\mathbf{F} = (xz, zy, z^2)$ and S is the part of the cone $z = \sqrt{x^2 + y^2}$ beneath the plane $z = 2$ in the first octant and is oriented upward
- (iv) $\mathbf{F} = (x, -z, y)$ and S the part of the sphere in the first octant and is oriented toward the origin
- (v) $\mathbf{F} = \mathbf{a} \times \mathbf{r}$, where \mathbf{a} is a constant vector and S is the sphere of radius R oriented outward and centered at the origin
- (vi) $\mathbf{F} = (2y + x, y + 2z - x, z - y)$, where S is the boundary of the cube with vertices $(\pm 1, \pm 1, \pm 1)$ and is oriented outward
- (vii) $\mathbf{F} = c\mathbf{r}/\|\mathbf{r}\|^3$, where $\mathbf{r} = (x, y, z)$, c is a constant, and S is the sphere of radius a that is centered at the origin and oriented inward
- (viii) $\mathbf{F} = (2y, x, -z)$ and S is the part of the paraboloid $y = 1 - x^2 - z^2$ in the first octant and is oriented upward
- (ix) $\mathbf{F} = (xy, zy, z)$ and S is the part of the plane $2x - 2y - z = 3$ that lies inside the cylinder $x^2 + y^2 = 1$ and is oriented upward
- (x) $\mathbf{F} = (x, y, z)$ and S is the part of the cylinder $x = z^2 + y^2$ that lies between the planes $x = 0$ and $x = 1$
- (xi) $\mathbf{F} = (x, y, z)$ and S is the sphere $x^2 + y^2 + z^2 = R^2$ oriented outward
- (xii) $\mathbf{F} = (f(x), g(y), h(z))$, where f , g , and h are continuous functions and S is the boundary of the rectangular box $[0, a] \times [0, b] \times [0, c]$ oriented outward
- (xiii) $\mathbf{F} = (y - z, z - x, x - y)$ and S is the part of the cone $x^2 + y^2 = z^2$, $0 \leq z \leq h$, oriented away from the z axis
- (3)** Use parametric equations of the specified surface S to evaluate the flux of the vector field across it:
- (i) $\mathbf{F} = (x, -y, z^2)$ and S is the part of the double cone $z^2 = x^2 + y^2$ between the planes $z = -1$ and $z = 1$
- (ii) $\mathbf{F} = (z^2 + y^2, x^2 + z^2, x^2 + y^2)$ and S is the boundary of the solid enclosed by the cylinder $x^2 + z^2 = 1$ and the planes $y = 0$ and $y = 1$; S is oriented outward
- (iii) $\mathbf{F} = (y, x, z)$ and S is the part of the sphere $x^2 + y^2 + z^2 = 4$ that lies outside the double cone $z^2 = 3(x^2 + y^2)$ and is oriented toward the origin
- (iv) $\mathbf{F} = (-y, x, z)$ and S is the torus with radii R and a that is oriented outward
- (v) $\mathbf{F} = (x^{-1}, y^{-1}, z^{-1})$ and S is the ellipsoid $(x/a)^2 + (y/b)^2 + (z/c)^2 = 1$ oriented outward
- (vi) $\mathbf{F} = (x^2, y^2, z^2)$ and S is the sphere $(x-a)^2 + (y-b)^2 + (z-c)^2 = R^2$ oriented outward

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, $\nabla \times \mathbf{F} = (\partial F_2/\partial x - \partial F_1/\partial y)\mathbf{e}_3$. This observation allows us to reformulate Green's theorem in the following vector form:

$$\oint_{\partial D} \mathbf{F} \cdot d\mathbf{r} = \iint_D (\text{curl } \mathbf{F}) \cdot \mathbf{e}_3 \, dA.$$

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. 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. 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 C 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 of S after the deformation and the orientations of the boundary of S and C can be compared. The boundary of S is *positively oriented* if it has the same orientation as C . The positively oriented boundary of S is denoted by ∂S . The procedure to define a positive orientation of the boundary of an oriented surface S is illustrated in Figure 15.13 (left panel).

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 (see the right panel of Figure 15.13).

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 the components of a vector field \mathbf{F} have continuous partial

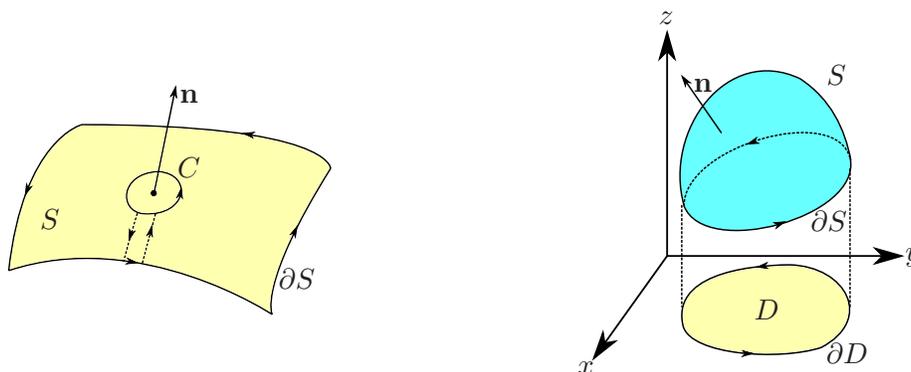


FIGURE 15.13. **Left:** The positive (or induced) orientation of the boundary of an oriented surface S . The surface S is oriented by its normal vector \mathbf{n} . Take a closed curve C in S that has counterclockwise orientation as viewed from the top of \mathbf{n} . Deform this curve toward the boundary of S . The boundary of S has positive orientation if it coincides with the orientation of C . **Right:** The surface S is the graph of a function on D . If S has upward orientation, then the positively oriented boundary ∂S is obtained from the positively (counterclockwise) oriented boundary ∂D .

derivatives 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 has continuous second-order partial derivatives 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 in which \mathbf{F} is replaced by $\nabla \times \mathbf{F}$:

$$\begin{aligned} \iint_S \operatorname{curl} \mathbf{F} \cdot \hat{\mathbf{n}} \, dS &= \iint_D (\operatorname{curl} \mathbf{F})_n \, dA, \\ (\operatorname{curl} \mathbf{F})_n &= -\left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\right) \frac{\partial z}{\partial x} - \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}\right) \frac{\partial z}{\partial y} \\ &\quad + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right), \end{aligned}$$

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. By the chain rule, $\mathbf{r}' = (x', y', g'_x x' + g'_y y')$. Therefore, $\mathbf{F} \cdot \mathbf{r}' = (F_1 + F_3 g'_x)x' + (F_2 + F_3 g'_y)y'$ and hence

$$\begin{aligned} \oint_{\partial S} \mathbf{F} \cdot d\mathbf{r} &= \int_a^b [(F_1 + F_3 g'_x)x' + (F_2 + F_3 g'_y)y'] dt \\ &= \oint_{\partial D} \left(F_1 + F_3 \frac{\partial z}{\partial x} \right) dx + \left(F_2 + F_3 \frac{\partial z}{\partial y} \right) dy \end{aligned}$$

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:

$$\begin{aligned} \oint_{\partial S} \mathbf{F} \cdot d\mathbf{r} &= \iint_D \left[\frac{\partial}{\partial x} \left(F_2 + F_3 \frac{\partial z}{\partial y} \right) - \frac{\partial}{\partial y} \left(F_1 + F_3 \frac{\partial z}{\partial x} \right) \right] dA \\ (15.14) \quad &= \iint_D (\operatorname{curl} \mathbf{F})_n dA = \iint_S \operatorname{curl} \mathbf{F} \cdot \hat{\mathbf{n}} dS, \end{aligned}$$

where the middle equality is verified by the direct evaluation of the partial derivatives using the chain rule. For example,

$$\begin{aligned} \frac{\partial}{\partial x} F_2(x, y, g(x, y)) &= \frac{\partial F_2}{\partial x} + \frac{\partial F_2}{\partial z} \frac{\partial g}{\partial x} \\ \frac{\partial}{\partial x} \left(F_3 \frac{\partial g}{\partial y} \right) &= \left(\frac{\partial F_3}{\partial x} + \frac{\partial F_3}{\partial z} \frac{\partial g}{\partial x} \right) \frac{\partial g}{\partial y} + F_3 \frac{\partial^2 g}{\partial x \partial y}. \end{aligned}$$

The terms containing the mixed derivatives $g''_{xy} = g''_{yx}$ are cancelled out owing to 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 in (15.14) holds by Theorem 14.22 ($dS = J dA$ and $\mathbf{n} = J\hat{\mathbf{n}}$). \square

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 .

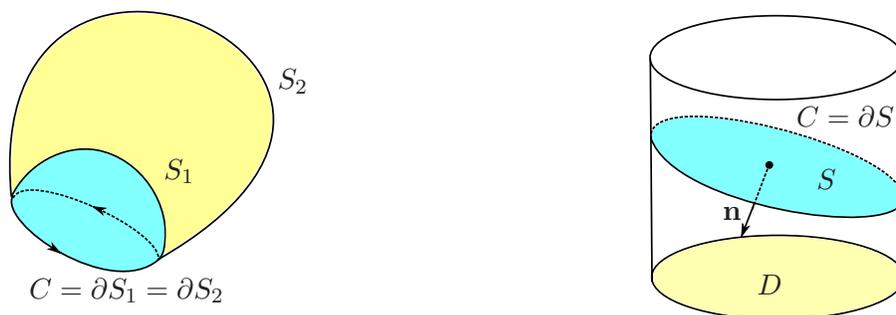


FIGURE 15.14. **Left:** Given the curve $C = \partial S$, the surface S may have any desired shape in Stokes' theorem. The surfaces S_1 and S_2 have the same boundaries $\partial S_1 = \partial S_2 = C$. The line integral over C can be transformed into the flux integral either across S_1 or S_2 . **Right:** An illustration to Example 15.10. The integration contour C is the intersection of the cylinder and a plane. When applying Stokes' theorem, the simplest choice of a surface, whose boundary is C , is the part of the plane that lies inside the cylinder.

This freedom should be used to make S as simple as possible (see the left panel of Figure 15.14).

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 $\mathbf{B} = \text{curl } \mathbf{F}$ and calculate the flux of \mathbf{B} across S .

EXAMPLE 15.10. 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 $x^2 + y^2 \leq 1$ (as shown in the right panel of Figure 15.14). 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 15.11. Evaluate the line integral of $\mathbf{F} = (z^2y, -z^2x, z)$ 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 $x^2 + y^2 \leq 1$ in the first quadrant (D is the vertical projection of the said part of the paraboloid onto the xy plane). 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,

$$\mathbf{B} = \nabla \times \mathbf{F} = \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ z^2y & -z^2x & z \end{pmatrix} = (2zx, 2zy, -2z^2)$$

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 $\mathbf{B} = \nabla \times \mathbf{F}$ 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 S_a be a simple region in the plane such that \mathbf{r}_0 is an interior point of S_a . Let a be the radius of the smallest disk centered at \mathbf{r}_0 that contains

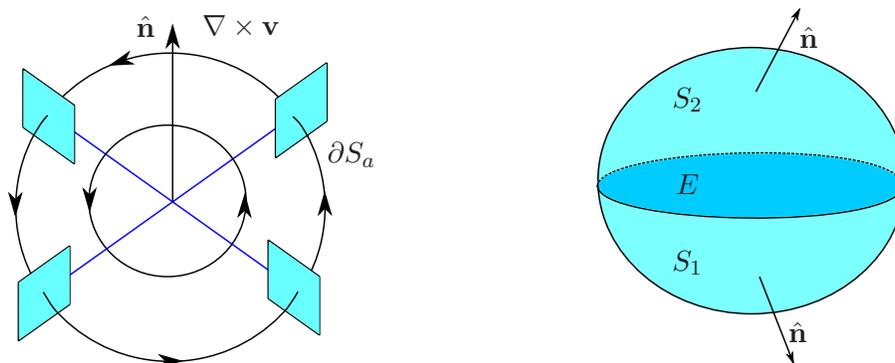


FIGURE 15.15. **Left:** An illustration to the mechanical interpretation of the curl. A small paddle wheel whose axis of rotation is parallel to \mathbf{n} is placed into a fluid flow. The work done by the pressure force along the loop ∂S_a through the paddles causes rotation of the wheel. It is determined by the line integral of the velocity vector field \mathbf{v} . The work is maximal (the fastest rotation of the wheel) when $\hat{\mathbf{n}}$ is aligned parallel with $\nabla \times \mathbf{v}$. **Right:** An illustration to Corollary 15.2. The flux of a vector field across the surface S_1 can be related to the flux across S_2 by the divergence theorem if S_1 and S_2 have a common boundary and their union encloses a solid region E .

S_a . If ΔS_a is the area of S_a , consider the *circulation of a vector field per unit area* at a point \mathbf{r}_0 defined as the ratio $\oint_{\partial S_a} \mathbf{F} \cdot d\mathbf{r} / \Delta S_a$ in the limit $a \rightarrow 0$, that is, in the limit when the region S_a shrinks to the point \mathbf{r}_0 . Then, by virtue of Stokes' theorem and the integral mean value theorem,

$$\lim_{a \rightarrow 0} \frac{1}{\Delta S_a} \oint_{\partial S_a} \mathbf{F} \cdot d\mathbf{r} = \lim_{a \rightarrow 0} \frac{1}{\Delta S_a} \iint_{S_a} \mathbf{B} \cdot \hat{\mathbf{n}} \, dS = \mathbf{B}_0 \cdot \hat{\mathbf{n}} = (\text{curl } \mathbf{F})_0 \cdot \hat{\mathbf{n}}.$$

Indeed, 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 $\Delta S_a f(\mathbf{r}_a)$. As $a \rightarrow 0$, $\mathbf{r}_a \rightarrow \mathbf{r}_0$ and, by the continuity of f , $f(\mathbf{r}_a) \rightarrow f(\mathbf{r}_0)$. Thus, *the circulation of a vector field per unit area is maximal if the normal to the area element is in the same direction as the curl of the vector field, and the maximal circulation equals the magnitude of the curl.*

This observation has the following mechanical interpretation illustrated in the left panel of Figure 15.15. 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 of rotation is directed $\hat{\mathbf{n}}$ along

$\hat{\mathbf{n}}$. The fluid exerts pressure on the paddles, causing the paddle wheel to rotate. The work done by the pressure force is determined by the line integral along the loop ∂S_a through the paddles. The more work done by the pressure force, the faster the wheel rotates. The wheel rotates fastest (maximal work) when its axis $\hat{\mathbf{n}}$ is parallel to $\text{curl } \mathbf{v}$ because, in this case, the normal component of the curl, $(\nabla \times \mathbf{v}) \cdot \hat{\mathbf{n}} = \|\nabla \times \mathbf{v}\|$, is maximal. For this reason, the curl is often called the *rotation* of a vector field and also denoted as $\text{rot } \mathbf{F} = \nabla \times \mathbf{F}$.

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 whose components have continuous second-order partial derivatives. Then \mathbf{F} can be decomposed into the sum of conservative and a rotational vector fields; that is, there is a function f and a vector field \mathbf{A} such that

$$\mathbf{F} = \nabla f + \nabla \times \mathbf{A}.$$

The vector field \mathbf{A} is called a *vector potential* of the field \mathbf{F} . The vector potential is not unique. It can be changed by adding the gradient of a function, $\mathbf{A} \rightarrow \mathbf{A} + \nabla g$, because

$$\nabla \times (\mathbf{A} + \nabla g) = \nabla \times \mathbf{A} + \nabla \times (\nabla g) = \nabla \times \mathbf{A}$$

for any g that has continuous second-order partial derivatives. Electromagnetic waves are rotational components of electromagnetic fields, while the Coulomb field created by static charges is conservative. The velocity vector field of an incompressible fluid (like water) is a rotational vector field.

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 throughout 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 $\text{curl } \mathbf{F} = \mathbf{0}$ throughout E , then, by Stokes' theorem,

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = \iint_S \text{curl } \mathbf{F} \cdot \hat{\mathbf{n}} dS = 0$$

for any simple closed curve C in E . By the path independence property, \mathbf{F} is conservative. The hypothesis that E is simply connected is crucial. For example, if E is the entire space with the z axis removed (see Study Problem 15.2), 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 .

114.6. Study Problem.

Problem 15.8. *Prove that the flux of a continuous rotational vector field \mathbf{F} vanishes across any smooth, closed, and orientable surface. What can be said about a flux in a flow of an incompressible fluid?*

SOLUTION: A continuous rotational vector field can be written as the curl of a vector field \mathbf{A} whose components have continuous partial derivatives, $\mathbf{F} = \nabla \times \mathbf{A}$. Consider a smooth closed simple contour C in a surface S . It cuts S into two pieces S_1 and S_2 . Suppose that S is oriented outward. Then the induced orientations of the boundaries ∂S_1 and ∂S_2 are opposite: $\partial S_1 = -\partial S_2$. The latter also holds if S is oriented inward. By virtue of Stokes' theorem,

$$\begin{aligned} \iint_S (\nabla \times \mathbf{A}) \cdot \hat{\mathbf{n}} \, dS &= \iint_{S_1} (\nabla \times \mathbf{A}) \cdot \hat{\mathbf{n}} \, dS + \iint_{S_2} (\nabla \times \mathbf{A}) \cdot \hat{\mathbf{n}} \, dS \\ &= \oint_{\partial S_1} \mathbf{A} \cdot d\mathbf{r} + \oint_{\partial S_2} \mathbf{A} \cdot d\mathbf{r} \\ &= \oint_{\partial S_1} \mathbf{A} \cdot d\mathbf{r} + \oint_{-\partial S_1} \mathbf{A} \cdot d\mathbf{r} = 0. \end{aligned}$$

Recall that the line integral changes its sign when the orientation of the curve is reversed. Since the flow of an incompressible fluid is described by a rotational vector field, the flux across a closed surface always vanishes in such a flow. \square

114.7. Exercises.

(1) Verify Stokes' theorem for the given vector field \mathbf{F} and surface S by calculating the circulation of \mathbf{F} along ∂S and the flux of $\nabla \times \mathbf{F}$ across S :

- (i) $\mathbf{F} = (y, -x, z)$ and S is the part of the sphere $x^2 + y^2 + z^2 = 2$ that lies above the plane $z = 1$
- (ii) $\mathbf{F} = (x, y, xyz)$ and S is the part of the plane $2x + y + z = 4$ in the first octant
- (iii) $\mathbf{F} = (y, z, x)$ and S is the part of the plane $x + y + z = 0$ inside the sphere $x^2 + y^2 + z^2 = a^2$

(2) Use Stokes' theorem to evaluate the line integral of the vector field \mathbf{F} along the specified closed contour C :

- (i) $\mathbf{F} = (x + y^2, y + z^2, z + x^2)$ and C is the triangle traversed as $(1, 0, 0) \rightarrow (0, 1, 0) \rightarrow (0, 0, 1) \rightarrow (1, 0, 0)$
- (ii) $\mathbf{F} = (yz, 2xz, e^{xy})$ and C is the intersection of the cylinder $x^2 + y^2 = 1$ and the plane $z = 3$ oriented clockwise as viewed from above
- (iii) $\mathbf{F} = (xy, 3z, 3y)$ and C is the intersection of the plane $x + y = 1$ and the cylinder $y^2 + z^2 = 1$
- (iv) $\mathbf{F} = (z, y^2, 2x)$ and C is the intersection of the plane $x + y + z = 5$ and the cylinder $x^2 + y^2 = 1$; the contour C is oriented counterclockwise as viewed from the top of the z axis
- (v) $\mathbf{F} = (-yz, xz, 0)$ and C is the intersection of the hyperbolic paraboloid $z = y^2 - x^2$ and the cylinder $x^2 + y^2 = 1$; C is oriented clockwise as viewed from the top of the z axis
- (vi) $\mathbf{F} = (z^2y/2, -z^2x/2, 0)$ and C is the boundary of the part of the cone $z = 1 - \sqrt{x^2 + y^2}$ that lies in the first quadrant; C is oriented counterclockwise as viewed from the top of the z axis.
- (vii) $\mathbf{F} = (y - z, -x, x)$ and C is the intersection of the cylinder $x^2 + y^2 = 1$ and the paraboloid $z = x^2 + (y - 1)^2$; C is oriented counterclockwise as viewed from the top of the z axis
- (viii) $\mathbf{F} = (y - z, z - x, x - y)$ and C is the ellipse $x^2 + y^2 = a^2$, $(x/a) + (z/b) = 1$, $a > 0$, $b > 0$, oriented positively when viewed from the top of the z axis
- (ix) $\mathbf{F} = (y + z, z + x, x + y)$ and C is the ellipse $x = a \sin^2 t$, $y = 2a \sin t \cos t$, $z = a \cos^2 t$, $0 \leq t \leq \pi$, oriented in the direction of increasing t
- (x) $\mathbf{F} = (y^2 - z^2, z^2 - x^2, x^2 - y^2)$ and C is the intersection of the surface of the cube $[0, a] \times [0, a] \times [0, a]$ by the plane $x + y + z = 3a/2$, oriented counterclockwise when viewed from the top of the x axis
- (xi) $\mathbf{F} = (y^2z^2, x^2z^2, x^2y^2)$, where C is the closed curve traced out by the vector function $\mathbf{r}(t) = (a \cos t, a \cos(2t), a \cos(3t))$ in the direction of increasing t

(3) Let C be a closed curve in the plane $\mathbf{n} \cdot \mathbf{r} = d$ that bounds a region of area A . Find

$$\oint_C (\mathbf{n} \times \mathbf{r}) \cdot d\mathbf{r}.$$

(4) Use Stokes' theorem to find the work done by the force \mathbf{F} in moving a particle along the specified closed path C :

- (i) $\mathbf{F} = (-yz, zx, yx)$ and C is the triangle $(0, 0, 6) \rightarrow (2, 0, 0) \rightarrow (0, 3, 0) \rightarrow (0, 0, 6)$
- (ii) $\mathbf{F} = (-yz, xz, z^2)$ and C is the boundary of the part of the paraboloid $z = 1 - x^2 - y^2$ in the first octant that is traversed clockwise as viewed from the top of the z axis
- (iii) $\mathbf{F} = (y + \sin x, z^2 + \cos y, x^3)$ and C is traversed by $\mathbf{r}(t) = (\sin t, \cos t, \sin(2t))$ for $0 \leq t \leq 4\pi$ (*Hint*: Observe that C lies in the surface $z = 2xy$.)

(5) Find the line integral of $\mathbf{F} = (e^{x^2} - yz, e^{y^2} - xz, z^2 - xy)$ along C , which is the helix $x = a \cos t, y = a \sin t, z = ht/(2\pi)$ from the point $(a, 0, 0)$ to the point $(a, 0, h)$. *Hint*: Supplement C by the straight line segment BA to make a closed curve and then use Stokes' theorem.

(6) Suppose that a surface S satisfies the hypotheses of Stokes' theorem and the functions f and g have continuous partial derivatives. Show that

$$\oint_{\partial S} (f \nabla g) \cdot d\mathbf{r} = \iint_S (\nabla f \times \nabla g) \cdot \hat{\mathbf{n}} dS.$$

Use the result to show that the circulation of the vector fields of the form $\mathbf{F} = f \nabla f$ and $\mathbf{F} = f \nabla g + g \nabla f$ vanishes along ∂S .

(7) Consider a rotationally symmetric solid. Let the solid be rotating about the symmetry axis at a constant rate ω (angular velocity). Let \mathbf{w} be the vector parallel to the symmetry axis such that $\|\mathbf{w}\| = \omega$ and the rotation is counterclockwise as viewed from the top of \mathbf{w} . If the origin is on the symmetry axis, show that the linear velocity vector field in the solid is given by $\mathbf{v} = \mathbf{w} \times \mathbf{r}$, where \mathbf{r} is the position vector of a point in the solid. Next, show that $\nabla \times \mathbf{v} = 2\mathbf{w}$. This gives another relation between the curl of a vector field and rotations.

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 15.12. Find the divergence of the vector field $\mathbf{F} = (x^3 + \cos(yz), y + \sin(x^2z), xyz)$.

SOLUTION: One has

$$\operatorname{div} \mathbf{F} = (x^3 + \cos(yz))'_x + (y + \sin(x^2z))'_y + (xyz)'_z = 3x^2 + 1 + yx.$$

□

COROLLARY 15.1. A rotational vector field whose components have continuous partial derivatives 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 the components of \mathbf{A} have continuous second-order partial derivatives because, by the hypothesis, the components of \mathbf{F} have continuous first-order partial derivatives. 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 the components of \mathbf{A} have continuous second-order partial derivatives (Clairaut's theorem holds for its components; see Section 111.4). □

Laplace Operator. Let $\mathbf{F} = \nabla f$. Then $\operatorname{div} \mathbf{F} = \nabla \cdot \nabla f = f''_{xx} + f''_{yy} + f''_{zz}$. The operator $\nabla \cdot \nabla = \nabla^2$ is called the *Laplace operator*.

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 $\mathbf{r}(t) = (x(t), y(t))$, $a \leq t \leq b$, traces out the boundary C of D in the positive (counterclockwise) direction, then

$$\hat{\mathbf{T}}(t) = \frac{1}{\|\mathbf{r}'(t)\|} (x'(t), y'(t)), \quad \hat{\mathbf{n}}(t) = \frac{1}{\|\mathbf{r}'(t)\|} (y'(t), -x'(t)),$$

$$\hat{\mathbf{T}} \cdot \hat{\mathbf{n}} = 0$$

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} ,

$$\oint_C \mathbf{F} \cdot \hat{\mathbf{n}} ds = \oint_C \mathbf{G} \cdot d\mathbf{r} = \iint_D \left(\frac{\partial G_2}{\partial x} - \frac{\partial G_1}{\partial y} \right) dA = \iint_D \left(\frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} \right) dA.$$

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 (e.g., the mass transfer by a planar flow across the boundary of 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 components of a vector field \mathbf{F} have continuous partial derivatives in 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 15.13. 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! \square

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 (see Figure 15.15, right panel). 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 . Note that S_1 and S_2 must be oriented so that their union is ∂E ; that is, it has outward orientation.

EXAMPLE 15.14. *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

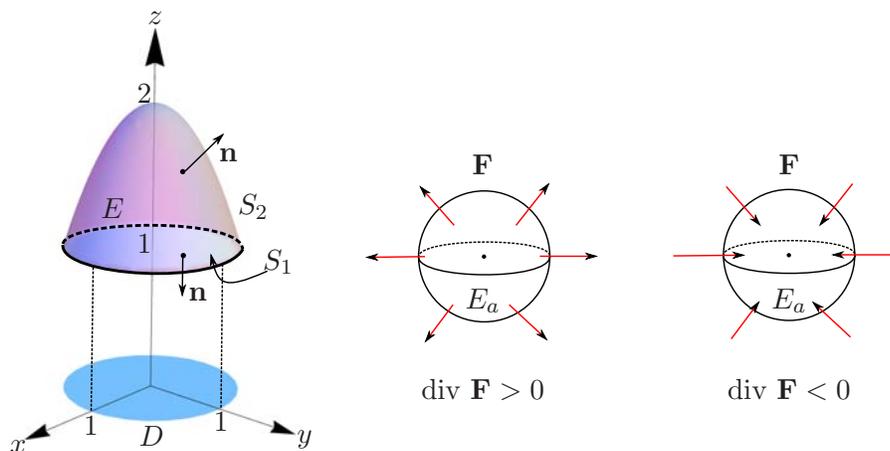


FIGURE 15.16. **Left:** An illustration to Example 15.14. The solid region E is enclosed by the paraboloid $z = 2 - x^2 - y^2$ and the plane $z = 1$. By Corollary 15.2, the flux of a vector field across the part S_2 of the paraboloid that has upward orientation can be converted to the flux across the part S_1 of the plane that has downward orientation. The union of S_1 and S_2 has outward orientation. **Right:** The divergence of a vector field \mathbf{F} determines the density of sources of \mathbf{F} . If $\operatorname{div} \mathbf{F} > 0$ at a point, then the flux of \mathbf{F} across a surface that encloses a small region E_a containing the point is positive (a “faucet”). If $\operatorname{div} \mathbf{F} < 0$ at a point, then the flux of \mathbf{F} across a surface that encloses a small region E_a containing the point is negative (a “sink”).

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 $E = \{(x, y, z) \mid z_{\text{bot}} = 1 \leq z \leq 2 - x^2 - y^2 = z_{\text{top}}, (x, y) \in D\}$.

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 the flux directly via the surface integral to appreciate the power of the divergence theorem!

COROLLARY 15.3. *The flux of a rotational vector field, whose components have continuous partial derivatives, 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:

$$\operatorname{div}(\nabla f + \nabla \times \mathbf{A}) = \nabla^2 f + \nabla \cdot (\nabla \times \mathbf{A}) = \nabla^2 f.$$

So the divergence of a vector field is determined by the action of the Laplace operator of the scalar potential f of the vector field. 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_a and an interior point \mathbf{r}_0 of E_a . Let a be the radius of the smallest ball that contains E_a and is centered at \mathbf{r}_0 . Let us calculate the outward flux *per unit volume* of a vector field \mathbf{F} across the boundary ∂E_a , which is defined by the ratio $\iint_{\partial E_a} \mathbf{F} \cdot \hat{\mathbf{n}} dS / \Delta V_a$ in the limit $a \rightarrow 0$, that is, when E_a shrinks to \mathbf{r}_0 . Suppose that components of \mathbf{F} have continuous partial derivatives. By virtue of the divergence theorem and the integral mean value theorem,

$$\lim_{a \rightarrow 0} \frac{1}{\Delta V_a} \iint_{\partial E_a} \mathbf{F} \cdot \hat{\mathbf{n}} dS = \lim_{a \rightarrow 0} \frac{1}{\Delta V_a} \iiint_{E_a} \operatorname{div} \mathbf{F} dV = \operatorname{div} \mathbf{F}(\mathbf{r}_0).$$

Indeed, 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 $\Delta V_a \operatorname{div} \mathbf{F}(\mathbf{r}_a)$. In the limit $a \rightarrow 0$, $\mathbf{r}_a \rightarrow \mathbf{r}_0$ and $\operatorname{div} \mathbf{F}(\mathbf{r}_a) \rightarrow \operatorname{div} \mathbf{F}(\mathbf{r}_0)$. 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. 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. In particular, the flux of the electric field \mathbf{E} across a closed surface S is determined by the total electric charge in the region enclosed by S . In contrast, the magnetic field \mathbf{B} is a rotational vector field and hence is divergence free. So there are no *magnetic charges* also known as *magnetic monopoles*. These two laws of physics are stated in the form:

$$\operatorname{div} \mathbf{E} = 4\pi\sigma, \quad \operatorname{div} \mathbf{B} = 0,$$

where σ is the density of electric charges. Flow lines of the magnetic field are closed, while flow lines of the electric field end at points where electric charges are located (as indicated by the arrows in the right panel of Figure 15.16).

115.5. Study Problem.

Problem 15.9. (Volume of a Solid as the Surface Integral).

Let E be bounded by a piecewise smooth surface $S = \partial E$ oriented by an outward unit normal vector $\hat{\mathbf{n}}$. Prove that the volume of E is

$$V(E) = \frac{1}{3} \iint_{\partial E} \hat{\mathbf{n}} \cdot \mathbf{r} \, dS.$$

SOLUTION: Consider three vector fields $\mathbf{F}_1 = (x, 0, 0)$, $\mathbf{F}_2 = (0, y, 0)$, and $\mathbf{F}_3 = (0, 0, z)$. Then

$$\operatorname{div} \mathbf{F}_1 = 1, \quad \operatorname{div} \mathbf{F}_2 = 1, \quad \operatorname{div} \mathbf{F}_3 = 1.$$

Then, by virtue of the equality $\mathbf{r} = \mathbf{F}_1 + \mathbf{F}_2 + \mathbf{F}_3$ and by the divergence theorem,

$$\begin{aligned} \iint_{\partial E} \hat{\mathbf{n}} \cdot \mathbf{r} \, dS &= \iint_{\partial E} \hat{\mathbf{n}} \cdot \mathbf{F}_1 \, dS + \iint_{\partial E} \hat{\mathbf{n}} \cdot \mathbf{F}_2 \, dS + \iint_{\partial E} \hat{\mathbf{n}} \cdot \mathbf{F}_3 \, dS \\ &= \iiint_E (\operatorname{div} \mathbf{F}_1 + \operatorname{div} \mathbf{F}_2 + \operatorname{div} \mathbf{F}_3) \, dV = 3 \iiint_E dV \\ &= 3V(E), \end{aligned}$$

and the required result follows. \square

115.6. Exercises.

(1) Find the divergence of the specified vector field:

- (i) $\mathbf{F} = \nabla f$, where $f = \sqrt{x^2 + y^2 + z^2}$
- (ii) $\mathbf{F} = \mathbf{r}/r$, where $r = \|\mathbf{r}\|$
- (iii) $\mathbf{F} = \mathbf{a}f(r)$, where $r = \|\mathbf{r}\|$ and \mathbf{a} is a constant vector
- (iv) $\mathbf{F} = \mathbf{r}f(r)$, where $r = \|\mathbf{r}\|$. When does the divergence vanish?
- (v) $\mathbf{F} = \mathbf{a}g$, where \mathbf{a} is a constant vector and g is a differentiable function. When does the divergence vanish?
- (vi) $\mathbf{F} = \mathbf{a} \times \mathbf{r}$, where \mathbf{a} is a constant vector
- (vii) $\mathbf{F} = \mathbf{a} \times \nabla g$, where \mathbf{a} is a constant vector. When does the divergence vanish?
- (viii) $\mathbf{F} = \mathbf{a} \times \mathbf{G}$, where \mathbf{a} is a constant vector. When does the divergence vanish?

(2) Prove the following identities, assuming that the appropriate partial derivatives of vector fields and functions exist and are continuous:

- (i) $\operatorname{div}(f\mathbf{F}) = f\operatorname{div} \mathbf{F} + \mathbf{F} \cdot \nabla f$
- (ii) $\operatorname{div}(\mathbf{F} \times \mathbf{G}) = \mathbf{G} \cdot \operatorname{curl} \mathbf{F} - \mathbf{F} \cdot \operatorname{curl} \mathbf{G}$
- (iii) $\operatorname{div}(\nabla f \times \nabla g)$
- (iv) $\operatorname{curl} \operatorname{curl} \mathbf{F} = \nabla(\operatorname{div} \mathbf{F}) - \nabla^2 \mathbf{F}$

(3) Let \mathbf{a} be a fixed vector and let $\hat{\mathbf{n}}$ be the unit normal to a planar closed curve C directed outward from the region bounded by C . Show that $\oint_C \mathbf{a} \cdot \hat{\mathbf{n}} \, ds = 0$.

(4) Let C be a simple closed curve in the xy plane and let $\hat{\mathbf{n}}$ be the unit normal to C directed outward from the region D bounded by C . If $A(D)$ is the area of D , find $\oint_C \mathbf{r} \cdot \hat{\mathbf{n}} \, ds$.

(5) Verify the divergence theorem for the given vector field \mathbf{F} on the region E :

- (i) $\mathbf{F} = (3x, yz, 3xz)$ and E is the rectangular box $[0, a] \times [0, b] \times [0, c]$

- (ii) $\mathbf{F} = (3x, 2y, z)$ and E is the solid bounded by the paraboloid $z = a^2 - x^2 - y^2$ and the plane $z = 0$

(6) Let \mathbf{a} be a constant vector and let S be a closed smooth surface oriented outward by the unit normal vector $\hat{\mathbf{n}}$. Prove that

$$\iint_S \mathbf{a} \cdot \hat{\mathbf{n}} \, dS = 0.$$

(7) Evaluate the flux of the given vector field across the specified closed surface S . In each case, determine the kind of source of \mathbf{F} in the region enclosed by S (sink or faucet):

- (i) $\mathbf{F} = (x^2, y^2, z^2)$ and S is the boundary of the rectangular box $[0, a] \times [0, b] \times [0, c]$ oriented outward
- (ii) $\mathbf{F} = (x^3, y^3, z^3)$ and S is the sphere $x^2 + y^2 + z^2 = R^2$ oriented inward
- (iii) $\mathbf{F} = (xy, y^2 + \sin(xz), \cos(yx))$ and S is bounded by the parabolic cylinder $z = 1 - x^2$ and the planes $z = 0, y = 0, y + z = 2$; S is oriented outward
- (iv) $\mathbf{F} = (-xy^2, -yz^2, zx^2)$ and S is the sphere $x^2 + y^2 + z^2 = 1$ with inward orientation
- (v) $\mathbf{F} = (xy, z^2y, zx)$ and S is the boundary of the solid region inside the cylinder $x^2 + y^2 = 4$ and between the planes $z = \pm 2$; S is oriented outward
- (vi) $\mathbf{F} = (xz^2, y^3/3, zy^2 + xy)$ and S is the boundary of the part of the ball $x^2 + y^2 + z^2 \leq 1$ in the first octant; S is oriented inward
- (vii) $\mathbf{F} = (yz, z^2x + y, z - xy)$ and S is the boundary of the solid enclosed by the cone $z = \sqrt{x^2 + y^2}$ and the sphere $x^2 + y^2 + z^2 = 1$; S is oriented outward
- (viii) $\mathbf{F} = (x + \tan(yz), \cos(xz) - y, \sin(xy) + z)$ and S is the boundary of the solid region between the sphere $x^2 + y^2 + z^2 = 2z$ and the cone $z = \sqrt{x^2 + y^2}$
- (ix) $\mathbf{F} = (\tan(yz), \ln(1 + z^2x^2), z^2 + e^{yx})$ and S is the boundary of the smaller part of the ball $x^2 + y^2 + z^2 \leq a^2$ between two half-planes $y = x/\sqrt{3}$ and $y = \sqrt{3}x, x \geq 0$; S is oriented inward
- (x) $\mathbf{F} = (xy^2, xz, zx^2)$ and S is the boundary of the solid bounded by two paraboloids $z = x^2 + y^2$ and $z = 1 + x^2 + y^2$ and the cylinder $x^2 + y^2 = 4$; S is oriented outward
- (xi) $\mathbf{F} = (x, y, z)$ and S is the boundary of the solid obtained from the box $[0, 2a] \times [0, 2b] \times [0, 2c]$ by removing the smaller box $[0, a] \times [0, b] \times [0, c]$; S is oriented inward

- (xii) $\mathbf{F} = (x - y + z, y - z + x, z - x + y)$ and S is the surface $|x - y + z| + |y - z + x| + |z - x + y| = 1$ oriented outward
- (xiii) $\mathbf{F} = (x^3, y^3, z^3)$ and S is the sphere $x^2 + y^2 + z^2 = x$ oriented outward

(8) Let S_1 and S_2 be two smooth orientable surfaces that have the same boundary. Suppose that $\mathbf{F} = \text{curl } \mathbf{A}$ and the components of \mathbf{F} have continuous partial derivatives. Compare the fluxes of \mathbf{F} across S_1 and S_2 .

(9) Use the divergence theorem to find the flux of the given vector field \mathbf{F} across the specified surface S by an appropriate deformation of S :

- (i) $\mathbf{F} = (xy^2, yz^2, zy^2 + x^2)$ and S is the top half of the sphere $x^2 + y^2 + z^2 = 4$ oriented toward the origin
- (ii) $\mathbf{F} = (z \cos(y^2), z^2 \ln(1 + x^2), z)$ and S is the part of the paraboloid $z = 2 - x^2 - y^2$ above the plane $z = 1$; S is oriented upward
- (iii) $\mathbf{F} = (yz, xz, xy)$ and S is the cylinder $x^2 + y^2 = a^2$, $0 \leq z \leq b$, oriented outward from its axis of symmetry
- (iv) $\mathbf{F} = (yz + x^3, x^2z^3, xy)$ and S is the part of the cone $z = 1 - \sqrt{x^2 + y^2}$ oriented upward

(10) The electric field \mathbf{E} and the charge density σ are related by the Gauss law $\text{div } \mathbf{E} = 4\pi\sigma$. Suppose the charge density is constant, $\sigma = k > 0$, inside the sphere $x^2 + y^2 + z^2 = R^2$ and 0 otherwise. Find the outward flux of the electric field across the ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$ in the two following cases: first, when R is greater than any of a, b, c ; second, when R is less than any of a, b, c .

(11) Let \mathbf{F} be a vector field such that $\text{div } \mathbf{F} = \sigma_0 = \text{const}$ in a solid bounded region E and $\text{div } \mathbf{F} = 0$ otherwise. Let S be a closed smooth surface oriented outward. Consider all possible relative positions of E and S in space (the solid region bounded by S may or may not have an overlap with E). If V is the volume of E , what are all possible values of the flux of \mathbf{F} across S ?

(12) Use the vector form of Green's theorem to prove Green's first and second identities:

$$\iint_D f \nabla^2 g dA = \oint_{\partial D} (f \nabla g) \cdot \mathbf{n} ds - \iint_D g \nabla^2 f dA,$$

$$\iint_D (f \nabla^2 g - f \nabla^2 f) dA = \oint_{\partial D} (f \nabla g - g \nabla f) \cdot \mathbf{n} ds,$$

where D satisfies the hypotheses of Green's theorem and the appropriate partial derivatives of f and g exist and are continuous.

(13) Use the result of Study Problem 15.9 to find the volume of a solid bounded by the specified surfaces:

- (i) The planes $z = \pm c$ and the parametric surface $x = a \cos u \cos v + b \sin u \sin v$, $y = a \cos u \sin v - b \sin u \cos v$, $z = c \sin u$
- (ii) The planes $x = 0$ and $z = 0$ and the parametric surface $x = u \cos v$, $y = u \sin v$, $z = -u + a \cos v$, where $u \geq 0$ and $a > 0$
- (iii) The torus $x = (R + a \cos u) \sin v$, $y = (R + a \cos u) \sin v$, $z = a \sin u$

(14) Use the results of Study Problem 15.4 to express the divergence of a vector field in cylindrical and spherical coordinates:

$$\begin{aligned}\nabla \cdot \mathbf{F} &= \frac{1}{r} \frac{\partial(rF_r)}{\partial r} + \frac{1}{r} \frac{\partial F_\theta}{\partial \theta} + \frac{\partial F_z}{\partial z}, \\ \nabla \cdot \mathbf{F} &= \frac{1}{\rho^2} \frac{\partial(\rho^2 F_\rho)}{\partial \rho} + \frac{1}{\rho \sin \phi} \left(\frac{\partial(\sin \phi F_\phi)}{\partial \phi} + \frac{\partial F_\theta}{\partial \theta} \right).\end{aligned}$$

Hint: Show $\partial \hat{\mathbf{e}}_\rho / \partial \phi = \hat{\mathbf{e}}_\phi$, $\partial \hat{\mathbf{e}}_\rho / \partial \theta = \sin \theta \hat{\mathbf{e}}_\theta$, and similar relations for the partial derivatives of other unit vectors.

(15) Use the results of Study Problem 15.4 to express the Laplace operator in cylindrical and spherical coordinates:

$$\begin{aligned}\nabla^2 f &= \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial f}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2} + \frac{\partial^2 f}{\partial z^2}, \\ \nabla^2 f &= \frac{1}{\rho^2} \frac{\partial}{\partial \rho} \left(\rho^2 \frac{\partial f}{\partial \rho} \right) + \frac{1}{\rho^2 \sin \phi} \frac{\partial}{\partial \phi} \left(\sin \phi \frac{\partial f}{\partial \phi} \right) + \frac{1}{\rho^2 \sin^2 \phi} \frac{\partial^2 f}{\partial \theta^2}.\end{aligned}$$

Hint: Show $\partial \hat{\mathbf{e}}_\rho / \partial \phi = \hat{\mathbf{e}}_\phi$, $\partial \hat{\mathbf{e}}_\rho / \partial \theta = \sin \theta \hat{\mathbf{e}}_\theta$, and similar relations for the partial derivatives of other unit vectors.

Acknowledgments

The author would like to thank his colleagues Dr. David Groisser and Dr. Thomas Walsh for their useful suggestions and comments that were helpful to improve the textbook.

